Topic-Enhanced Personalized Retrieval-Based Chatbot

Hongjin Qian and Zhicheng Dou(✉)

Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
{ian,dou}@ruc.edu.cn

Abstract. Building a personalized chatbot has drawn much attention recently. A personalized chatbot is considered to have a consistent personality. There are two types of methods to learn the personality. The first mainly model the personality from explicit user profiles (*e.g.*, manually created persona descriptions). The second learn implicit user profiles from the user's dialogue history, which contains rich, personalized information. However, a user's dialogue history can be long and noisy as it contains long-time, multi-topic historical dialogue records. Such data noise and redundancy impede the model's ability to thoroughly and faithfully learn a consistent personality, especially when applied with models that have an input length limit (*e.g.*, BERT). In this paper, we propose deconstructing the long and noisy dialogue history into topic-dependent segments. We only use the topically related dialogue segment as context to learn the topic-aware user personality. Specifically, we design a **Top**ic-enhanced personalized **R**etrieval-based **C**hatbot, TopReC. It first deconstructs the dialogue history into topic-dependent dialogue segments and filters out irrelevant segments to the current query via a Heter-Merge-Reduce framework. It then measures the matching degree between the response candidates and the current query conditioned on each topic-dependent segment. We consider the matching degree between the response candidate and the cross-topic user personality. The final matching score is obtained by combining the topic-dependent and cross-topic matching scores. Experimental results on two large dataset show that TopReC outperforms all previous state-of-the-art methods.

Keywords: Personalization · Dialogue systems

1 Introduction

Developing an open-domain chatbot is a long-lasting task in the AI domain. The main reason is that an open-domain chatbot enables human-machine interactions via text from any domain irrespective of any constraints, which is considered as an ultimate goal of AI [3]. Methods for building an open-domain chatbot can be divided into two categories: generation-based and retrieval-based. The former leverages models (*e.g.*, encoder-decoder) to generate a new response [9,16,24].

J. Kamps et al. (Eds.): ECIR 2023, LNCS 13981, pp. 79–93, 2023.
https://doi.org/10.1007/978-3-031-28238-6_6

The latter retrieves a set of response candidates and chooses the most matched one as the output [10,12]. In this paper, we focus on the retrieval-based chatbot.

For an open-domain chatbot, a consistent personality is crucial as personality inconsistency might bring a sense of unpredictability and untrustworthiness [3]. To this end, many works seek to develop personalized chatbots that have consistent personalities. Previous works about personalized chatbots can be divided into three groups: (1) Early works assign a trainable user embedding to each user, which is updated during training and can be used to guide response retrieval or generation [8]; (2) some works model the user personality from explicit user profiles which are usually persona descriptions or attributes [15,19]; (3) recent works propose learning implicit user profiles from the user's dialogue history [11,13,23]. As discussed in [13], learning implicit user profiles from the dialogue history is advantageous regarding flexibility and effectiveness. First, a user's dialogue history is easy to obtain and update. Second, a user's dialogue history contains rich personalized information, such as the user's preferences and preferred expressions, which are essential for personality modeling.

However, a user's dialogue history contains long-time and multi-topic dialogue records, which might be redundant and noisy. Directly modeling the raw dialogue history has two challenges: (1) the redundant dialogue history contains a large number of historical dialogues. Feeding the whole dialogue history into a neural model might lead to model capacity overflow, especially when applying pre-trained language models with token length limits (*e.g.*, BERT has 512 length limits); (2) a user might have dynamic preferences over different topics. Modeling such topical preference dynamism is challenging to maintain the consistent personality of a personalized chatbot.

Most previous methods that learn implicit user profiles fail to overcome the two challenges. They usually learn several user representations directly from the whole dialogue history to guide response selection or generation. In this paper, we instead propose **deconstructing the long and noisy dialogue history into topic-dependent dialogue segments from which we learn the topic-aware implicit user profiles for personalized chatbot**. Modeling the implicit user profiles from the topic-dependent dialogue segments has three advantages: (1) as the long dialogue history is split into short dialogue segments, we can model the implicit user profile from each segment separately, which greatly reduces the required model capacity; (2) the topic-dependent dialogue segments are less noisy than the whole dialogue history. The reason is that the data noise in the dialogue history is primarily caused by its varied topics. And a user might have different personal preferences over various topics. For example, regarding the topic of organic vegetables, a vegetarian is likely to show great interest while a meatatarian would not; (3) given an input query, we can further measure the topical relevance between the topic-dependent dialogue segments and the query and filter out the irrelevant dialogue segments.

We design TopReC, which learns topic-aware user personality from topic-dependent dialogue segments. TopReC comprises two modules, the Topic-dependent Context Deconstruction module, and the Personalized Topic Matching module.

