Persona-aware Dialogue Generation with Enriched Profile

Yinhe Zheng\textsuperscript{1,2}, Guanyi Chen\textsuperscript{3}, Minlie Huang\textsuperscript{2}, Song Liu\textsuperscript{1}, Xuan Zhu\textsuperscript{1}
\textsuperscript{1} Samsung Research China – Beijing (SRC-B), \textsuperscript{2} Tsinghua University, \textsuperscript{3} Utrecht University
yh.zheng@samsung.com, g.chen@uu.nl, aihuang@tsinghua.edu.cn, s0101.liu@samsung.com, xuan.zhu@samsung.com

Abstract

Endowing a dialogue system with particular persona is essential to deliver more human-like conversations. However, due to the difficulties of embodying personalities in natural language, this problem is still far from well studied. This paper proposes a novel task of generating dialogue responses conditioned on explicit personal profiles with rich attributes. A dataset is constructed to facilitate the proposed task and a persona-aware dialogue generation model is also introduced. In this model, a structured personal profile (in key-value pairs) is transformed and composed into a representation vector using a persona fusion module, and several different fusion methods are tested. Two techniques, namely persona-aware attention and persona-aware bias, are devised to capture and incorporate various persona attributes in the decoding process. Experiments and case studies demonstrate that our model is able to address proper attributes in different contexts.

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 persona, so that the system can gain users’ confidence and trust (Shum et al., 2018). The user engagement level of a dialogue agent increases when the agent is conditioned on various persona settings (Zhang et al., 2018), including age, gender, language, location, level of knowledge, or even a proper accent. The ability of exhibiting a certain persona with various attributes is essential for conversational systems to well interact with users in a more natural and coherent way (Qian et al., 2018; Li et al., 2016b; Kottur et al., 2017).

Recent studies in this area can be broadly classified into two types. One is implicit models, in which a speaker’s persona is modeled implicitly from the dialogue data and represented using a speaker embedding (Li et al., 2016b; Kottur et al., 2017; Zhang et al., 2017). In this approach, it is unclear what types of persona are captured and how personas are interpreted. Moreover, these methods also suffer from the data sparsity issue: there should be a sufficient amount of dialogues from each speaker to train a reliable model. The other is explicit models, in which a speaker’s persona is given explicitly, i.e., the generated responses are explicitly conditioned either on a given profile (Qian et al., 2018), or on a text-described persona (Zhang et al., 2018). However, these methods are limited to either manually-labeled data (Qian et al., 2018) or crowdsourced dialogues (Zhang et al., 2018; Dinan et al., 2019; Wolf et al., 2018), thereby not scalable to large-scale dialogue datasets.

As a matter of fact, the persona of a speaker can be viewed as a composite of various attributes. During conversations, people may reveal some of these attributes (e.g. Age or Gender), consciously or unconsciously. For example, as shown in the conversations of Figure 1, speaker B uses the word “tomboy” in response to speaker A’s comment. It can be inferred that speaker B is a female. Similarly, based on the second and third turns of this session, we can easily infer that both speaker A and B are...
Figure 1: An example dialogue session and the personal profile of each speaker. Words in response are in the same color with the corresponding attributes.

living in Shenzhen (a city in Guangdong province, China). Therefore, we argue that an intelligent conversational agent can be more user-engaging if it is equipped with a persona with various attributes and is able to decide which attribute to express in different contexts.

To address the above issues, we propose a novel task and a persona-aware model for generating dialogue responses conditioned on an explicit personal profile with various attributes. A dataset is also constructed to facilitate the study of the proposed task, which is unique in several aspects: First, the persona of each speaker is presented explicitly in a structured manner (using key-value pairs). With such structured persona, the dialogue data across speakers with same attributes can be shared to alleviate the data sparsity issue; Second, although the persona is given explicitly, the use of such information can be captured implicitly by scalable, data-driven methods. For instance, a female speaker may not necessarily say the word “female” to reveal her gender in every utterance she responds with, instead, she may consciously or unconsciously use related words that may reveal her gender in particular contexts. This differs from prior explicit models (Qian et al., 2018) which require that values of the given personal profile must appear in a generated response and demand for manually-labeled data. Third, personas with various attributes are modeled. In fact, expressing persona in natural language is usually subtle and implicit (Bamman et al., 2014), it is interesting to study how personas with various attributes are expressed and revealed in dialogues.

