Partner Matters! An Empirical Study on Fusing Personas for Personalized Response Selection in Retrieval-Based Chatbots

Jia-Chen Gu¹, Hui Liu², Zhen-Hua Ling¹, Quan Liu¹,³, Zhigang Chen³, 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

gjc@mail.ustc.edu.cn, {hui.liu,xiaodan.zhu}@queensu.ca,{zhling,quanliu}@ustc.edu.cn,zgchen@iflytek.com

ABSTRACT
Persona can function as the prior knowledge for maintaining the consistency of dialogue systems. Most of previous studies adopted the self persona in dialogue whose response was about to be selected from a set of candidates or directly generated, but few have noticed the role of partner in dialogue. This paper makes an attempt to thoroughly explore the impact of utilizing personas that describe either self or partner speakers on the task of response selection in retrieval-based chatbots. Four persona fusion strategies are designed, which assume personas interact with contexts or responses in different ways. These strategies are implemented into three representative models for response selection, which are based on the Hierarchical Recurrent Encoder (HRE), Interactive Matching Network (IMN) and Bidirectional Encoder Representations from Transformers (BERT) respectively. Empirical studies on the Persona-Chat dataset show that the partner personas neglected in previous studies can improve the accuracy of response selection in the IMN- and BERT-based models. Besides, our BERT-based model implemented with the context-response-aware persona fusion strategy outperforms previous methods by margins larger than 2.7% on original personas and 4.6% on revised personas in terms of hits@1 (top-1 accuracy), achieving a new state-of-the-art performance on the Persona-Chat dataset.

CCS CONCEPTS
• Information systems → Personalization;

KEYWORDS
Persona fusion, self and partner speakers, multi-turn response selection, retrieval-based chatbot

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

1 INTRODUCTION
Building human-like conversational systems has been a long-standing goal in artificial intelligence, where one of the major challenges is to present a consistent personality [38]. Personalized response selection, which aims to select an appropriate response from a set of candidates given the conversation context and personas of speakers, is an important technique to present personalities of dialogue agents in retrieval-based chatbots [8, 9, 18, 35, 37]. Retrieval-based chatbots are commonly used to build dialogue agents [4–7, 16, 28, 29, 33, 34, 36, 40]. Nowadays, many companies have built retrieval-based virtual assistants due to its promising potentials and alluring commercial values [2, 11, 39]. Zhang et al. [35] constructed a Persona-Chat dataset for building personalized dialogue agents, where each persona was represented as multiple sentences of profile description. A raw example dialogue conditioned on given profiles from the Persona-Chat dataset is shown in Table 1.

Although great progress has been made for building personalized dialogue agents [8, 9, 14, 18, 35, 37], they are still in their infancy. Most of previous studies focused on the self speaker’s persona in dialogue who was about to utter a response, while the contribution of the partner speaker’s persona to dialogue was rarely noticed. For a conversation conditioned on personas, if a dialogue agent has no access to the partner persona, it often over-focuses on retrieving responses related to the agent itself, which sometimes deviates from the ground truth of how a conversation really goes. For example, given a conversation about hobbies, if the agent only has access to the self persona profile “I like playing basketball”, it often over-weights response candidates such as “I like sports”. However, if the agent also has access to the partner persona profile “I often play various instruments”, it gives models more flexibility to not only focus on continuously talking about the agent itself, but also conducting more collaborative communication, e.g., asking questions such as “Who is your favorite musician”, as the real conversations often proceed.

Whether and how personas of different speakers in dialogue contribute to building coherent personalized dialogue models is a fundamental problem. In order to compare the ability of different personas for selecting an appropriate response directly, we first perform the preliminary experiments by ablating the context information and searching for an appropriate response with only the given self or partner persona information. Hierarchical
Table 1: An example dialogue from the Persona-Chat dataset.

| Persona 1                                      | Persona 2                                      |
|-----------------------------------------------|-----------------------------------------------|
| Original                                      | Original                                      |
| I just bought a brand new house.             | I love to meet new people.                    |
| I like to dance at the club.                 | I have a turtle named timothy.                |
| I run a dog obedience school.                | My favorite sport is ultimate frisbee.        |
| I have a big sweet tooth.                    | My parents are living in bora bora.           |
| I like taking and posting selkies.           | Autumn is my favorite season.                 |
| Revised                                      | Revised                                      |
| I have purchased a home.                     | I like getting friends.                       |
| Just go dancing at the nightclub, it is fun!| Reptiles make good pets.                      |
| I really enjoy animals.                      | I love to run around and get out my energy.   |
| I enjoy chocolate.                           | My family lives on a island.                  |
| I pose for pictures and put them online.     | I love watching the leaves change colors.     |

Dialogue
Person 1: Hello, how are you doing tonight?
Person 2: I am well and loving this interaction how are you?
Person 1: I am great. I just got back from the club.
Person 2: This is my favorite time of the year season wise.
Person 1: I would rather eat chocolate cake during this season.
Person 2: What club did you go to? Me an timothy watched tv.
Person 1: I went to club chino. What show are you watching?
Person 2: LOL oh okay kind of random.
Person 1: I love those shows. I am really craving cake.
Person 2: Why does that matter any? I went outdoors to play frisbee.
Person 1: It matters because I have a sweet tooth.
Person 2: So? LOL I want to meet my family at home in bora.
Person 1: My family lives in alaska. It is freezing down there.
Person 2: I bet it is oh I could not.