In the Topic-dependent Context Deconstruction module, the dialogue history is first reassembled into topic-dependent segments concerning the topical inter-relations among the historical dialogues. We then filter out the topic segments according to their relevance to the current query and only keep the topically related dialogue segments to model personality. When deconstructing the dialogue history into topic-dependent dialogue segments, one challenge is that the number of topics in each user's dialogue history is dynamic. To tackle such dynamism, we propose a Heter-Merge-Reduce method that can flexibly deconstruct the dialogue history into topic-dependent segments without deciding the topic number in advance. **In the Personalized Topic Matching module**, we measure the matching degree between the response candidate and each topic-dependent topic-dependent dialogue segment to obtain topic-dependent matching scores. Besides, we also measure the relevance between the response candidate and the cross-topic user profile to get the cross-topic matching score. The final matching scores are obtained by fusing the topic-dependent and cross-topic matching scores.

To verify the effectiveness of the proposed model TopReC, we conduct extensive experiments on two publicly available datasets for personalized response selection. The empirical results show that our model achieves the best performance overall baseline models. Our contributions are three-fold: (1) We point out that a user's dialogue history might reflect multi-faceted user interests, which indicates that a user's personalized preferences can be dynamic in the dialogue history; (2) We propose TopReC that deconstructs the user's dialogue history into topic-dependent segments via the Herter-Merge-Reduce method and performs personalized response selection by learning topic-aware implicit user profile from the topic-dependent dialogue segments; (3) Comprehensive experiments show that our model outperforms the state-of-the-art models.

2 Related Work

2.1 Retrieval-Based Chatbot

A retrieval-based chatbot aims to select a proper response from the response candidates given the current query. Early works mainly focus on single-turn dialogue, which takes the current query as the dialogue context. Afterwards, many works turn attention to the multi-turn dialogue, which takes a series of follow-up dialogues as the context [10]. To model the multi-turn dialogues, early works directly encode the multi-turn dialogues into hidden states via RNN and use the last hidden states to perform matching [10]. Later works mainly improve the multi-turn dialogue task by either obtaining deep context representation (*e.g.*, DAM [18]) or selecting useful dialogue context (*e.g.*, MSN [24]). With the huge success of the pre-trained language model (*e.g.*, BERT), recent works further improve the effectiveness of the multi-turn dialogue model. For example, Han et al. design self-supervised tasks to continue training BERT [6] and Xu et al. split the dialogue context into segments and feed them into BERT to compute relevance scores [17].

2.2 Personalized Chatbot

For open-domain chatbots, inconsistent personalities bring unpredictability and untrustworthiness to the end-user. Maintaining a consistent personality is the ultimate goal for the domain. To endow consistent personality to the open-domain chatbots, early works assign a user embedding to each user, which can be updated during training [8]. Inspired by the PERSONA-CHAT dataset [20], which contains user descriptions for each user, many works explore directly modelling the explicit user profile (*e.g.*, personality descriptions or user attributes). For example, DGMN [22] lets the dialogue context and the user profile interact with each other to learn a user representation. Some works also claim that the explicit user profile contains noise which might undermine the user modeling. Hence, models like CSN [25] and RSM-DCK [7] propose context selection to denoise the explicit user profile. Besides, Gu et al. concatenate the dialogue context, user profile, and the current query into a long sequence to feed into BERT to obtain the matching representation [5]. Though the explicit user profile can partly reflect the user's personality, it suffers from inflexibility and limited personalized information. Therefore, recent works propose learning implicit user profiles from the user's dialogue history [11,13]. In this paper, we argue that the user's dialogue history might be long and noisy. Directly learning the implicit user profile from the whole dialogue history might limit the model's performance. Therefore, we propose deconstructing the dialogue history into topic-dependent segments and learn topic-dependent user representations from the topic-dependent segment.

3 Methodology

3.1 Preliminary

For a retrieval-based chatbot, the major goal is to return the best response from a response repository given an input query. Formally, let $g(\cdot,\cdot)$ be a scoring model evaluating the matching degree of a candidate response r for an input query q under the context $\mathcal{C}$. The chatbot will choose the response r^* with the highest scores of g from a repository of responses $\mathcal{R}$ as the output. Hence, following [13], we have:

$$r^* = \arg\max_{r \in \mathcal{R}} g(q, r, C),$$

where the context $\mathcal{C}$ can be versatile. Taking the personalized chatbot as an example, $\mathcal{C}$ is the user profile that portrays the personality of the user. As mentioned in Sect. 1, the personalized chatbot can either learn the user personality from explicit user descriptions [7,19,25] or implicitly learn the personality from the dialogue history [11,13]. Inspired by recent works that highlight the effectiveness and availability of learning implicit user profiles from the user's dialogue history, we can define a mapping function $\mathcal{F}(\cdot)$ that learns the implicit user profile from the dialogue history. Formally, we have $\mathcal{C} = \mathcal{F}(H)$, where $H = \{(p_j, r_j)\}, j \in [1, t]$ represents the dialogue history of the user and (p_j, r_j) refers to the j-th historical post-response pair.