In this paper, we employ the sequence to sequence learning framework (Sutskever et al., 2014) and devise a persona-aware dialogue model. Specifically, a persona fusion module is used to model different attributes of a speaker, i.e., each attribute is encoded as an embedding vector and different attributes are merged to produce an integrated persona representation. Two approaches are devised to leverage the persona representation in the generation process: the first approach introduces a persona-aware attention mechanism where the persona representation contributes to compute the attention weights to obtain the context vector at each decoding position, and the second approach adds a persona-aware bias to the word generation distribution. Automatic and manual evaluations indicate that our model and dataset help to incorporate proper personas when generating responses in different contexts.

2 Related Work

Traditional studies focus on psychology inspired persona-aware dialogue generation, i.e., modeling “Big Five” (Goldberg, 1993) (i.e., Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism) of speakers (Mairesse and Walker, 2007; Gill et al., 2012). However, annotating dialogues with “Big Five” labels is extremely difficult and such metric is very subjective. Therefore, “Big Five” is not suitable for building large-scale persona-aware dialogue systems, particularly with data-driven neural models.

Thanks to the success of applying neural models to dialogue systems, some data-driven persona-aware dialogue models and corresponding datasets have been introduced (Huang et al., 2019). Early studies focused on modeling characters in movie dialogues (Banchs, 2012). More recently, inspired by the successful application of social media data (Ritter et al., 2011; Serban et al., 2015) and the sequence to sequence learning framework (Sutskever et al., 2014; Serban et al., 2016; Li et al., 2016b), introduced
An overview of our persona-aware dialogue generation model is shown in Figure 2. Given a personal profile \( T \), a persona fusion module is used to merge attributes in \( T \) into a persona representation \( v_p \). Three approaches are proposed in this fusion module, namely attention, average, and concatenation. Two methods are devised to incorporate \( v_p \) into the decoding process: the first method introduces a persona-aware attention, namely, using \( v_p \) to generate the attention weights at each decoding position such that the context vector computed at each position is conditioned on \( v_p \); the second method applies a persona-aware bias directly in estimating the generation distribution.

3 Model

Our task can be formulated as follows: Given a post \( X \) and a personal profile \( T \) for the responder, the system should generate a response \( Y \) that embodies the given profile:

\[
Y^* = \arg \max_Y P(Y|X,T)
\]

Note that the profile \( T \) consists of a set of attributes \( T = \{t_1, t_2, ..., t_N\} \) and each attribute \( t_i \) is given as a key-value pair \( t_i = (k_i, v_i) \). The exact value \( v_i \) is not required to appear in \( Y \). Moreover, personas of speakers who make posts are not modeled in our task. We leave this as the future work.

An overview of our persona-aware dialogue generation model is shown in Figure 2. Given a personal profile \( T \), a persona fusion module is used to merge attributes in \( T \) into a persona representation \( v_p \in \mathbb{R}^{d_p} \). Three approaches are proposed in this fusion module, namely attention, average, and concatenation. Two methods are devised to incorporate \( v_p \) into the decoding process: the first method introduces a persona-aware attention, namely, using \( v_p \) to generate the attention weights at each decoding position such that the context vector computed at each position is conditioned on \( v_p \); the second method applies a persona-aware bias directly in estimating the generation distribution.