Table 2: Performance of persona-response matching on the Persona-Chat dataset conditioned on the original persona.

| Model | Persona | hits@1 | MRR  |
|-------|---------|--------|------|
| HRE   | Self    | 23.9   | 40.1 |
|       | Partner | 8.7    | 23.4 |
| IMN   | Self    | 48.8   | 60.7 |
|       | Partner | 19.3   | 34.2 |
| BERT  | Self    | 50.6   | 62.5 |
|       | Partner | 20.6   | 35.6 |

Recurrent Encoder (HRE) [25], Interactive Matching Network (IMN) [6] and Bidirectional Encoder Representations from Transformers (BERT) [3] are chosen as the matching models and the results are shown in Table 2. The results show that the single persona-response matching can achieve a comparable performance, which shows the usefulness of utilizing persona information to select an appropriate response. Meanwhile, it can be seen that although the partner persona is less important than the self persona, it can still contribute to response selection to some extent, which is consistent with our assumption mentioned above. However, few studies noticed the role of partner in dialogue and under what conditions the partner speaker’s persona can contribute more has not been studied too much yet.

To this end, we make an attempt to explore the impact of utilizing personas that describe either self or partner speakers on the task of personalized response selection. Four persona fusion strategies, i.e., none-aware (NA), context-aware (CA), response-aware (RA) and context-response-aware (CRA) ones, are designed based on whether or not considering the interactions between personas and contexts as well as the interactions between personas and responses. For a thorough comparison and analysis, these four strategies are implemented into three representative models for response selection, which are based on the HRE, IMN and BERT models respectively. HRE follows the sentence-encoding-based framework for response selection, which encodes contexts and responses separately without interactions between them and obtains their embeddings separately. As a representative model under the cross-attention-based framework, IMN performs the interactive matching between contexts and responses to derive the matching information between them. Meanwhile, IMN shares the most similar architecture with HRE we implemented in this paper, so that we can explore the effect of interactions between contexts and responses on persona fusion. The BERT-based response selection model not only performs interactions between contexts and responses, but also incorporates generic distributional semantics and other knowledge through pretraining.

We introduce our models and test the proposed persona fusion methods on the Persona-Chat dataset [35] which is the largest public dataset to date containing multi-turn dialogues conditioned
on personas. Experimental results show that the partner persona contributes to the performance when using the IMN- and BERT-based models. Besides, the pretraining algorithms can help to capture deep semantics given more context. Furthermore, compared with previous methods, our BERT-based model implemented with the context-response-aware persona fusion strategy improves hits@1 (top-1 accuracy) by 2.7% on original personas and by 4.6% on revised personas, achieving a new state-of-the-art performance on this dataset.

In summary, the contributions of this paper are two-fold. First, four persona fusion strategies are designed and implemented into three models, aiming to explore the impact of utilizing the personas of not only self but also partner speakers on response selection. Second, experimental results demonstrate that our proposed models outperform the existing state-of-the-art models by large margins on the widely used Persona-Chat response selection benchmark.

2 RELATED WORK

Chit-chat models suffer from a lack of a consistent personality as they are typically trained over many dialogues, each with different speakers, and a lack of explicit long-term memory as they are typically trained to produce an utterance given only a very recent dialogue history. Existing methods used to build a dialogue agent can be generally categorized into generation-based [14, 25, 26] and retrieval-based methods [4–7, 16, 28, 29, 33, 34, 36, 40]. Nowadays, many companies have built personalized virtual assistants due to its promising potentials and alluring commercial values [2, 11, 39]. Li et al. [14] proposed a persona-based neural conversation model to capture individual characteristics such as background information and speaking style. Miller et al. [20] proposed the key-value memory network, where the keys were dialogue histories, i.e., contexts, and the values were next dialogue utterances. Zhang et al. [35] constructed a Persona-Chat dataset for building personalized dialogue agents, which is the largest public dataset to date containing multi-turn dialogues conditioned on personas. It also established many baselines for this benchmark, such as the profile memory network by considering the dialogue history as input and then performing attention over the context-response-aware persona fusion strategy. It should be emphasized that the focus of this paper is not so much on designing drastically new models, but instead on investigating the conditions under which the self and partner personas can work. We aim to comprehensively understand the impact of utilizing the personas from not only the self but also the partner speakers on personalized response selection. Thus, we design four persona fusion strategies, and choose three very representative models to apply these strategies into these models in order to verify the effectiveness of these strategies. We can certainly choose other models because models are just a testbed for applying these strategies, which are not the focus of this paper. Instead, exploring the conditions under which the self and partner personas can work is our focus. We hope our work help shed some light on combining self and partner personas to further improve performance.