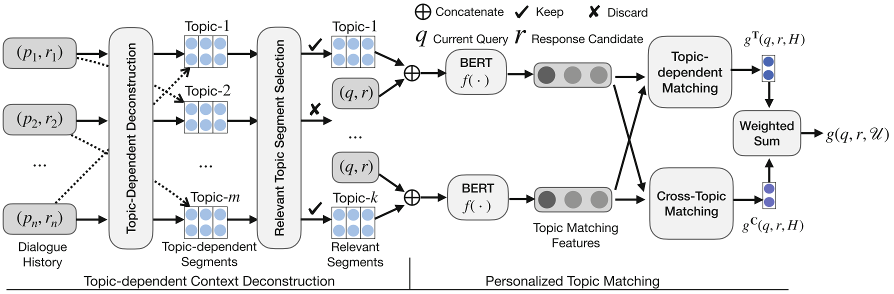

Fig. 1. The overview of TopReC.

3.2 The Proposed Model: TopReC

When learning the implicit user profile from the user's dialogue history, data noise and redundancy of dialogue history are two major issues we need to address. The data noise undermines the faithfulness of the learned user personality. And the data redundancy might lead to model capacity overflow (*e.g.*, BERT has 512 length limits). Our TopReC proposes deconstructing the long dialogue history into topic-dependent dialogue segments and filtering out dialogue segments that are irrelevant to the current query. Afterward, TopReC performs relevance matching between the response candidates and the topically-related dialogue segments to obtain the relevance scores.

Figure 1 shows the overview of TopReC. Specifically, TopReC comprises two modules: the Topic-dependent Context Deconstruction module and the Personalized Topic Matching module. The former deconstruct the dialogue history $H = \{(p_j, r_j)\}, j \in [1, t]$ into topic-dependent segments $\{H_1, \cdots, H_m\}, m \leq t$ and filter out irrelevant ones to get $\{\tilde{H}_1, \cdots, \tilde{H}_k\}, k \leq m$ that are topically-related to the current query. The latter applies a pre-trained encoder (*e.g.*, BERT) to perform topic-dependent matching and cross-topic matching to obtain the topic-dependent feature $g^{\mathbf{T}}(q, r, H)$ and the cross-topic matching feature $g^{\mathbf{C}}(q, r, H)$, respectively. The final matching score is computed by fusing the topic-dependent and cross-topic matching features.

3.3 Topic-Dependent Context Deconstruction

The Topic-dependent Context Deconstruction module obtains the topic-dependent dialogue segments via three steps: Heter-Merge-Reduce. We illustrate the procedures in Fig. 2. For a user's dialogue history $H = \{(p_j, r_j)\}, j \in [1, t]$ and the current query q, the goal of the module is to first deconstruct the dialogue history H into m topic-dependent segments $\{H_1, \cdots, H_m\}, m \leq t$ and then select k topic segments $\{\tilde{H}_1, \cdots, \tilde{H}_k\}, k \leq m$ that are topically-related to the current query q. We then will explain the details of each step of Heter-Merge-Reduce.

In the Heter step, we seek to decide the number of topics of a user's dialogue history. The difficulty of this step is that the number of topics in a user's dialogue history is changeable. Therefore, we cannot preset a fixed number of topics for all users. TopReC applies a soft margin to dynamically control the number of topics of each user's dialogue history. We achieve the goal by choosing m historical posts $P^{\text{topic}} = \{\hat{p}_1, \cdots, \hat{p}_m\}$ as the topic centers in which the mutual similarities of any two posts are smaller than a threshold γ. Specifically, we feed all the historical posts $\{p_1, \cdots, p_t\}$ into a pretrained encoder (*e.g.*, BERT). And we use the [CLS] token's hidden states of the i-th post p_i as its sentence representation $\mathbf{p}_i$. We then compute the point-wise similarities M of the historical posts:

$$M = \{\theta(p_i, p_j)\}, i, j \in [1, t], i \neq j, \tag{1}$$

$$\theta(p_i, p_j) = \frac{\mathbf{p}_i \cdot \mathbf{p}_j}{\|\mathbf{p}_i\|_2 \cdot \|\mathbf{p}_j\|_2}, \tag{2}$$

$$\mathbf{p} = \text{Pool}_{cls}(\text{BERT}(p)), \tag{3}$$

where $\theta(\cdot, \cdot)$ is the cosine similarity function.

After obtaining the point-wise similarities of all historical posts, we can choose the topic center posts $P^{\text{topic}} = \{\hat{p}_1, \cdots, \hat{p}_m\}$ by:

$$P^{\text{topic}} = \{\hat{p}_m\}, \theta(\hat{p}_m, \hat{p}_j) < \gamma, \hat{p}_m \neq \hat{p}_j. \tag{4}$$

In the merge step, we assign each historical dialogue (p_j, r_j) to a topic segment $H_n, n \in [1, m]$ of which the topic center $\hat{p}_n$ is the most similar to p_j:

$$(p_j, r_j) \rightarrow H_n; n = \text{argmax}_{n \in [1,m]} \theta(p_j, \hat{p}_n). \tag{5}$$

In the reduce step, we remove the negative impact of the irrelevant topic segments. Thus, we prune the topic segments $\{H_1, \cdots, H_m\}$ to $\{\tilde{H}_1, \cdots, \tilde{H}_k\}, k \leq m$ by measuring the similarity $\theta(q, \hat{p}_n), n \in [1, m]$ between the topic center $\hat{p}_n, n \in [1, m]$ and the current query q. We keep the k topic segments with the highest similarity score $\theta(q, \hat{p}_n)$. We will discuss the impact of the choice of k in Sect. 4.5.