3.1 Sequence to Sequence Framework

The backbone of our model is the sequence to sequence (Seq2Seq) framework (Sutskever et al. 2014), which usually consists of an encoder and a decoder. The encoder encodes the post sequence \( X = x_1, x_2, ..., x_n \) into a sequence of hidden representations \( (h_1, h_2, ..., h_n) \), \( h_i \in \mathbb{R}^{d_h} \). The decoder will sample a word from a generation distribution over the vocabulary at each decoding step. The generation distribution is conditioned on the preceding state of the decoder, the previously generated word, and the context vector which is computed with an attention mechanism.

In this study, we use the attention mechanism proposed by Bahdanau et al. (2014), which produces a context representation \( c_t \in \mathbb{R}^{d_c} \) at each decoding step \( t \) by attending to the encoder’s outputs, at the
same time conditioned on the previous state of the decoder \( s_{t-1} \in \mathbb{R}^{d_s} \). Formally, we have:

\[
c_t = \sum_{i=1}^{n} \alpha_i h_i, \quad \alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \quad (1)
\]

where \( V \), \( W^1 \) and \( W^2 \) are trainable parameters.

In general Seq2Seq models, the generation probability \( P_t \) at step \( t \) of the decoder is produced by a softmax function:

\[
P_t = \text{softmax}(W^1 o_t + b_{out}), \quad s_t = \text{RNN}(s_{t-1}, c_t, w_{t-1}). \quad (2)
\]

where \( w_{t-1} \) is the word vector of the decoded word from previous time step. \( W^1_o \) and \( b_{out} \) are trainable parameters.

### 3.2 Persona Fusion

The persona fusion module is used to compute an integrated persona representation \( v_p \). We first map each attribute \( t_i \) in \( T \) to an embedding representation \( v_{t_i} \) using its corresponding attribute encoder. In this study, attribute encoders are implemented using look-up tables since each attribute considered (i.e., \( \text{Age} \), \( \text{Gender} \) and \( \text{Location} \)) only has one unique value for each speaker. Note that other kinds of attribute can also be modeled if proper encoders are provided. For instance, an LSTM encoder can be applied to model one-sentence self-descriptions of speakers. After encoding all the attributes in \( T \) into a set of attribute representations \( \{ v_{t_1}, v_{t_2}, \ldots, v_{t_N} \} \), we can merge them using a persona fusion function to obtain \( v_p \). In this paper, three different fusion methods are investigated:

**Attributes Attention.** Merge \( v_{t_i} \) \((i = 1, 2, \ldots, N)\) based on an attention mechanism. Specifically, an attention weight \( \alpha'_i \) is computed for \( v_{t_i} \) based on the state from the previous decoding step \( s_{t-1} \), and \( v_p \) is obtained as a weighted sum of all attribute representations:

\[
v_p = \sum_{i=1}^{N} \alpha'_i v_{t_i}, \quad \alpha'_i = \frac{\exp(e'_i)}{\sum_{j=1}^{N} \exp(e'_j)}, \quad e'_i = V^T \cdot \tanh(W^1_o s_{t-1} + W^2_o v_{t_i}) \quad (3)
\]

where \( V \), \( W^1_o \) and \( W^2_o \) are trainable parameters. The weight \( \alpha'_i \) indicates how much the current context favors attribute \( t_i \). It allows us to make proper combination of persona attributes with respect to different contexts.

**Attributes Average.** Average all the attribute representations in \( T \): \( v_p = \frac{1}{N} \sum_{i=1}^{N} v_{t_i} \), which is a special case of the Attributes Attention, where all attributes are weighted equally.

**Attributes Concatenation.** Concatenate each attribute representation in \( T \) to produce \( v_p \). In this case, the length of \( v_p \) (i.e., \( d_p \)) is divisible by \( N \).