3 TASK DEFINITION

Given a dialogue dataset $D$ with personas, an example of the dataset can be represented as tuple $(c, p, r, y)$ and is shown in Table 1. Specifically, $c = \{u_1, u_2, ..., u_n\}$ represents a context with $(u_m)_{m=1}^{n_c}$ as its utterances and $n_c$ as the utterance number. $p = \{p_1, p_2, ..., p_n\}$ represents a persona with $(p_n)_{n=1}^{n_p}$ as its profile sentences and $n_p$ as the profile number. $r$ represents a response candidate. $y \in \{0, 1\}$ denotes a label, $y = 1$ indicates that $r$ is a proper response for $(c, p)$; otherwise, $y = 0$. Our goal is to learn a matching model $g(c, p, r)$ from $D$. For any context-persona-response triple $(c, p, r)$, $g(c, p, r)$ measures the matching degree between $(c, p)$ and $r$.

4 PERSONA FUSION FOR RESPONSE SELECTION

Capturing personas of different speakers in dialogue is the key for developing personalized dialogue agents. In order to thoroughly explore the impact of both self and partner personas on dialogue, we design four persona fusion strategies that assume personas interact with contexts or responses in different ways and implement them into three models, which are sentence-encoding-based, cross-attention-based and pretraining-based ones. Details about the model structures and the strategies are presented in this section.
Figure 1: Comparison of the model architectures for (a) HRE and (b) IMN.

4.1 Sentence-Encoding-Based Model

A representative model under the sentence-encoding-based framework for multi-turn dialogue is Hierarchical Recurrent Encoder-Decoder (HRED) [25] which was originally proposed for dialogue generation. Here, we only need the encoder part to obtain the encoded embedding, so we name it Hierarchical Recurrent Encoder (HRE) in this paper.

Figure 1 (a) shows an overview of the architecture. First, we follow the setting used in IMN [6], which constructs word representations by combining general pretrained word embeddings, those estimated on the task-specific training set, as well as character-level embeddings, in order to deal with the out-of-vocabulary issue. Formally, embeddings of the $m$-th utterance in a context, the $n$-th profile sentence in a persona and a response candidate are denoted as $U_m = \{u_{m,i}\}_{i=1}^{l_m}$, $P_n = \{p_{n,j}\}_{j=1}^{l_n}$ and $R = \{r_k\}_{k=1}^{l_r}$ respectively, where $l_m$, $l_n$ and $l_r$ are the numbers of words in $U_m$, $P_n$ and $R$ respectively. Each $u_{m,i}$, $p_{n,j}$, or $r_k$ is an embedding vector.

Then, context utterances, persona profiles and response candidate are encoded by bidirectional long short-term memories (BiLSTMs) [10]. Detailed calculations of BiLSTM are omitted due to limited space. We denote the calculations as:

$$\hat{u}_{m,i} = \text{BiLSTM}(U_m, i), i \in \{1, ..., l_m\},$$
$$\hat{p}_{n,j} = \text{BiLSTM}(P_n, j), j \in \{1, ..., l_n\},$$
$$\hat{r}_k = \text{BiLSTM}(R, k), k \in \{1, ..., l_r\},$$

where $U_m = \{\hat{u}_{m,i}\}_{i=1}^{l_m}$, $P_n = \{\hat{p}_{n,j}\}_{j=1}^{l_n}$ and $R = \{\hat{r}_k\}_{k=1}^{l_r}$. The parameters of these three BiLSTMs are shared in our implementation. Each $u_{m,i}$, $p_{n,j}$, or $r_k$ is an embedding vector.

The matching matrix $U_m$, $P_n$, and $R$ are aggregated by the max and last-hidden-state pooling operations to derive their embedding vectors as

$$u_{m,agr} = \{u_{m,max}: \hat{u}_{m,i}\}_{i=1}^{l_m}, m \in \{1, ..., n_c\},$$
$$p_{n,agr} = \{p_{n,max}: \hat{p}_{n,j}\}_{j=1}^{l_n}, n \in \{1, ..., n_p\},$$
$$\hat{r}_{agr} = [\hat{r}_{max}; \hat{r}_1].$$