Taking Fig. 2 as an example, we explain the Heter-Merge-Reduce method. In the Heter step, we choose the historical posts $\{p_1, p_5, p_6\}$ as the topic centers. We assign historical post-response pairs to the most similar topic segment in the Merge step. In the Reduce step, we keep $k = 2$ topic segments and filter out the segments centered by p_3.

3.4 Personalized Topic Matching

As mentioned in Sect. 1, a user's personal preferences can be dynamic over topics. Therefore, instead of modelling the user personality from the whole dialogue history, we propose modelling the topic-aware personality from topic-dependent

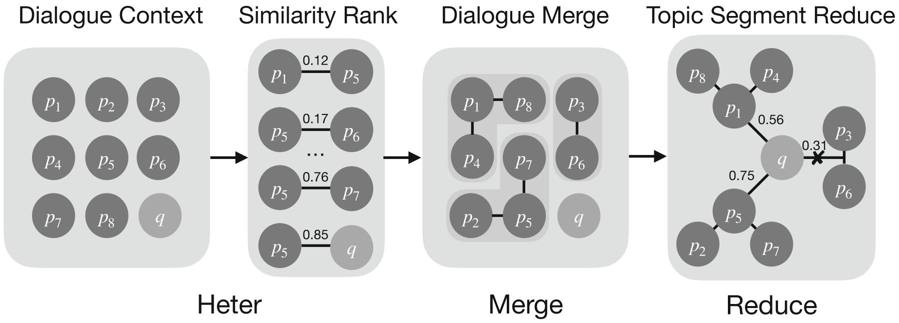

Fig. 2. The Heter-Merge-reduce method

dialogue segments, which can greatly avoid the personality bias brought by such preference discrepancy. In the Topic-dependent Context Deconstruction module, we obtain k topic-dependent segments $\{\tilde{H}_1, \cdots, \tilde{H}_k\}$ that are relevant to the current query q. We then perform matching between the current query q and the response candidate r under the context of each topic-dependent segment $\tilde{H}$ via the Personalized Topic Matching module. Formally, given the k topic-aware segments $\{\tilde{H}_1, \cdots, \tilde{H}_k\}$, the current query q and response candidate r, we seek to compute k topic-dependent matching features $\{\mathbf{e}_1, \cdots, \mathbf{e}_k\}$ which represent the topic-dependent relevance between the topic segments and the response candidate given the current query q. Taking the k-th topic segments $\tilde{H}_k = \{(p_{k,1}, r_{k,1}), \cdots, (p_{k,n_k}, r_{k,n_k})\}$ as an example, we first concatenate the topic-dependent segment with the current query q and the response candidate r into a token sequence S_k:

$$S_k = [\mathrm{CLS}], p_{k,1}, r_{k,1}[\mathrm{SEP}], \cdots, [\mathrm{SEP}], q, [\mathrm{SEP}], r \tag{6}$$

And we then feed the token sequence S_k into a pretrained encoder $\phi(\cdot)$ (*e.g.*, BERT) to get the token representations. We use the representation of the first token ([CLS]) as the sequence representation which is fed into a Multi-layer Perceptron (MLP) to obtain the k-th matching representations $\mathbf{e}_k$:

$$\mathbf{e}_k = \mathrm{MLP}_1(\mathrm{Pooling}_{\mathrm{CLS}}(\phi(S_k))), \tag{7}$$

where $\mathrm{MLP}_1 \in \mathbb{R}^{d \times d}$ and d is the hidden size.

Likewise, we perform the topic-dependent matching over each topic segments respectively and obtain k topic-dependent matching representations $\mathbf{E} = \{\mathbf{e}_1, \cdots, \mathbf{e}_k\}, \mathbf{E} \in \mathbb{R}^{k \times d}$. The k matching representations measure the matching degree between the response candidate r and the k topic-dependent segments given the current query q.

Furthermore, we think that the impact of the topic-aware segments is different as their topical relatedness to the current query is not the same. The topic-dependent segments with larger relevance scores to the current query should be more important. Thus, we perform self-attention to compute the relative importance of each segment by:

$$\mathbf{S} = \mathrm{softmax}(\frac{\mathbf{E} \cdot \mathbf{E}^\top}{\sqrt{d}}), \tag{8}$$

where $\mathbf{S} \in \mathbb{R}^{k \times k}$. We then computed the weighted topic segment matching scores by:

$$g^{\mathbf{T}} = \sum \underset{dim=-1}{\mathrm{mean}}(\mathbf{S}) \cdot \mathrm{MLP}_2(\mathbf{E}), \tag{9}$$

where $\mathrm{MLP}_2 \in \mathbb{R}^{d \times 1}$.