### 3.3 Decoding with Persona Representation

In order to incorporate the persona representation \( v_p \) in our decoding process, we develop the following two methods:

**Persona-Aware Attention (PAA).** The first method extends the computation of attention weights (Equation[1]). The attention weight is now dependent on not only the decoder’s state, but also on \( v_p \), namely,

\[
e_i = V^T \cdot \tanh(W^1_o s_{t-1} + W^2_o h_t + W^3_o v_p) \quad (4)
\]

where \( V \), \( W^1_o \), \( W^2_o \), and \( W^3_o \) are trainable parameters. The score \( e_i \) is then used to compute attention weights. This approach helps our decoder to attend to different contexts based on the persona representation \( v_p \), i.e., being persona-aware.

**Persona-Aware Bias (PAB).** The second method tries to incorporate \( v_p \) in the output layer of the decoder. Specifically, we extend Equation[2] to include a persona-aware bias. A gate \( g_t \) is also devised to balance the original term and the persona bias term:

\[
P_t = \text{softmax}(g_t \cdot W^1_o s_t + (1 - g_t) \cdot W^2_o v_p + b_{out}), \quad g_t = \sigma(V^T_o \cdot s_t). \quad (5)
\]
Table 1: Basic statistics of PERSONAL DIALOG.

<table>
<thead>
<tr>
<th>Statistics</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of dialogues</td>
<td>20.83 M</td>
</tr>
<tr>
<td>Total number of speakers</td>
<td>8.47M</td>
</tr>
<tr>
<td>Total number of utterances</td>
<td>56.25 M</td>
</tr>
<tr>
<td>Dialogues with more than 4 utterances</td>
<td>3.43 M</td>
</tr>
<tr>
<td>Average utterances per dialogue</td>
<td>2.70</td>
</tr>
<tr>
<td>Average tokens per utterance</td>
<td>9.35</td>
</tr>
</tbody>
</table>

where \( W^1_o, W^2_o, b_{out} \) and \( V_o \) are trainable parameters. Note that the computed scalar variable \( g_t \in [0, 1] \) works as a gate to control how much persona features should be incorporated at each step. It decides whether to use persona related word or semantic related word, and thus makes the generated response more consistent. This approach is similar to modules used in many previous works (Zhou et al., 2018; Jaech and Ostendorf, 2018; Luo et al., 2019) and it is assumed to be more direct in influencing the generation distribution.

4 Experiments

4.1 Dataset

In order to facilitate the study of the proposed task, we constructed a large-scale dialogue dataset: PERSONAL DIALOG, which contains millions of dialogues (in single turn and multi turns) and various persona information for a large number of speakers. Specifically, this dataset comes from Weibo (a large social media in China). Each dialogue in it is composed of a Weibo post and its following replies. The personal profile of each speaker is collected and three persona attributes (i.e., Gender, Age, and Location) are approached in this study. The scripts used to construct this dataset is available upon requests. Researchers that are interested can utilize these scripts to collect their own dialogue data and the associated persona. A basic statistics of PERSONAL DIALOG is shown in Table 1. We randomly extracted 10M sessions of single-turn dialogues for training, and 20K sessions for validation.

Coarse-grained labels were considered for Age and Location attributes. Specially, Age had four labels: “post-70s” (born within 1970-1979), “post-80s” (1980-1989), “post-90s” (1990-1999), and “post-00s” (born after 2000). This simplification was made because previous studies had proved the impracticality of predicting exact ages of speakers with text-only inputs (Eckert, 2017). Similarly, ten Location labels were chosen based on the geolinguistic theory about dialect area distributions of Chinese (Cao and Liu, 2008), i.e., districts in similar dialect areas were merged.

4.2 Attribute Classifiers

To check if dialogues in PERSONAL DIALOG indeed carry persona related features. We built a classifier for each persona attribute that takes in dialogue texts and predicts the associated attribute labels of speakers.

The inputs to our classifiers were constructed by concatenating every 20 dialogue utterances issued by speakers with a same attribute label. This was a commonly used strategy in persona perception tasks on social media data (Flekova et al., 2016) because speakers wouldn’t reveal their persona in every single utterances they issued (Nguyen et al., 2014). Specifically, taking Gender classifier as an example: we first constructed two sets of utterances that were issued by males and females, respectively. Then for each set, we concatenated every 20 utterances and used the resulting concatenations as inputs to our gender classifier.