Next, the sequences of $u_{m,agr}$ and $p_{n,agr}$ are further aggregated to get the embedding vectors for the context and the persona respectively. As the utterances in a context are chronologically ordered, the utterance embeddings $U_{agr} = \{u_{m,agr}\}_{m=1}^{n_c}$ are sent into another BiLSTM following the chronological order of utterances in the context. Combined max pooling and last-hidden-state pooling operations are then performed to obtain the context embeddings as

$$\hat{u}_{m} = \text{BiLSTM}(U_{agr}, m), m \in \{1, ..., n_c\},$$
$$\hat{c}_{agr} = [\hat{u}_{max}; \hat{u}_{n_c}].$$

Similarly, given the sequence of profile embeddings $\{p_{n,agr}\}_{n=1}^{n_p}$, the aggregated persona embedding $\hat{p}_{agr}$ is obtained by persona fusion. In this paper, four persona fusion strategies are designed based on whether or not considering the interactions between personas and contexts, and the interactions between personas and responses.

4.1.1 None-Aware Persona Fusion. In this strategy, the persona fusion is independent of both contexts and responses. A self-attention-based aggregation is designed to derive the persona embedding as follows,

$$a_n = w^T \cdot \hat{p}_{agr} + b,$$
$$\hat{p}_{agr} = \sum_{n=1}^{n_p} \frac{e^{a_n}}{\sum_{k=1}^{n_p} e^{a_k}} \hat{p}_{agr},$$

where $w$ and $b$ are parameters that need to be estimated during training. Then, the aggregated persona embedding is fused as a part of the final matching feature as shown in Eq. (17). This persona fusion strategy is not aware of any information of context and response, so we name it accordingly as none-aware (NA) persona fusion in this paper.

4.1.2 Context-Aware Persona Fusion. In order to make aware of the context information when fusing the persona, we design a context-aware (CA) persona fusion strategy by computing similarities between the context embedding and each profile embedding, and then performing the attention operation to obtain the aggregated
person embedding $\hat{\mathbf{p}}^{agr}$ as
\begin{equation}
\alpha_n = e^{agr}^\top \mathbf{p}_n^{agr},
\end{equation}
\begin{equation}
\hat{\mathbf{p}}^{agr} = \sum_{n=1}^{n_p} \frac{e^{agr}_n}{\sum_{k=1}^{n_p} e^{agr}_k} \mathbf{p}_n^{agr}.
\end{equation}

This persona fusion strategy is aware of the context information by attaching different importance to profile embeddings according to their similarities to the context dynamically, so we name it context-aware (CA) persona fusion in this paper.

### 4.1.3 Response-Aware Persona Fusion

Similarly, we design a response-aware (RA) persona fusion strategy by computing similarities between the response embedding and each profile embedding, and then performing the attention operation to obtain the aggregated persona embedding $\hat{\mathbf{p}}^{agr}$ as follows,
\begin{equation}
\alpha_n = e^{agr}^\top \mathbf{p}_n^{agr},
\end{equation}
\begin{equation}
\hat{\mathbf{p}}^{agr} = \sum_{n=1}^{n_p} \frac{e^{agr}_n}{\sum_{k=1}^{n_p} e^{agr}_k} \mathbf{p}_n^{agr}.
\end{equation}

Then the same attention operation as Eq. (10) is performed to obtain $\hat{\mathbf{p}}^{agr}$.

### 4.1.4 Context-Response-Aware Persona Fusion

In order to make aware of both the context and the response information simultaneously, we design a context-response-aware (CRA) persona fusion strategy. This strategy first concatenates the context and the response embedding, and then transforms it to the same dimension of profile embeddings with a linear transformation. Similarities are computed between it and each profile embedding. Then the same attention operation is performed to obtain $\hat{\mathbf{p}}^{agr}$. Mathematically, we have
\begin{equation}
\alpha_n = (\mathbf{w}^\top \cdot [\mathbf{c}^{agr}; \mathbf{r}^{agr}] + b)^\top \cdot \mathbf{p}_n^{agr},
\end{equation}
\begin{equation}
\hat{\mathbf{p}}^{agr} = \sum_{n=1}^{n_p} \frac{e^{agr}_n}{\sum_{k=1}^{n_p} e^{agr}_k} \mathbf{p}_n^{agr}.
\end{equation}

Lastly, after obtaining the aggregated persona embedding $\hat{\mathbf{p}}^{agr}$, the final matching feature vector is the concatenation of the context, persona and response embeddings as
\begin{equation}
\mathbf{m} = [\mathbf{c}^{agr}; \mathbf{p}^{agr}; \mathbf{r}^{agr}].
\end{equation}

The final matching feature vector is then sent into a multi-layer perceptron (MLP) classifier. Here, the MLP classifier is designed to predict whether a context-response-persona triple $(c, p, r)$ match appropriately based on the derived matching feature vector and returns a score denoting the matching degree of this triple. Finally, a softmax output layer is adopted in the MLP to return a probability distribution over all response candidates. Models are learnt by minimizing the MLP cross-entropy loss. Let $\Theta$ denote the model parameters. The learning objective $\mathcal{L}(\mathcal{D}, \Theta)$ is formulated as
\begin{equation}
\mathcal{L}(\mathcal{D}, \Theta) = - \sum_{(c,p,r,y) \in \mathcal{D}} y \log g(c, p, r)).
\end{equation}

### 4.2 Cross-Attention-Based Model

A representative model under the cross-attention-based framework for multi-turn dialogue is Interactive Matching Network (IMN) [6], which was originally proposed for multi-turn response selection. Another reason to choose this model is that it shares the most similar architecture with HRE we implemented in this paper, so that we can explore the effect of interactions between contexts and responses on persona fusion.