Besides the matching signal among topic-dependent segments, we also want to model the matching signal from the cross-topic user profile. Therefore, we compute the cross-topic matching representation by using a residual connection with an MLP to get a fused representation $\tilde{\mathbf{E}}$:

$$\tilde{\mathbf{E}} = \mathrm{MLP}_3(\hat{\mathbf{E}}) + \hat{\mathbf{E}}, \quad \hat{\mathbf{E}} = \mathbf{E} \cdot \mathbf{S} + \mathbf{E}, \tag{10}$$

where $\mathrm{MLP}_3 \in \mathbb{R}^{d \times d}$. We then pool the weighted matching representation $\tilde{\mathbf{E}}$ to obtain the cross-topic matching feature and feed it into a MLP to obtain the matching score of the cross-topic user profile.

$$g^{\mathbf{C}} = \mathrm{MLP}_2(\underset{dim=-1}{\mathrm{mean}}(\tilde{\mathbf{E}})). \tag{11}$$

We combine the two scores by:

$$g = \alpha \cdot g^{\mathbf{C}} + (1 - \alpha) \cdot g^{\mathbf{T}}, \tag{12}$$

where α is a trainable parameter and is initialized by 0.5.

We use cross-entropy loss to train the model:

$$\mathcal{L}(\theta) = -\frac{1}{|D|} \sum_{D} [y \log(g) + (1 - y) \log(1 - g)]. \tag{13}$$

4 Experiments

4.1 Dataset and Evaluation

We explore learning implicit user profiles from the user's dialogue history. Therefore, we require datasets with user identifications to construct users' dialogue history. We use two public datasets: Weibo and Reddit. Specifically, the Weibo dataset is derived from the PChatbotW dataset, in which all posts and responses have timestamps and user IDs [14]. The Reddit dataset is released by [21], which is crawled from the Reddit forum from Dec. 1, 2015, to Oct. 30, 2018. By traversing the chain-like responses, we can obtain post-response pairs with timestamps and user IDs. We first aggregate the user's dialogue history for the two datasets and then filter out users who have less than fifteen historical dialogues. Besides, we limit the length of all utterances by 50 tokens. Following previous works [10,13], we use the latest post as the current query and create a list of ten response candidates in which the negative samples are mined via a BM25 engine. The candidate list contains: (1) the ground-truth response made by the user; (2) other user's responses under the same post (non-personalized response); (3) retrieved response candidates via a retrieval engine (hard negative samples). The statistic information of the two datasets is shown in Table 1.

Table 1. The statistics of the two datasets.

	Weibo	Reddit
Number of users	420,000	280,642
Average history length	32.3	85.4
Average length of post	24.9	10.5
Average length of response	10.1	12.4
Number of response candidates	10	10
Number of training samples	3,000,000	2,000,000
Number of validation samples	600,000	403,210
Number of testing samples	600,000	403,210

To evaluate our proposed TopReC and all baseline models, we use $\mathbf{R_n@k}$ (recall at position k in n candidates) and **MRR** (Mean Reciprocal Rank) as evaluation metrics. As the ground-truth response is the personalized response, the two metrics can directly evaluate the model's ability to output a response that is consistent with the user's personality.

4.2 Baseline Models

In the task, the user's dialogue history comprises many single-turn dialogues. Besides, the dialogue history can be considered as the multi-turn context. Hence, except for the two types of personalized baseline models, we also consider the single-turn and multi-turn models as the baseline: (1) Single-turn models: **Conv-KNRM** [1]: The model utilizes a kernel-based ranking method with CNN to learn soft n-gram matches for ad-hoc matching; **BERT-adhoc** [2]: We fine-tune the BERT model with single-turn dialogue data. (2) Multi-turn models: **DAM** [24]: The model stacks multiple attentive modules to extract deep semantic interactive semantics. **IOI** [16]: The model designs a chain of deep interactive blocks to perform semantic interactions. **MSN** [18]: The model filters irrelevant dialogue context and performs matching at multi-grained. (3) Explicit user profile-based models: **DIM** [4]: The model separately encodes the context, user profile, and response candidates and then performs interactions; **RSM-DCK** [7]: The model performs context selection over dialogue context and then perform response selection; **CSN** [25]: The model uses a content selection network to select relevant dialogue context and then perform matching; (4) Implicit user profile-based models: **IMPChat** [13]: The model proposes learning implicit user profile from the user's dialogue history. **BERT** [2]: We fine-tune the BERT model with all users' dialogue history.

Table 2. Evaluation results of all models on both Weibo and Reddit corpus. "†" denote the TopReC is significantly better than all baselines in t-test with $p < 0.05$ level. The best results are in bold.

