Less is More: Learning to Refine Dialogue History for Personalized Dialogue Generation

Hanxun Zhong\(^1\), Zhicheng Dou\(^{2,*}\), Yutao Zhu\(^3\), Hongjin Qian\(^2\), Ji-Rong Wen\(^2\)

\(^1\)School of Information, Renmin University of China, Beijing, China
\(^2\)Gaoling School of Artificial Intelligence, Renmin University of China, Beijing, China
\(^3\)University of Montreal, Quebec, Canada

{hanxun_zhong,dou,ian}@ruc.edu.cn
{yutaozhu94,jirong.wen}@gmail.com

Abstract

Personalized dialogue systems explore the problem of generating responses that are consistent with the user’s personality, which has raised much attention in recent years. Existing personalized dialogue systems have tried to extract user profiles from dialogue history to guide personalized response generation. Since the dialogue history is usually long and noisy, most existing methods truncate the dialogue history to model the user’s personality. Such methods can generate some personalized responses, but a large part of dialogue history is wasted, leading to sub-optimal performance of personalized response generation. In this work, we propose to refine the user dialogue history on a large scale, based on which we can handle more dialogue history and obtain more abundant and accurate persona information. Specifically, we design an MSP model which consists of three personal information refiners and a personalized response generator. With these multi-level refiners, we can sparsely extract the most valuable information (tokens) from the dialogue history and leverage other similar users’ data to enhance personalization. Experimental results on two real-world datasets demonstrate the superiority of our model in generating more informative and personalized responses.

1 Introduction

Recent years have witnessed great progress in building personalized dialogue systems. In general, previous work explores building a personalized dialogue system via two pathways: (1) directly modeling user personality from predefined persona descriptions or user attribute (Qian et al., 2018; Zhang et al., 2018; Song et al., 2019); and (2) implicitly modeling the user personality from the user’s dialogue history (Li et al., 2016c; Ma et al., 2021). The latter is considered superior as the dialogue history is easy to obtain and comprises rich personalized information. In this paper, we follow the second pathway that automatically learns implicit user profiles from the user’s dialogue history to assist in personalized response generation.

It is challenging to model user personality directly from the dialogue history. The main reason is that a user’s dialogue history may contain massive historical dialogues, which are too heavy to load in the model and likely to be noisy. A straightforward solution is to truncate the dialogue history, as done by existing work (Ma et al., 2021; Qian et al., 2021a). However, as tremendous information has been wasted, the model’s performance is also influenced. On the other hand, we observe that the dialogue history from other users may also be helpful in generating a personalized response for the current user. For example, users with the same interest in “soccer” may talk about similar things on such a topic. This has been overlooked by existing methods. Intuitively, the problem of “data explosion” is even more severe in the latter case (when other similar users’ dialogue history is also considered). To alleviate these problems, we propose using a hierarchical refiner structure to sparsely extract the most valuable query-aware persona information from both the current and other similar users’ dialogue history. By this means, more dialogue history can be utilized for learning user personality and improving response generation.

Our model is called MSP, which stands for Modeling and Selecting user Personality from the dialogue history for generating personalized responses. Instead of attending to all dialogue history, MSP refines the most valuable historical information that can well portray the user’s personality and guide the response generation. Specifically, MSP consists of three personal information refiners working at different levels and a personalized response generator. At first, a user refiner is designed to select a group of users who have similar interests to the current user. By refining dialogue history at the user level, we can obtain similar data.
to share information with similar users and avoid mutual interference with other users. Then, a topic refiner filters out the current and similar users’ dialogue history that has different topics with the current query at the sentence level. Next, we design a token refiner to extract the most valuable query-aware user profiles from the remaining dialogue history at the token level. Finally, a personalized response generator combines user profiles and the current query to generate responses. Given that there is no explicit supervisory signal guiding the refiner to extract an exemplary user profile, we design a supplementary sentence matching task and a joint training method. The generator will construct a pseudo-label to guide the refiner’s extraction.

Our contributions are three-fold:

1. We design an MSP model to tackle the data noise problem. It can efficiently refine user profiles through dialogue history and generate personalized responses. By this means, our method can capture abundant user profiles while keeping away from noisy data.

2. We design a refiner structure to extract the query-aware profile at three levels. Similar users’ information is taken into account, which can help improve the personality of the response.

3. We design a joint training method for the refiner and generator. The refiner provides the generator with user profiles to assist in generating responses, while the generator constructs a pseudo-label for the refiner to assist in selecting user profiles.

2 Related Work

Personalized Dialogue Generation Open-domain dialogue generation has been extensively studied (Koehn et al., 2003; Vinyals and Le, 2015; Serban et al., 2016; Zhang et al., 2019a,b; Liu et al., 2020; Xiao et al., 2020; Zhu et al., 2020). Recently, personalized dialogue systems have attached more and more attention. Typical methods include: (1) explicitly using predefined persona descriptions or attributes as users’ profile to generate personalized responses (Qian et al., 2018; Zhang et al., 2018; Olabiyi et al., 2019; Song et al., 2019); (2) using user ID embeddings to enhance personalized dialogue generation (Li et al., 2016c; Chan et al., 2019); and (3) extracting implicit user profile from users’ dialogue history to generate personalized responses (Al-Rfou et al., 2016; Bak and Oh, 2019; Ma et al., 2021). Since manually collecting user profiles is impractical for large-scale datasets and the user ID embeddings perform badly, in this study, we focus on the last group of methods for personalized response generation.