Figure 1 (b) shows an overview of the architecture. IMN shares with HRE the same modules of word representation, sentence encoding, aggregation, persona fusion and prediction. In addition, IMN is equipped with an interaction module which performs the global and bidirectional cross-attention operation between the context and the response to capture the matching information between them. We introduce the interaction module briefly as follows and readers could refer to [6] for more details of IMN. First, after looking up the wording embedding table and encoded by the sentence encoder to derive the set of utterance representations $\{\mathbf{U}_m\}_{m=1}^{n_c}$ and the response representations $\mathbf{R}$, the context representation $\mathbf{C} = \{e_c\}_{i=1}^{l_c}$ with $l_c = \sum_{m=1}^{n_c} l_{um}$ is formed by concatenating the set of utterance representations $\{\mathbf{U}_m\}_{m=1}^{n_c}$. Then, IMN matches the response with the whole context in a global and bidirectional way, i.e., considering the whole context as a single sequence. The global context-response matching can help select the most relevant parts of the whole context and neglect the irrelevant parts.

An attention-based alignment is employed to collect information between the context and the response by computing the attention weight between each $(\mathbf{c}_i, \mathbf{r}_k)$ tuple as
\begin{equation}
e_{ik} = (\mathbf{c}_i)^\top \cdot \mathbf{r}_k.
\end{equation}

For a word in the response, its response-to-context relevant representation is composed as a weighted summation of $\{e_c\}_{i=1}^{l_c}$. The same calculation is performed for each word in a context to compose its context-to-response representation as a weighted summation of $\{\mathbf{r}_k\}_{k=1}^{l_r}$. To further enhance the collected information, the element-wise differences and products with their representations after the sentence encoder are calculated and are then concatenated to obtain the enhanced representations. Finally, the concatenated context representations need to be converted back to separate utterance representations which are sent for further aggregation.

It is notable that the persona aggregation in IMN is identical to that in HRE. Readers can refer to Section 4.1 for details.

### 4.3 Pretraining-Based Model

A representative model under the pretraining-based framework is Bidirectional Encoder Representations from Transformers (BERT) [3]. Due to space limitation, we omit an exhaustive background description of BERT. Readers can refer to [3] for details. The four persona fusion strategies are implemented by adjusting BERT to fit the task of personalized response selection in different ways.

#### 4.3.1 None-Aware Persona Fusion

In this strategy, we propose a dual matching architecture which is composed of two encoding pipelines. One is used to derive the matching feature between contexts and responses, and the other is used to derive the persona
After encoded by stacked Transformer blocks [30], the embedding is aware of the context information by interactions between the first token to form the sequence B in BERT. Then, these two sequences are used to form the sequence A in BERT, and the context is used to form the sequence A in BERT, and the response is used to form the sequence B. Then these two sequences are concatenated with a [SEP] token. In order to further distinguish between them, three subtypes of embedding are added to the corresponding token representations, in addition to the original sequence A/B embeddings, which are parameters updated during the fine-tuning process. The encoded embedding of the first token [CLS] of each concatenated sequence is used as the aggregated representation for a persona-context-response triple classification. This embedding captures the matching information in this triple. In this strategy, the persona fusion is aware of both contexts and responses by interactions with both of them. Finally, this embedding is sent into a MLP classifier, and returns a score denoting the matching degree of this triple.

4.3.3 Response-Aware Persona Fusion. The strategy is similar to the context-aware persona fusion strategy in BERT, except that it replaces the context with the response when deriving the persona fusion feature. Figure 2 (b) shows an overview of the architecture.