	Weibo Corpus				Reddit Corpus			
	R$_{10}$@1	**R$_{10}$@2**	**R$_{10}$@5**	**MRR**	**R$_{10}$@1**	**R$_{10}$@2**	**R$_{10}$@5**	**MRR**
(1) Conv-KNRM	0.323	0.520	0.893	0.538	0.576	0.711	0.917	0.712
(1) BERT (adhoc)	0.342	0.545	0.966	0.561	0.668	0.797	0.991	0.787
(2) DAM	0.438	0.644	0.966	0.635	0.605	0.748	0.965	0.741
(2) IOI	0.442	0.651	0.969	0.639	0.620	0.764	0.974	0.753
(2) MSN	0.355	0.554	0.931	0.567	0.555	0.733	0.977	0.715
(3) DIM	0.388	0.557	0.835	0.571	0.678	0.813	0.979	0.794
(3) RSM-DCK	0.428	0.627	0.947	0.623	0.615	0.753	0.972	0.748
(3) CSN	0.387	0.560	0.842	0.572	0.681	0.807	0.976	0.794
(4) IMPChat	0.460	0.665	0.963	0.651	0.691	0.820	0.982	0.804
(4) BERT	0.445	0.653	0.967	0.641	0.727	0.849	0.991	0.830
(4) TopReC	**0.486**†	**0.695**†	**0.972**†	**0.677**†	**0.750**†	**0.868**†	**0.992**	**0.852**†

4.3 Implementation Details

We employ the *bert-base-uncased* and *chinese-bert-wwm-ext* as the backbone of TopReC for the Reddit and Weibo datasets, respectively. The codes are implemented based on the PyTorch-Lightning[1] and Transformers[2] libraries. We train the TopReC on 4 T V100 16GB GPUs for 3 epochs. We set the batch size as 128, and the learning rate as 1e-5. For the number of topic segments k and the sequence length l of each topic segment, we set $k = 3, l = 256$, and $k = 4, l = 128$ for the Weibo and Reddit dataset, respectively. The reason that we keep a longer sequence length for the Weibo dataset is that dialogues on Weibo are usually longer than on Reddit (see Table 1). The further analysis of the choice of k and l can refer to Sect. 4.5. We use the history length of 15 for all baseline models and TopReC. The detailed analysis of the choice of history length can refer to Sect. 4.5. We tune TopReC and all baseline models on the dev set and evaluate the models on the test set. The codes will be released at https://github.com/qhjqhj00/ECIR23-TopReC.

[1] https://github.com/PyTorchLightning/pytorch-lightning.
[2] https://github.com/huggingface/transformers.

4.4 Experimental Results

Table 2 shows the experiment results from which we have the following findings: **First**, **the proposed TopReC outperforms all baseline models regarding all evaluation metrics.** And TopReC lead statistically significant improvement regarding all metrics on the Weibo dataset and most metrics on the Reddit dataset (t-test with $p < 0.05$). It proves the effectiveness of TopReC's ability to find the most proper response that is consistent with the user's personality. **Second**, all models perform worse in the Weibo dataset than the Reddit dataset, which implies that the dialogue history in the Weibo dataset might contain more noise than the Reddit dataset. Impacted by such noise, in the Weibo dataset, the pre-trained model BERT performs worse than the IMPChat, which does not benefit from the pre-trained language model. The reason might be that IMPChat conduct reweighs the importance of the historical dialogues, which alleviate the impact of data noise. Compared to BERT, TopReC models the user's personality concerning the topical inter-relations inside the dialogue history and prunes the topically irrelevant dialogue history. As a result, TopReC can be partially immune to the negative effect of the data noise and therefore booster the performances; **Third**, regarding the model types, we find that the models learning implicit user profile from the dialogue history perform better than the rest types of models. It demonstrates the superior effectiveness of learning implicit user profiles from the dialogue history. A fundamental problem of learning implicit user profiles is how to use the dialogue history properly. In this paper, TopReC uses the dialogue history from a topic-aware perspective which is empirically effective. Future works might explore more promising perspectives to use the dialogue history and provide better performances and explainability.

4.5 Discussion

Ablation Study. To verify the effect of the Topic-dependent Context Deconstruction module, we randomly deconstruct the dialogue history into the same number of dialogue segments for comparison. To study the effect of the Personalized Topic Matching module, we respectively remove the topic-dependent matching scores and the cross-topic matching score. Table 3 shows the results. We find: (1) removing any module of TopReC would bring performance decline, implying that any module of TopReC captures orthogonal information that is indispensable to the overall model; (2) randomly deconstructing the dialogue history lead to big performance decline, verifying the validity of our idea that models the user personality from topic-dependent dialogue segments; (3) removing any of the topic-dependent matching scores and the cross-topic matching score would lead to performance decline, which implies that the two matching scores capture the personalized information from different perspectives (*e.g.*, local and global).

Impact of History Length. We conduct experiments with our TopReC and previous SOTA model IMPChat to study the impact of history length. Figure 3 shows the results: (1)the model performances show an increasing tendency with

Table 3. Ablation results on the Reddit dataset.