DHAP (Ma et al., 2021) is the state-of-the-art method in personalized dialogue generation. It uses a transformer-based structure to model the user’s dialogue history and extract personal information for response generation. Unfortunately, this model can only handle a limited number of user dialogue histories, wasting a lot of valuable information. Our method has two main differences from DHAP: (1) We propose a refiner structure in our model so that more dialogue history can be handled and the most valuable information can be extracted for improving response generation; (2) With our proposed refiner, we can further incorporate more dialogue history from other users (having similar interests) to facilitate personalized dialogue generation.

Retrieval-guided Natural Language Generation Retrieval-based methods can collect relevant information for language generation (Yang et al., 2019). It has been widely applied in many tasks such as text style transfer (Li et al., 2018) and dialogue generation (Wu et al., 2019; Cai et al., 2019). The idea of using a retrieval system to get useful information inspires our study. We use a refiner to automatically extract personal information from dialogue history and guide the personalized generation.

3 Methodology

In this section, we first formalize our problem and provide an overview of our proposed MSP model. Then, we describe the details of each component and the model optimization.

3.1 Problem Statement and Overview

Considering a set of users \( U = \{u_1, \ldots, u_l\} \), for any user \( u_i \), we have their dialogue history with others \( U^i = \{ (q^i_1, r^i_1), \ldots, (q^i_l, r^i_l) \} \), where \( q_j^i \) is a query issued by others, while \( r_j^i \) is the corresponding response given by \( u_i \).\(^1\) Our target is to generate a personalized response \( r^i \) for the user \( u_i \) to reply a new query \( q \). As we introduced earlier, the personalized information can be obtained from the dialogue history \( U^i \) of the user \( u_i \) and other dialogue history \( U^j \) from similar users \( u_j (j \neq i) \).

\(^1\)Here we use the term “query” to denote the utterance given by others. Generally, the query can be either one utterance in single-turn dialogues, or several history utterances in multi-turn dialogues.
The overview of our MSP model is shown in Figure 1. MSP consists of three personal information refiners working on different levels and a personalized response generator. Specifically, the first refiner is working at the user level. By comparing the dialogue history of the current user $u_i$ with others, MSP can select a group of users having similar interests with $u_i$. After obtaining a group of similar users, we further refine their dialogue history according to the relevance to the current query’s topic. Moreover, we add the last refiner to extract several tokens so that the most fine-grained personal information can be extracted from the relevant utterances. Finally, the query and the extracted tokens are fed into the generator and construct a personalized response.

3.2 User Refiner

The dialogue history of users with similar interests may share much personal information. Therefore, our first target is to select a group of users with similar interests to the current user. We design a user refiner to achieve this. Since the users’ interest is usually contained in their dialogues with others, we consider both the queries and responses in the dialogue history to select similar users. Specifically, for the user $u_i$’s dialogue history $U_i$, we apply a pre-trained BERT (Devlin et al., 2019) and represent them by the embedding of [CLS] token:

$$U_q^i = \sum_{j=1}^{t} \text{BERT}(q_j^i), \quad U_r^i = \sum_{j=1}^{t} \text{BERT}(r_j^i).$$

Then, we can select $k_u$ users that have similar interest with the current user $u_i$ as:

$$a^\text{sim} = \text{TopK}(U_q^i \cdot U_{j_u}^j),$$

$$U^i = [U_q^i; U_r^i],$$

where TopK($\cdot$, $\cdot$) is the top-$k$ selection operation.

After the user refiner, we can obtain the dialogue history of the similar users $\{u_j\}_{j=1}^{k_u}$. It is worth noting that, since the number of users is large in the datasets, we choose to use the dot-product to compute the similarity of the users so that the whole process can be implemented by dense retrieval libraries, such as Faiss (Johnson et al., 2021), which is very efficient.

3.3 Topic Refiner

The users’ dialogue history often contains many dialogues with others. These dialogues have various topics, which may be irrelevant to the current query. Therefore, we propose a topic refiner to select relevant dialogue history for personalized response generation. Specifically, we use a topic classifier to compute the topic distribution of the current query.
where \( t, t_j^i \in \mathbb{R}^{d_t \times 1} \), and \( d_t \) is the number of topic. Then, we filter out the dialogue history \(< q_j^i, r_j^i >\) that has different topics with the current query, i.e., \( \max(t_j^i) \neq \max(t) \).

In the topic refining process, we compare the queries in the history dialogues with the current query to filter out topic-irrelevant dialogues. This process can further reduce the noise and make our model more lightweight. Both the dialogue history of the current user and that of the similar users (obtained in the former step) are refined. In the next step, we will use the responses in these selected history dialogues and extract the most valuable tokens for personalized response generation.