4.3.4 Context-Response-Aware Persona Fusion. Different from the strategies mentioned above which derive the context-response matching feature and the persona fusion feature respectively, in this strategy, we propose a simple yet effective method to derive a feature which contains both types of information simultaneously. Figure 2 (c) shows an overview of the architecture. Specifically, the persona and the context are concatenated to form the sequence A, and the response is used to form the sequence B. Then these two sequences are concatenated with a [SEP] token. To ensure results are comparable, we used the same evaluation metrics as in the previous work [9, 35]. Each model aimed to select the best-matched response from available candidates for the given context c and persona p. We calculated the recall of the true positive replies, denoted as hits@1. In addition, the mean reciprocal rank (MRR) was also adopted to take the rank of the correct response over all candidates into consideration.
5.3 Training Details

For building HRE, IMN, and their persona fusion models, the ratio of positive and negative responses was set to 1:19 in the training set, with a softmax output layer over all response candidates. The Adam method [12] was employed for optimization with a batch size of 16. The initial learning rate was 0.001 and was exponentially decayed by 0.96 every 5000 steps. Dropout [27] with a rate of 0.2 was applied to the word embeddings and all hidden layers. The maximum number of training epochs was set to 10. The word representation is a concatenation of a 300-dimensional GloVe embedding [23], a 100-dimensional embedding estimated on the training set using the Word2Vec algorithm [19], and 150-dimensional character-level embeddings with window sizes [3, 4, 5], each consisting of 50 filters. The word embeddings were not updated during training. All hidden states of the LSTM have 200 dimensions. The MLP at the prediction layer have 256 hidden units with ReLU [21] activation. The maximum number of characters in a word, that of words in a context utterance, of utterances in a context, of words in a persona profile, of profiles in a persona, and of words in a response were set to 18, 20, 15, 15, 5, and 20, respectively. We padded with zeros if the number of utterances in a context was less than 15; otherwise, we kept the last 15 utterances. Similarly, we padded with zeros if the number of profile sentences in a persona was less than 5. We used the validation set to select the best model for testing.

For building BERT and its persona fusion models, we employed the base version of BERT and most hyper-parameters of the original BERT were followed [3] except the following configurations. The initial learning rate was set to 2e-5 and was linearly decayed by L2 weight decay. A dynamic negative sampling strategy was adopted that the ratio of positive and negative responses was set to 1:1 in the training set, and it used different negative responses at each epoch. Thus, the maximum number of training epochs was set to 19. The maximum sequence length was set to 320. The training batch size was set to 12. The MLP at the prediction layer was a single-layer feed-forward neural network with sigmoid activation.

All code was implemented in the TensorFlow framework [1] and is published to help replicate our results.\(^1\)

5.4 Comparison Methods

Non-pretraining-based methods. IR baseline, Starspace, Profile and KV Profile were baselines established in Zhang et al. [35] who released the Persona-Chat dataset. DGMN [37], DIM [9] and FIRE [8] were follow-up studies which did not employ any pretraining.

Pretraining-based methods. FT-PC [18] employed the “pretrain and fine-tune” framework by first pretraining on a domain-specific corpus, dialogues of which were extracted from Reddit, and then fine-tuning on the Persona-Chat. TransferTransfo [32] and P2Bot [15] were both initialized with the pretrained language model of GPT [24] which were pretrained on a large general corpus, and then fine-tuned on the Persona-Chat as well.

5.5 Experimental Results

For comparison, Table 3 presents the performance of HRE, IMN and BERT models together with previous methods on this task without using any personas. In our experiments, all reported results were the top one out of four runs under the same model configuration. Thus, the reported baseline results were the best we attain.\(^2\)

Table 4 presents the evaluation results of our proposed persona fusion strategies implemented into three models on the Persona-Chat dataset under the original persona configuration. Numbers marked with * denote that the gains or losses after adding persona conditions are statistically significant (t-test with p-value < 0.05) comparing with the corresponding baseline models in Table 3. Numbers in bold denote the persona fusion strategy that achieves the best performance.

| Table 3: Evaluation results of our reimplemented HRE, IMN and BERT models together with previous methods on the Persona-Chat dataset without using any personas. |
|----------------|----------------|----------------|
|                | hits@1 | MRR  |                  |
|----------------|--------|------|------------------|
| IR baseline [35] | 21.4   | -    |                  |
| Starspace [35]   | 31.8   | -    |                  |
| Profile [35]     | 31.8   | -    |                  |
| KV Profile [35]  | 34.9   | -    |                  |
| HRE [25]         | 42.7   | 60.0 |                  |
| IMN [6]          | 63.8   | 75.8 |                  |
| BERT[3]          | 70.7   | 80.8 |                  |

\(^1\)https://github.com/JasonForJoy/Personalized-Response-Selection

\(^2\)Our reimplemented HRE, IMN and BERT were the same as those presented in [3, 6, 25] on their datasets.
and responses could help to determine the importance of different parts of a persona and then fuse them. Third, the RA persona fusion strategy performs best among four strategies in the sentence-encoding-based and cross-attention-based models and the CRA fusion strategy performs best in the pretraining-based model. We consider this is because responses are closely related to personas while pretraining algorithms seem to help capture more semantics given additional contexts.