	$\mathbf{R_{10}@1}$	$\mathbf{R_{10}@2}$	$\mathbf{R_{10}@5}$	**MRR**
TopReC	**0.750**	**0.868**	**0.992**	**0.852**
w/o cross	0.739	0.858	0.991	0.838
w/o topic	0.736	0.857	0.990	0.836
Random segment	0.732	0.852	0.986	0.833
BERT	0.727	0.849	0.991	0.830

longer dialogue history, which indicates that longer dialogue history can provide more personalized information; (2) our TopReC outperforms IMPChat after the history length of 15, which proves that TopReC is more effective when modeling user personality from long dialogue history. Before the history length of 15, TopReC is more sensitive to data insufficiency than IMPChat, as the latter is designed to learn multi-grained user representations from the whole dialogue history. Such saturated fitting is effective for short dialogue history but also limits the model capacity for longer dialogue history; (3) the increasing tendency slow down after the history length of 15, the reason might be that the experiments setting[3] limits TopReC's capacity for longer dialogue history, which indicates that longer dialogue history contains more dialogue topics, and correspondingly, we should increase the choice of the max number of topics. Figure 3 middle shows the impact of the max number of topics (the k value in Sect. 3.3). The model performance peaks at $k = 4$ and then decreases, verifying the effectiveness of using topically-related dialogue segments as context. And it also proves that less relevant topic segments might undermine the model performance.

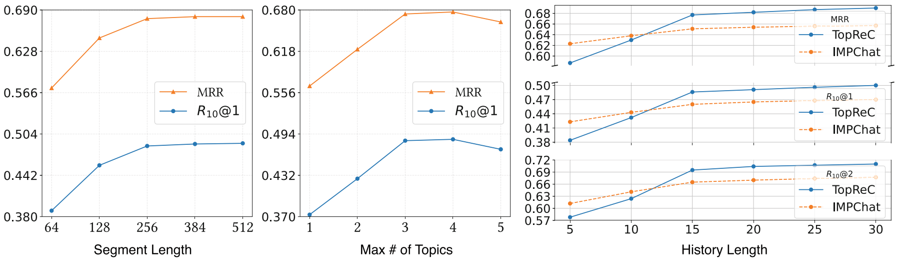

Fig. 3. Left shows the impact of segment length, middle shows the impact of max number of topics to keep, and right shows the impact of history length.

[3] For TopReC, we set the max segment length as 256 and the max number of topics as 4. For IMPChat, we feed all dialogue history into the model without truncation.

Impact of Sequence Length. Figure 3 left shows the impact of the length of topic segment (the l value in Sect. 4.3). We find that the model performance steadily increase with longer segment length. The reason is that less context would be truncated when using longer segment length. But in the meantime, the required computing resources greatly increase with longer segment length (*e.g.*, BERT's complexity exponentially increases with the sequence length), for which we choose to use relatively small l value[4] in this paper.

5 Conclusion

This paper explores learning implicit user profiles from dialogue history for a personalized chatbot. We observe that a user's dialogue history might be long and noisy as the dialogue history contains the user's long-term, multi-topic dialogue records. To reduce the data noise and increase the model's capacity to adapt long dialogue history, we propose deconstructing the user's dialogue history into topic-dependent segments and filtering out irrelevant dialogue segments. We design a model TopReC, which first performs dialogue history deconstructions via a Heter-Merge-Reduce method and learns the topic-aware personality from each topic-dependent segment. Besides, TopReC also explores a cross-topic personalized matching feature that measures the matching degree of the response candidate from a general perspective. The final response is selected by fusing the topic-dependent and cross-topic matching scores. Experimental results verify the effectiveness of the proposed TopReC. The limitations of this work are: (1) we conduct experiments on datasets that come from social media, which might not reflect how people usually talk; (2) we prune noisy topical segments by measuring the similarities, which might be biased by data noise and therefore lack interpretability. In the future, we will further explore how TopReC performs in more dialogue datasets and how to better learn an implicit user profile from the dialogue history to enhance personalized response selection regarding both effectiveness and interpretability.

Acknowledgments. Zhicheng Dou is the corresponding author. This work was supported by the National Natural Science Foundation of China No. 62272467 and No. 61872370, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, the Fundamental Research Funds for the Central Universities, the Research Funds of Renmin University of China NO. 22XNKJ34, Public Computing Cloud, Renmin University of China, Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China, and the Outstanding Innovative Talents Cultivation Funded Programs 2023 of Renmin Univertity of China. The work was partially done at Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education, and Beijing Key Laboratory of Big Data Management and Analysis Methods.

[4] $l = 128$ and $l = 256$ for Reddit and Weibo.