### 3.4 Token Refiner

After the previous two refiners, we obtain a collection of historical responses. Though we can directly add them into the generation process, our preliminary experiments indicate that they perform poorly. A major reason is the noisy quality of the responses. Indeed, existing studies (Xing et al., 2017; Zhu et al., 2020) have demonstrated the effectiveness of using informative tokens to improve the response generation. Inspired by these studies, we further devise a token refiner to extract the most fine-grained information (tokens) from the historical responses. Specifically, we compute an attention map \( A \) between the query \( q \) and the historical responses \( r^{\text{sim}} \) and \( r^{\text{cur}} \) (they are from the similar users and the current user respectively) as:

\[
A = \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right),
\]

\[
Q = \text{TRM}_{\text{enc}}(q) \cdot W_Q,
\]

\[
K = \text{TRM}_{\text{enc}}(r) \cdot W_K,
\]

where \( \text{TRM}_{\text{enc}}(\cdot) \) is a transformer encoder. \( r \) refers to \( r^{\text{sim}} \) or \( r^{\text{cur}} \), and correspondingly, \( A \) refers to the similar user matching map \( A^{\text{sim}} \) and current user matching map \( A^{\text{cur}} \). \( W_Q, W_K \in \mathbb{R}^{d \times d} \) are parameters, and \( d \) is the dimension of the hidden state. After obtaining the attention matching map \( A \), we select tokens to form the similar users’ profile and current user’s profile according to each token’s attention weight:

\[
c^{\text{sim}} = \text{TopK} \left( \text{Max} \left( A^{\text{sim}}, k_p \right) \right),
\]

\[
c^{\text{cur}} = \text{TopK} \left( \text{Max} \left( A^{\text{cur}}, k_p \right) \right),
\]

where \( k_p \) is a hyper-parameter to control the number of profile tokens. Its influence will be investigated in our experiments.

### 3.5 Generator

We use a transformer decoder as to generate a personalized response by using the similar users’ profile \( c^{\text{sim}} \), current user’s profile \( c^{\text{cur}} \), and query information \( q \) as input. The decoding process can be defined as:

\[
y = \text{TRM}_{\text{dec}}(x),
\]

\[
x = [c^{\text{sim}}, c^{\text{cur}}, q],
\]

where \([;] \) is the concatenation operation and \( y \) is the word generation probability.

### 3.6 Training and Optimization

The generator is optimized by maximizing the generation probability of the ground-truth \( y \):

\[
L^g = -y \log \hat{y}.
\]

In practice, we find that the token refiner is hard to train. We speculate the reason is a missing of direct supervision signals. In this case, it is difficult to tell whether the training errors stem from the generation process or the refining process. To tackle this problem, we propose a supplementary sentence matching task to assist the token selection.

#### Supplementary Sentence Matching Task

The core idea of this task is to train the token refiner directly by introducing supervision signals so that it can automatically mine valuable tokens. Specifically, we design a sentence matching task to match the query with the dialogue history. The task’s target is to find the history sentences that help generate personalized responses. We consider using the query-history cross attention weight \( A \) to generate a matching representation and then use this representation to finish the task. In this way, once the matching task has been well-finished, we can use attention map \( A \) to identify the most informative and valuable parts of a history sentence that are helpful to generate a personalized response.

To achieve our idea, we design a matching process. First, we calculate the matching representations \( H \) by the cross-attention map \( A \) and then
apply a CNN with a max-pooling operation to aggregate token information:

\[ S = \text{Maxpool}(\text{CNN}(H)) \]  
(13)

\[ H = A \cdot V, \]  
(14)

\[ V = \text{TRM}_{\text{enc}}(r) \cdot W_V. \]  
(15)

Next, we flatten \( S \) and applies a LSTM to aggregate the sentence information:

\[ h = \text{LSTM}(\text{Flatten}(S)). \]  
(16)

Finally, we use the sentence matching vector \( h \) to compute the matching score:

\[ \hat{g} = \text{Sigmoid}(\text{MLP}(h)). \]  
(17)

For guiding the sentence matching task, we design a pseudo-label \( g \) to measure the matching goodness of each history sentence. We expect the history with more persona profile information can achieve a higher score. Therefore, we use the difference between the personalized ground-truth \( y \) and a non-personalized generated response \( \hat{y} \) to measure the persona profile and create the pseudo-label:

\[ g = \begin{cases} 1, & g_{\text{soft}} \geq \alpha \\ 0, & g_{\text{soft}} < \alpha, \end{cases} \]  
(18)

\[ g_{\text{soft}} = \text{Sum}(\text{Max}(y - \hat{y}) \cdot r))/d_y, \]  
(19)

where \( d_y \) is the length of \( y \), and \( \alpha \) is a threshold. Finally, we minimize the binary cross entropy loss between \( g \) and \( \hat{g} \):

\[ \mathcal{L}^g = g \log \hat{g} + (1 - g) \log(1 - \hat{g}). \]  
(20)

**Joint Training** To facilitate the learning with the above gradient approximation approach, we design a joint training process to train the refiner and generator in turn. Specifically, in each training iteration, we first sample a batch of query \( q \), response \( y \), similar and current users’ dialogue history \( r^{\text{sim}} \) and \( r^{\text{cur}} \) from dataset \( D \). Then, we generate a non-personalized response \( \hat{y} \) and create the pseudo-label \( g \) (Eq. 18) through a non-personalized generator. This pseudo-label is used to train the token refiner by optimizing the loss \( \mathcal{L}^g \) (Eq. 20). Further, we sample another batch \( D^p \) from \( D \). After extracting similar user profile \( e^{\text{sim}} \) and current user profile \( e^{\text{cur}} \) (Eq. 8, Eq. 9), we generate the personalized response \( \hat{y} \) and update the generator by optimizing the loss \( \mathcal{L}^g \) (Eq. 12). To avoid misleading the generation (by poor profile) at the beginning of the training process, we pre-train the refiner for \( N_f \) steps before extracting the profile for the generator. The detailed training process is summarized in Algorithm 1.