Table 5 presents the evaluation results of our proposed and previous methods on the Persona-Chat dataset under various persona configurations. Our BERT-based model implemented with the context-response-aware persona fusion strategy achieves a new state-of-the-art performance. We can see that incorporating the generic distributional semantics and external knowledge learned from pretraining rendered improvements on both hits@1 and MRR conditioned on various personas. Compared with the FT-PC model [18], BERT-CRA and BERT-RA outperformed it by margins of 18.7% and 16.4% respectively in terms of hits@1 conditioned on revised self personas. The results show that the generic distributional semantics and other knowledge learning from pretraining is beneficial for building personalized dialogue agents. Compared with TransferTransfo [32] and P² Bot [15] which were also equipped with the generic distributional semantics, BERT-CRA and BERT-RA still outperformed them, which shows the effectiveness of our proposed persona fusion strategies. Lastly, BERT-CRA and BERT-RA outperformed all previous pretraining-based and non-pretraining-based methods by margins larger than 2.7% and 1.0% respectively hits@1 conditioned on original self personas, and margins larger than 4.6% and 2.3% respectively in terms of hits@1 conditioned on revised self personas. The results show that our proposed models achieved superiority on original personas and greater advantages on revised personas, achieving a new state-of-the-art performance of response selection on the Persona-Chat dataset.

### 5.6 Analysis

**Subtype Embeddings.** In order to demonstrate the importance of subtype embeddings used in BERT-CRA, an ablation test was further performed and the results were shown in the last row in Table 5. We can see that the subtype embeddings contribute to the performance of BERT-CRA, which shows its effectiveness to distinguish personas, contexts and responses from each other.

**Retrieval Time.** The retrieval time is very critical for retrieval-based chatbots. Thus, we tested the time complexity by recording the inference time over the whole validation set using a GeForce RTX 2080 Ti GPU. The results were reported by averaging two runs as shown in Table 6. Although BERT-based models take more time, it is acceptable compared to the performance they achieved. Meanwhile, some recent studies have explored the methods of accelerating inference of BERT for the deployment in real-time.

| Efficiency (cases/second) |
|---------------------------|
| HRE-NA 4660.1 | IMN-NA 1661.4 | BERT-NA 53.29 |
| HRE-CA 4956.9 | IMN-CA 1666.5 | BERT-CA 53.29 |
| HRE-RA 4626.9 | IMN-RA 1674.6 | BERT-RA 53.29 |
| HRE-CRA 4643.5 | IMN-CRA 1688.0 | BERT-CRA 92.67 |

Table 6: The efficiency (cases/second) of four persona fusion strategies implemented into three models by recording the inference time over the whole validation set on the Persona-Chat dataset under the original persona configuration.
Table 7: Performance of response generation conditioned on the original persona. Numbers marked with * denote that the gains or losses after adding persona conditions are statistically significant (t-test with p-value < 0.05). Numbers in bold denote the persona fusion strategy that achieves the best performance.

| Model-Persona          | Relevance | Diversity | Length |
|------------------------|-----------|-----------|--------|
|                         | BLEU      | Dist-1    | Dist-2 |
| MiniLM                  | 3.54      | 94.47     | 99.49  | 8.73  |
| MiniLM-CA-Self          | 4.50*     | 94.94*    | 99.59  | 8.59  |
| MiniLM-CA-Partner       | 3.65*     | 93.80*    | 99.48  | 9.05* |

5.7 Discussion on Response Generation

Although fusing personas for dialogue response generation is not the focus of this paper, we conducted a preliminary experiment to show that self or partner personas also contribute differently to response generation.

It is notable that only context is available while response candidate is not during inference in response generation. Thus in this section, we explored the impact of self and partner personas on only the context-aware persona fusion strategy. Furthermore, we implemented the strategy into pretrained-based models. In our experiment, we adopt a lightweight model MiniLM [31] for the consideration of time and space complexity. Due to space limitation, we omit the introduction of MiniLM here and readers can refer to Wang et al. [31] for details.

The overview of the model architecture for MiniLM with the context-aware persona fusion strategy is shown in Figure 3. In this strategy, we follow the setting in previous studies [15, 38] in which all the persona profiles are concatenated directly and the context utterances are concatenated and separated with the [SEP] token. Then, the persona and the context are concatenated with a [SEP] token. The concatenated persona-context combination are fed into the model as the input for the Seq2Seq generation.

We take two sets of widely used metrics to evaluate the relevance and the diversity of the generated responses. For relevance, we use BLEU [22] which is a weighted summation of BLEU 1-4 and the length. For diversity, we calculate the uni-gram and bi-gram distinct ratios (DIST-1, DIST-2) at the instance level [13].