References

1. Dai, Z., Xiong, C., Callan, J., Liu, Z.: Convolutional neural networks for soft-matching n-grams in ad-hoc search. In: Proceedings of the 11th WSDM. ACM (2018). https://doi.org/10.1145/3159652.3159659
2. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
3. Gao, J., Galley, M., Li, L.: Neural Approaches to Conversational AI: Question Answering. Task-oriented Dialogues and Social Chatbots. Now Foundations and Trends (2019)
4. Gu, J.C., Ling, Z.H., Zhu, X., Liu, Q.: Dually interactive matching network for personalized response selection in retrieval-based chatbots. In: Proceedings of the EMNLP-IJCNLP 2019 (2019)
5. Gu, J.C., Liu, H., Ling, Z.H., Liu, Q., Chen, Z., Zhu, X.: Partner matters! an empirical study on fusing personas for personalized response selection in retrieval-based chatbots. In: Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 565–574 (2021)
6. Han, J., Hong, T., Kim, B., Ko, Y., Seo, J.: Fine-grained post-training for improving retrieval-based dialogue systems. In: Proceedings of NAACL 2021. ACM (2021). https://www.aclweb.org/anthology/2021.naacl-main.122
7. Hua, K., Feng, Z., Tao, C., Yan, R., Zhang, L.: Learning to detect relevant contexts and knowledge for response selection in retrieval-based dialogue systems. In: Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pp. 525–534 (2020)
8. Li, J., Galley, M., Brockett, C., Spithourakis, G.P., Gao, J., Dolan, W.B.: A persona-based neural conversation model. In: Proceedings of the ACL 2016. ACL (2016)
9. Liu, Y., Qian, H., Xu, H., Wei, J.: Speaker or listener? the role of a dialog agent. In: Findings of the Association for Computational Linguistics: EMNLP 2020. pp. 4861–4869. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.findings-emnlp.437, https://aclanthology.org/2020.findings-emnlp.437
10. Lowe, R., Pow, N., Serban, I.V., Pineau, J.: The ubuntu dialogue corpus: a large dataset for research in unstructured multi-turn dialogue systems. In: SIGDIAL 2015, pp. 285–294 (2015). https://doi.org/10.18653/v1/w15-4640
11. Ma, Z., Dou, Z., Zhu, Y., Zhong, H., Wen, J.R.: One chatbot per person: creating personalized chatbots based on implicit user profiles. In: SIGIR 2021 (2021)
12. Mao, K., Dou, Z., Qian, H.: Curriculum contrastive context denoising for few-shot conversational dense retrieval. In: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2022, , pp. 176–186. Association for Computing Machinery, New York (2022). https://doi.org/10.1145/3477495.3531961
13. Qian, H., Dou, Z., Zhu, Y., Ma, Y., Wen, J.R.: Learning implicit user profile for personalized retrieval-based chatbot. In: Proceedings of the 30th CIKM, pp. 1467–1477 (2021)
14. Qian, H., et al.: Pchatbot: a large-scale dataset for personalized chatbot. In: Proceedings of the 44th SIGIR. ACM, Virtual Event (2021). https://doi.org/10.1145/3404835.3463239, https://doi.org/10.1145/3404835.3463239

15. Qian, Q., Huang, M., Zhao, H., Xu, J., Zhu, X.: Assigning personality/profile to a chatting machine for coherent conversation generation. In: IJCAI, pp. 4279–4285 (2018)
16. Tao, C., et al.: 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 ACL, pp. 1–11 (2019). https://doi.org/10.18653/v1/P19-1001, https://www.aclweb.org/anthology/P19-1001
17. Xu, Y., Zhao, H., Zhang, Z.: Topicaware multi-turn dialogue modeling. In: The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-2021) (2021)
18. Yuan, C., et al.: Multi-hop selector network for multi-turn response selection in retrieval-based chatbots. In: Proceedings of EMNLP-IJCNLP 19. ACL (2019). https://doi.org/10.18653/v1/D19-1011, https://www.aclweb.org/anthology/D19-1011
19. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: i have a dog, do you have pets too? In: Proceedings of the ACL 2018, pp. 2204–2213. ACL (2018)
20. Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., Weston, J.: Personalizing dialogue agents: i have a dog, do you have pets too? In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, vol. 1: Long Papers, pp. 2204–2213. Association for Computational Linguistics, Melbourne (2018). https://doi.org/10.18653/v1/P18-1205, https://aclanthology.org/P18-1205
21. Zhang, Y., Sun, S., Galley, M., Chen, Y., Brockett, C., et al.: DIALOGPT: large-scale generative pre-training for conversational response generation. In: Proceedings of the ACL 2020 (2020)
22. Zhao, X., Tao, C., Wu, W., Xu, C., Zhao, D., Yan, R.: A document-grounded matching network for response selection in retrieval-based chatbots (2019). https://doi.org/10.48550/ARXIV.1906.04362, https://arxiv.org/abs/1906.04362
23. Zhong, H., Dou, Z., Zhu, Y., Qian, H., Wen, J.R.: Less is more: learning to refine dialogue history for personalized dialogue generation. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 5808–5820. Association for Computational Linguistics, Seattle (2022). https://doi.org/10.18653/v1/2022.naacl-main.426, https://aclanthology.org/2022.naacl-main.426
24. Zhou, X., et al.: Multi-turn response selection for chatbots with deep attention matching network. In: Proceedings of the 56th ACL, pp. 1118–1127 (2018). https://doi.org/10.18653/v1/P18-1103, https://www.aclweb.org/anthology/P18-1103
25. Zhu, Y., Nie, J.-Y., Zhou, K., Du, P., Dou, Z.: Content selection network for document-grounded retrieval-based chatbots. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12656, pp. 755–769. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72113-8_50