---

**Algorithm 1 Joint Training Process**

**Input:** \( M \) dialogue triplets: \( D = \{ \langle q_i, y_i, r_i \rangle \}_{i=1}^N \)

**Output:** A personalized dialogue model

1: Init the refiner and generator module
2: while not converge do
3: Sample \( n_u \) dialogue triplets \( D^{u} = \{ \langle q_i, y_i, r_i \rangle \}_{i=1}^{n_u} \)
4: Get \( \hat{y} = \{ \hat{y}_i \}_{i=1}^{n_u} \) on \( D^{u} \) from \( p_{\hat{g}}(\hat{y} | q) \)
5: Get pseudo-label \( g \) on \( D^{u} \)
6: Train refiner by optimizing \( \mathcal{L}^g \) on \( D^p \)
7: if Current Step > \( N_f \) then
8: Sample \( n_v \) dialogue triplets \( D^{v} = \{ \langle q_i, y_i, r_i \rangle \}_{i=1}^{n_v} \)
9: Extract \( C^{v} = \{ c_i \}_{i=1}^{n_v} \) on \( D^{v} \) from \( p_{\hat{g}}(c | q, r) \)
10: Train generator by optimizing \( \mathcal{L}^g \) on \( D^{u} \cup C^{v} \)
11: end if
12: end while

---

**Table 1:** Statistics of Reddit and Weibo datasets.

<table>
<thead>
<tr>
<th></th>
<th>Reddit</th>
<th>Weibo</th>
</tr>
</thead>
<tbody>
<tr>
<td># Users</td>
<td>78,031</td>
<td>46,973</td>
</tr>
<tr>
<td>Avg. history length</td>
<td>72.4</td>
<td>30.8</td>
</tr>
<tr>
<td>Avg. # words of query</td>
<td>19.8</td>
<td>22.9</td>
</tr>
<tr>
<td>Avg. # words of response</td>
<td>9.1</td>
<td>9.6</td>
</tr>
<tr>
<td># Training samples</td>
<td>5,734,129</td>
<td>1,495,149</td>
</tr>
<tr>
<td># Validation samples</td>
<td>10,000</td>
<td>10,000</td>
</tr>
<tr>
<td># Testing samples</td>
<td>10,000</td>
<td>10,000</td>
</tr>
</tbody>
</table>

---

**4 Experiments**

**4.1 Datasets**

To evaluate our model’s performance, we conduct experiments on a Chinese Weibo dataset (Qian et al., 2021b) and an English Reddit dataset. Both are collected from open-domain social media platforms. On these platforms, users can post various topics, and other users can respond to them. We compare user-id and timestamps to associate the query with its corresponding response and the current user’s dialogue history. As a result, each training sample contains a query, a response, and a sequence of dialogue history. Finally, the dataset is divided into training, validation, and test sets in chronological order. The statistics of the datasets are provided in Table 1.

**4.2 Baseline Methods**

We compare our proposed model with eight highly correlated and strong baselines. They can be categorized into four groups:

**Non-personalized Methods**

1. Seq2Seq-Attention (Sutskever et al., 2014) is a vanilla sequence-to-sequence model with attention mechanism (Luong et al., 2015).
2. MMI (Li et al., 2015).
3. LSTM (Hochreiter and Schmidhuber, 1997).
4. Attention (Vaswani et al., 2017).

---

Table 2: Criteria of human annotation.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Rating</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Readability</strong></td>
<td>3: Fluent and easy to read</td>
</tr>
<tr>
<td></td>
<td>2: Grammatically formed</td>
</tr>
<tr>
<td></td>
<td>1: Not a complete sentence or hard to read</td>
</tr>
<tr>
<td><strong>Informativeness</strong></td>
<td>3: Have clear and specific meaning</td>
</tr>
<tr>
<td></td>
<td>2: Contain few informative words</td>
</tr>
<tr>
<td></td>
<td>1: Meaningfulness sentence</td>
</tr>
<tr>
<td><strong>Personalization</strong></td>
<td>1: Reflect personal information contained in user history</td>
</tr>
<tr>
<td></td>
<td>0: Does not resemble any user history</td>
</tr>
</tbody>
</table>

2016a) is based on Seq2seq and use maximum mutual information as an extra loss to improve diversity. (3) DialoGPT (Zhang et al., 2019b) is a variant of GPT-2 (Radford et al., 2019) designed for dialogue generation.

**Predefined Profile-based Methods** Since there is no persona description in the datasets, we test these methods by using the user’s dialogue history as a simulation of predefined persona profiles. (4) GPMN (Zhang et al., 2018) enhances the Seq2seq model with a memory module, which encodes and stores the persona profile as memory representations. (5) PerCVAE (Zhao et al., 2017) encodes predefined personalized sentences as a conditional representation and uses CVAE to generate a personalized response.