Table 8: An example from the generated responses that demonstrates the different contributions of the self and partner personas. Given the conversation context, ✗ denotes an inappropriate response and ✓ denotes an appropriate one.

| Self Persona | Partner Personas |
|--------------|------------------|
| I like to go hunting. | I live in Ohio. |
| I am a handyman. | I like to go hiking. |
| I am allergic to shellfish. | I am a single mom of two boys. |
| I restore classic cars. | I work as an accountant. |

**Context:**
how are you tonight, i just got back from hiking.

**Response with self persona:**
i am good, i just got back from hunting.

**Response with partner persona:**
that sounds fun, i just got back from a hike. ✗

Hiking in Ohio must be very interesting. ✓

Table 7 presents the performance of MiniLM with context-aware (CA) persona fusion strategy on the task of response generation using the original version of persona. As we can see, the conclusions are consistent with those on response selection. First, compared to partner persona, self persona can help achieve better results on both the relevance and diversity metrics, which demonstrates that self persona can provide more fundamental information about the speaker who is about to utter a response. Second, although the partner persona was mostly thought to be not useful in previous studies, our results show that partner persona contribute to the performance on response generation. When partner persona is given, the BLEU score is improved on the MiniLM-based models.

We assume that self and partner personas contribute differently in response generation, while an example shown in Table 8 can verify our assumption to some extent. As we can see, if we consider the partner persona as the self persona equally, the generated response will confuse the information from self or partner. Designing effective personas fusion strategies for self and partner on the task of response generation is a large scope and we will leave it to our future work.

6 CONCLUSIONS

In this paper, we propose four persona fusion strategies to explore the impact of self and partner personas on personalized response selection in retrieval-based chatbots. These strategies are implemented into three representative models for evaluation and comparison. Empirical studies on the Persona-Chat dataset show that the partner persona neglected in previous studies can still improve the performance under certain conditions. Besides, our proposed models improve the accuracy of response selection, outperforming previous methods by large margins and achieving a new state-of-the-art performance of response selection on the Persona-Chat dataset. In the future, we will work on exploring the impact of self and partner personas for dialogue response generation to further verify the usefulness of partner in dialogue.
ACKNOWLEDGEMENTS
We thank anonymous reviewers for their valuable comments.

REFERENCES
[1] Martin Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard et al. 2016. TensorFlow: A System for Large-Scale Machine Learning. In 12th (USENIX) Symposium on Operating Systems Design and Implementation (OSDI). 162–283.
[2] Ana Berdasco, Gustavo López, Ignacio Díaz, Luis Quezada, and Luis A. Guerrero. 2019. User Experience Comparison of Intelligent Personal Assistants: Alexa, Google Assistant, Siri and Cortana. In Multidisciplinary Digital Publishing Institute Proceedings, Vol. 31. 51.
[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 4171–4186.
[4] Jia-Chen Gu, Tianda Li, Zhen-Hua Ling, Quan Liu, Zhiming Su, Yu-Ping Ruan, and Xiaodan Zhu. 2021. Deep Contextualized Utterance Representations for Response Selection and Dialogue Analysis. IEEE/ACM Transactions on Audio, Speech, and Language Processing (2021).
[5] Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2041–2044.
[6] 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 on Information and Knowledge Management. 2321–2324.
[7] Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2020. Utterance-to-Utterance Interactive Matching Network for Multi-Turn Response Selection in Retrieval-Based Chatbots. IEEE/ACM Transactions on Audio, Speech, and Language Processing 28 (2020), 369–379.
[8] Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, Zhihang Chen, and Xiaodan Zhu. 2020. Filtering before Iteratively Referring for Knowledge-Grounded Response Selection in Retrieval-Based Chatbots. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP): Findings. 1412–1422.
[9] Jia-Chen Gu, Zhen-Hua Ling, Xiaodan Zh, and Quan Liu. 2019. Dually Interactive Matching Network for Personalized Response Selection in Retrieval-Based Chatbots. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 1845–1854.
[10] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. In Proceedings of the 1997 Conference of the International Joint Conference on Neural Networks. 2775–2783.
[11] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and their Compositionality. In Advances in neural information processing systems. 3111–3119.
[12] Vijay Shetty and Xiandong Zhang. 2017. Improved Language Understanding by Generative Pre-Training. In Proceedings of the 2017 Conference of the Association for Computational Linguistics (Volume 1: Long Papers). 1007–1017.
[13] Emma Smith, Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, Zhigang Chen, and Xiaodan Zhu. 2021. Deep Contextualized Utterance Representations for Response Selection and Dialogue Analysis. In Proceedings of the 2021 Conference of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 994–1003.
[14] Jia-Chen Gu, Zhen-Hua Ling, Quan Liu, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-Aware BERT for Multi-Turn Response Selection in Retrieval-Based Chatbots. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management. 2041–2044.
[15] Qian Liu, Yihong Chen, Bei Chen, Jin-Guang Lou, Zixuan Chen, Bin Zhou, Jia-Chen Gu, Zhen-Hua Ling, and Quan Liu. 2019. Interactive Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 1845–1854.
[16] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. In Proceedings of the 1997 Conference of the International Joint Conference on Neural Networks. 2775–2783.