**User ID-based Methods** (6) Speaker (Li et al., 2016c) is based on seq2seq while using user-ID embedding as user representation to facilitate the response generation. (7)Persona WAE (Chan et al., 2019) uses WAE (Wasserstein autoencoder) for response generation. It maps user-ID embeddings to a personalized Gaussian mixture distribution and then samples the personalized vector to guide the response generation.

**User Dialogue History-based Methods** (8) DHAP (Ma et al., 2021) uses history memory to store and construct the dynamic query-aware user profile from dialogue history and then uses a personalized decoder to generate a response. Since this model also learns the user profile directly from the dialogue history, it is the most relevant baseline of our method.

4.3 Implementation Details

We experiment with multiple sets of hyperparameters to select the best model, and the final hyperparameters are listed as follows: The dimensions of the embeddings and Transformer hidden units are 768. The number of heads in the Transformer is 12. The number of layers in the Transformer is set as 2 and 12 respectively in the query encoder and decoder. The topic number is 15, and the similar user number is set as 10. The selected profile token number is 200 for the Weibo dataset and 30 for the Reddit dataset. The batch size is 128. Following (Holtzman et al., 2020), we adopt nucleus sampling as our decoding strategy. We use the Adam (Kingma and Ba, 2015) optimizer for training the refinners and AdamW (Loshchilov and Hutter, 2019) with a warm-up method for the generator. Our code is publicly available.3

4.4 Evaluation

**Metric-based Evaluation** We first evaluate all methods by several metrics with respect to different aspects. (1) BLEU-1/2 (Papineni et al., 2002) and ROUGE-L (Lin and Och, 2004) are typical word overlap-based metrics for measuring the similarity between the generated response and the ground-truth.4 (2) Distinct-1/2 (Li et al., 2016b) consider the number of uni- or bi-grams in the generated response, which is commonly used for evaluating the diversity. (3) The embedding-based metrics (i.e., average, extrema, and greedy) (Chan et al., 2019) are introduced to measure the semantic similarity between the generated response and the ground-truth one. (4) As a personalized dialogue model, following previous studies (Ma et al., 2021), two tailored metrics are adopted to measure how much information is included in the dialogue history that can be reflected in the response. Persona-F1 (P-F1) (Lian et al., 2019) calculates the F1 value to measure the uni-grams co-occurring in both the generated response and the dialogue history. Persona Coverage (P-Cover) (Song et al., 2019) calculates the IDF-weighted word overlap between the generated response and the golden one so that the importance of different words can be taken into account.

**Human Annotation** Due to the variability of human language, a response that differs from the ground-truth may also be appropriate. Following previous studies (Chan et al., 2019), we conduct a human evaluation of all methods. Concretely, we

4The results of BLEU-3/4 and ROUGE-1/2 are provided in Appendix A.
Table 3: Metric-based evaluation result on Weibo and Reddit dataset. The best results are in **bold**. “†” indicates that our model achieves significant improvement in t-test with p-value < 0.05.

<table>
<thead>
<tr>
<th></th>
<th>Overlap-based Metric</th>
<th>Diversity</th>
<th>Embedding Metric</th>
<th>Persona Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>BLEU-1</td>
<td>BLEU-2</td>
<td>ROUGE-L</td>
<td>Dist-1</td>
</tr>
<tr>
<td>Weibo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>3.330†</td>
<td>0.294†</td>
<td>8.95†</td>
<td>0.931†</td>
</tr>
<tr>
<td>MMI</td>
<td>3.631†</td>
<td>0.095†</td>
<td>5.26†</td>
<td>10.71†</td>
</tr>
<tr>
<td>DiagOFT</td>
<td>6.068†</td>
<td>0.741†</td>
<td>8.45†</td>
<td>15.32†</td>
</tr>
<tr>
<td>GPNN</td>
<td>4.899†</td>
<td>0.696†</td>
<td>7.78†</td>
<td>11.72†</td>
</tr>
<tr>
<td>PerCVAE</td>
<td>5.114†</td>
<td>0.299†</td>
<td>7.38†</td>
<td>14.09†</td>
</tr>
<tr>
<td>Speaker</td>
<td>4.994†</td>
<td>0.113†</td>
<td>7.86†</td>
<td>6.03†</td>
</tr>
<tr>
<td>PersonaWAE</td>
<td>3.510†</td>
<td>0.155†</td>
<td>10.54†</td>
<td>2.55†</td>
</tr>
<tr>
<td>DHAP</td>
<td>9.324†</td>
<td>0.894†</td>
<td>14.12†</td>
<td>15.17†</td>
</tr>
<tr>
<td>MSP (Ours)</td>
<td>11.875</td>
<td>5.108</td>
<td>15.56†</td>
<td>24.20†</td>
</tr>
</tbody>
</table>

| Reddit         |        |        |         |        |       |         |        |        |      |       |
|----------------|        |        |         |        |       |         |        |        |      |       |
| Seq2Seq        | 1.820† | 0.023† | 4.06†   | 5.20†  | 19.48† | 0.545†  | 0.554† | 0.470† | 0.051† | 0.029† |
| MMI            | 2.065† | 0.011† | 3.78†   | 5.91†  | 31.09† | 0.543†  | 0.607† | 0.454† | 0.085† | 0.038† |
| DiagOFT        | 4.735† | 0.397† | 8.94†   | 6.35†  | 29.10† | 0.604†  | 0.733† | 0.448† | 0.137† | 0.040† |
| GPNN           | 2.686† | 0.376† | 4.77†   | 12.32† | 35.76† | 0.406†  | 0.331† | 0.358† | 0.189† | 0.037† |
| PerCVAE        | 5.933† | 0.576† | 8.11†   | 9.63†  | 40.21† | 0.637†  | 0.649† | 0.499† | 0.212† | 0.040† |
| Speaker        | 2.642† | 0.054† | 4.46†   | 8.95†  | 34.18† | 0.538†  | 0.606† | 0.457† | 0.115† | 0.031† |
| PersonaWAE     | 2.637† | 0.113† | 8.19†   | 1.75†  | 25.91† | 0.629†  | 0.685† | 0.442† | 0.206† | 0.032† |
| DHAP           | 6.858† | 0.737† | 11.72†  | 18.70† | 66.93† | 0.709†  | 0.721† | 0.539† | 0.227† | 0.111† |
| MSP (Ours)     | 7.174† | 0.883† | 12.17†  | 21.24† | 68.89† | 0.716†  | 0.764† | 0.545† | 0.276† | 0.137† |

Sample 100 (query, response, user dialogue history) triplets and hire three well-educated annotators to score the responses generated by different models. Three aspects, *i.e.*, readability, informativeness, and personalization, are considered. The first two factors are scored on a scale of [1, 3] for their quality, while the third is assessed on a scale of [0, 1], indicating whether the response can accurately reflect the user’s personality. The detailed scoring criteria are shown in Table 2.

### 4.5 Experimental Results

#### Metric-based Evaluation

Table 3 shows all models’ performance under different metrics. On both datasets, it is clear to see that our MSP model outperforms baselines on all metrics. The improvement is statistically significant (t-test with p-value < 0.05). These findings indicate that our model is capable of generating more fluent, diverse, and personalized responses than all baselines. In particular, we can observe: (1) MSP achieves better performance on overlap-based metrics. This suggests that our model can provide responses that are more similar to the ground-truth with the help of the selected tokens. (2) For diversity metrics, the higher distinct values show that our generated responses are more diverse. Additionally, predefined profile-based methods and user dialogue history-based methods outperform others. This shows that incorporating external information can aid in the generation of more informative responses. (3) In addition to generating more overlapped words with the ground-truth response, the improvements in embedding metrics reflect that our model generates more semantically relevant responses. (4) Finally, the increase in personalized metrics implies that our approach can incorporate more user-specific information into the generation. Furthermore, the significant improvement over DHAP demonstrates that our model can extract more meaningful personalized information from the user dialogue history.

#### Human Annotation

The result of human annotation on the Weibo dataset is shown in Table 4. The Fleiss Kappa is around 0.62, indicating a substantial agreement achieved by three annotators. In general, the results of human annotation are consistent with those of the metric-based evaluation. Both of them demonstrate our model’s superiority in generating more fluent, informative, and personalized responses. Compared to non-personalized methods, user id-based methods can enhance personalization at the expense of readability. User dialogue history-based methods (i.e., DHAP and MSP) can largely improve the personalization of the response while retaining a high level of readability and informativeness. We attribute this to the abundant personal information contained in the user dialogue history.

Table 4: The result of human evaluation on Weibo dataset. “†” indicates that our model achieves significant improvement in t-test with $p$-value < 0.05.

<table>
<thead>
<tr>
<th>Model</th>
<th>Readability</th>
<th>Informativeness</th>
<th>Personality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>1.76</td>
<td>1.37†</td>
<td>0.11†</td>
</tr>
<tr>
<td>MMI</td>
<td>1.96</td>
<td>1.88†</td>
<td>0.19†</td>
</tr>
<tr>
<td>DialoGPT</td>
<td>2.33</td>
<td>2.10†</td>
<td>0.32†</td>
</tr>
<tr>
<td>GPIM</td>
<td>2.01†</td>
<td>2.16†</td>
<td>0.35†</td>
</tr>
<tr>
<td>PerCVAE</td>
<td>2.10†</td>
<td>2.01†</td>
<td>0.39†</td>
</tr>
<tr>
<td>Speaker</td>
<td>1.89†</td>
<td>1.44†</td>
<td>0.24†</td>
</tr>
<tr>
<td>PersonaWAE</td>
<td>1.81†</td>
<td>2.01†</td>
<td>0.32†</td>
</tr>
<tr>
<td>DHAP</td>
<td>2.29†</td>
<td>2.19†</td>
<td>0.55†</td>
</tr>
<tr>
<td>MSP (Ours)</td>
<td>2.37</td>
<td>2.39</td>
<td>0.67</td>
</tr>
<tr>
<td>Ground-Truth</td>
<td>2.71</td>
<td>2.66</td>
<td>0.76</td>
</tr>
</tbody>
</table>

Table 5: The results of ablation experiments on Weibo dataset.

<table>
<thead>
<tr>
<th>Models</th>
<th>BLEU-1</th>
<th>BLEU-2</th>
<th>P-Cover</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSP (Full)</td>
<td>11.875</td>
<td>5.108</td>
<td>0.297</td>
</tr>
<tr>
<td>w/o User Refiner</td>
<td>6.093</td>
<td>0.757</td>
<td>0.151</td>
</tr>
<tr>
<td>w/o Topic Refiner</td>
<td>6.163</td>
<td>0.839</td>
<td>0.178</td>
</tr>
<tr>
<td>w/o Token Refiner</td>
<td>4.213</td>
<td>0.609</td>
<td>0.116</td>
</tr>
<tr>
<td>w/o Current U’s Profile</td>
<td>9.365</td>
<td>3.146</td>
<td>0.238</td>
</tr>
<tr>
<td>w/o Similar Us’ Profile</td>
<td>6.413</td>
<td>0.871</td>
<td>0.245</td>
</tr>
<tr>
<td>w/o Joint Training</td>
<td>6.070</td>
<td>0.749</td>
<td>0.130</td>
</tr>
</tbody>
</table>

5 Further Analysis

We further conduct a series of analyses to elaborate our model. All analyses here are based on the result of the Weibo dataset, while similar results can be observed on the Reddit dataset.

Ablation Study  To investigate the impact of different modules in MSP, we conduct an ablation study by removing or using different strategies in each module.

We first study the influence of the refiners at three levels: (1) We remove the user refiner and train our model using randomly sampled users. We can see the performance of all metrics drops. This illustrates that our MSP model can select users that share the same interests as the current user and thereby improving response quality. (2) We remove the topic refiner and supply the token refiner with full dialogue history. The performance degradation demonstrates that various topics in dialogue history introduce lots of noise, misleading the token refiner on extracting valuable tokens, thus impairing the personalized response generation. (3) We eliminate the token refiner and feed all dialogue history sentences directly into the generator. The decline in performance implies the effectiveness and necessity of token selection. It is worth noting that, as compared to using the complete history, our selection strategy can reduce training time by 41.6%, considerably increasing efficiency. All of the aforementioned experimental results suggest that MSP’s advantage stems from high-quality personalized information extraction rather than simply introducing additional information.

We then explore the impact of personalized information from two sources, i.e., the current user’s profile and the similar users’ profile. Removing either of them results in decreased performance. This exemplifies their usefulness. Specifically, compared with similar users’ profiles, eliminating the current user’s profile will hurt the personalization effect heavily. This result shows that, for personalization, the current user’s profile is more essential than that of similar users, which is quite intuitive. However, the similar users’ profile has a significant effect on BLEU-1/2, implying that such a profile can provide abundant information for response generation. Consequently, integrating both types of profiles significantly improves response generation.

Finally, we conduct an experiment to validate our proposed joint training for the token refiner. The declining performance indicates that the token refiner is unable to extract useful information in the absence of additional supervision signals. Indeed, when the sentence matching task is removed, the token refiner extracts tokens that are relevant to the current query, which is less useful for generating a personalized response.

Influence of Selection Mechanism  To validate the effectiveness of our proposed selection mechanism, we replace the refiner with a traditional retrieval method (i.e., BM25 (Robertson and Walker, 1994)). Specifically, we use the query to retrieve
15 relevant responses and feed them into our model for training. The experimental results are shown in Figure 2. We can observe that the retrieval strategy achieves comparable performance with our model on word-overlap and embedding-based metrics. This suggests that the relevant dialogue history for the query can provide valuable information for response generation. However, the retrieval strategy performs poorly on diversity and personalization metrics. This demonstrates that, without careful selection, the retrieved information is too generic and thus less helpful for personalized response generation.

Influence of Personalized Tokens Amount In MSP, three refiners are designed to extract personalized tokens for response generation. Intuitively, the amount of the tokens will influence the refiner’s performance. We report this influence in Figure 3. As we can see in the left part, the quality of response generation improves with more tokens used. This is because fewer tokens are incapable of covering sufficient personalized information for response generation. Our MSP model performs optimally when about 200 personalized tokens are selected. When more tokens are introduced, the performance degrades. The potential reason is that more tokens would bring the noise to the generation. This is consistent with our speculation that the dialogue history is noisy and the information selection is both effective and necessary.

To provide a more qualitative view of our method, we conduct a case study in Appendix B.

6 Conclusion

In this work, we propose an MSP model for personalized response generation. Unlike previous related work, we utilize a refiner structure to extract query-aware persona information from large-scale dialogue history. The multi-level refiners can sparsely extract valuable information from dialogue history and leverage similar users’ information to enhance the current user’s personalization. Experimental results confirm the effectiveness of our model in generating informative and personalized responses.

Acknowledgement

Zhicheng Dou is the corresponding author. This work was supported by National Natural Science Foundation of China No. 61872370 and No. 61832017, Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the “Double-First Class” Initiative, Renmin University of China. We also acknowledge the support provided and contribution made by Public Policy and Decision-making Research Lab of RUC.

References


Table 6: The results of other evaluation metrics on Weibo and Reddit dataset. “†” indicates that our model achieves significant improvement in t-test with p-value < 0.05. The best results are in bold.

<table>
<thead>
<tr>
<th></th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>ROUGE-1</th>
<th>ROUGE-2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weibo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>0.011†</td>
<td>0.001†</td>
<td>8.740†</td>
<td>0.373†</td>
</tr>
<tr>
<td>MMI</td>
<td>0.046†</td>
<td>0.007†</td>
<td>5.316†</td>
<td>0.105†</td>
</tr>
<tr>
<td>DialoGPT</td>
<td>0.114†</td>
<td>0.027†</td>
<td>9.414†</td>
<td>0.632†</td>
</tr>
<tr>
<td>GPMN</td>
<td>0.359†</td>
<td>0.066†</td>
<td>8.086†</td>
<td>0.753†</td>
</tr>
<tr>
<td>PerCVAE</td>
<td>0.466†</td>
<td>0.089†</td>
<td>7.946†</td>
<td>0.485†</td>
</tr>
<tr>
<td>Speaker</td>
<td>0.107†</td>
<td>0.041†</td>
<td>7.997†</td>
<td>0.155†</td>
</tr>
<tr>
<td>PersonaWAE</td>
<td>0.889†</td>
<td>0.155†</td>
<td>11.341†</td>
<td>0.358†</td>
</tr>
<tr>
<td>DHAP</td>
<td>1.170†</td>
<td>0.401†</td>
<td>14.131†</td>
<td>3.608†</td>
</tr>
<tr>
<td>MSP (Ours)</td>
<td>3.973</td>
<td>3.522</td>
<td>16.249</td>
<td>5.812</td>
</tr>
<tr>
<td>Reddit</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Seq2Seq</td>
<td>0.007†</td>
<td>0.001†</td>
<td>3.989†</td>
<td>0.233†</td>
</tr>
<tr>
<td>MMI</td>
<td>0.007†</td>
<td>0.003†</td>
<td>3.960†</td>
<td>0.245†</td>
</tr>
<tr>
<td>DialoGPT</td>
<td>0.054†</td>
<td>0.010†</td>
<td>8.977†</td>
<td>0.610†</td>
</tr>
<tr>
<td>GPMN</td>
<td>0.039†</td>
<td>0.006†</td>
<td>4.896†</td>
<td>0.330†</td>
</tr>
<tr>
<td>PerCVAE</td>
<td>0.068†</td>
<td>0.009†</td>
<td>8.004†</td>
<td>0.540†</td>
</tr>
<tr>
<td>Speaker</td>
<td>0.021†</td>
<td>0.005†</td>
<td>4.017†</td>
<td>0.245†</td>
</tr>
<tr>
<td>PersonaWAE</td>
<td>0.029†</td>
<td>0.007†</td>
<td>8.247†</td>
<td>0.517†</td>
</tr>
<tr>
<td>DHAP</td>
<td>0.079†</td>
<td>0.013†</td>
<td>10.680†</td>
<td>0.697†</td>
</tr>
<tr>
<td>MSP (Ours)</td>
<td>0.106</td>
<td>0.019</td>
<td>11.078</td>
<td>0.745</td>
</tr>
</tbody>
</table>

A Additional Experimental Results

As n-gram word overlap metrics can reflect user speaking style more accurately, we evaluate the BLEU-3/4 (Papineni et al., 2002), and the result is shown in Table 6. It is consistent with other evaluations that our model outperforms every indicator. This demonstrates that user profiles also contain speaking style information, and our model can use the information to achieve a personalized response.

B Case Study

To show the effect of our model more concretely, we adopt a case study, and the results are shown in Table 7. It shows that our model can extract profiles from both current and similar users and generate informative and personalized responses. Specifically, in his dialogue history, he mentioned sports H1, H2 and music H3, H4 topics. Firstly, we can select similar users who also talk about sports and music, using the user refiner. Then, as the query is related to the music topic, we extract the current and sim users’ dialogue history responses about the music topic R1, R2 by the topic refiner. Furthermore, the token refiner selects some meaningful and personalized words from long sentences of reference. In this case, we can find that the token refiner extracts some compliment words (like, beautiful, enjoying) from the current user’s history sentence since the current user likes listening to music. And the token refiner captures more concrete tokens from sim users’ history sentences, such as “Angela Leung” and “Ember”. By combining two profiles, our personalized generator gets an informative and personalized response close to ground-truth. In contrast, DialoGPT generates a fluent but meaningless response to the query.
Table 7: A case study. Due to space limitation, we omit some user history responses and references.

| History                          | H1: Liverpool’s configuration has the life of a champion, support it!  
|                                 | H2: Champion Liverpool!  
|                                 | H3: I like to listen to music.  
|                                 | H4: The songs on this album are really beautiful, so worth enjoying.  
| Query                           | New album "How Am I? - The sun rises as it always does", the first single "Ember" heals the system, go listen!  
| Persona Reference               | R1: I like to listen to music.  
|                                 | R2: The songs on this album are really beautiful, so worth enjoying.  
| Sim Reference                   | R1: Quite like this type of song.  
|                                 | R2: Angela Leung’s "Ember" is really good.  
| Persona Profile                 | like, song, album, beautiful, enjoying  
| Sim Profile                     | like, song, Angela Leung, good, Ember  
| Response                        | DialoGPT: My little heart has flown.  
|                                 | MSP(Ours): Angela Leung’s song is very good.  
|                                 | Golden: Angela Leung’s songs must be listened to.  