We constructed the dataset for each classifier using 10M dialogue sessions sampled from PERSONAL DIALOG and balanced the classes in each dataset with oversampling, and random split each dataset into train, validation and test set. For each classifier, we tried several commonly used classification models, including CNN (Kim, 2014), LSTM, and RCNN (Lai et al., 2015). RCNN obtain the best performance among other classifiers. Concretely, it reached accuracy at 90.61%, 78.32%, and 62.04% for gender, age, and location classification, respectively.
Table 2: Automatic evaluation on the unbiased test set with perplexity (ppx.), Distinct-1, Distinct-2 and attribute accuracy (acc.).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>84.07</td>
<td>0.0226</td>
<td>0.0599</td>
<td>50.2</td>
<td>25.3</td>
<td>10.2</td>
</tr>
<tr>
<td>Gen GLBA</td>
<td>79.05</td>
<td>0.0287</td>
<td>0.0764</td>
<td>73.5</td>
<td>25.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Age GLBA</td>
<td>79.21</td>
<td>0.0285</td>
<td>0.0743</td>
<td>50.1</td>
<td>42.0</td>
<td>10.0</td>
</tr>
<tr>
<td>Loc GLBA</td>
<td>80.04</td>
<td>0.0276</td>
<td>0.0689</td>
<td>50.1</td>
<td>25.1</td>
<td>19.6</td>
</tr>
<tr>
<td>Avg.+PAA</td>
<td>81.47</td>
<td>0.0271</td>
<td>0.0746</td>
<td>63.5</td>
<td>30.2</td>
<td>15.4</td>
</tr>
<tr>
<td>Con.+PAA</td>
<td>82.37</td>
<td>0.0272</td>
<td>0.0735</td>
<td>63.4</td>
<td>30.6</td>
<td>15.8</td>
</tr>
<tr>
<td>Att.+PAA</td>
<td>82.26</td>
<td>0.0259</td>
<td>0.0707</td>
<td>70.1</td>
<td>29.2</td>
<td>14.3</td>
</tr>
<tr>
<td>Avg.+PAB</td>
<td>79.46</td>
<td>0.0287</td>
<td>0.0741</td>
<td>76.7</td>
<td>37.2</td>
<td>20.7</td>
</tr>
<tr>
<td>Con.+PAB</td>
<td>81.51</td>
<td>0.0279</td>
<td>0.0779</td>
<td>77.9</td>
<td>37.5</td>
<td>20.8</td>
</tr>
<tr>
<td>Att.+PAB</td>
<td>78.44</td>
<td><strong>0.0293</strong></td>
<td><strong>0.0805</strong></td>
<td>77.1</td>
<td>38.9</td>
<td>22.2</td>
</tr>
</tbody>
</table>

4.3 Biased Test Sets

To test how well our models perform in different contexts, we built four test sets: unbiased set, gender-biased, age-biased, and location-biased set, each of which included 10k dialogue sessions. Specifically, unbiased set were randomly sampled, whereas biased sets were deliberately selected to provide us different contexts under which speakers tend to reveal certain persona attribute of themselves. For example, the post-response pair “Are you a boy or a girl?” and “I am a girl” are Gender biased because most speakers tend to reveal their gender in response to gender-related questions. It would be interesting to see if our model can capture this behaviour. Biased sets were constructed by selecting biased responses that were most likely to be correctly classified by our attribute classifiers (with high confidence).

Formally, for a biased set, we first randomly sampled a set of 50K post-response pairs $D$. Then for each response $r \in D$, we constructed $M$ classifier inputs $S_j(r)$, $j = 1, \ldots, M$ containing $r$. Specifically, $S_j(r)$ was a concatenation of 20 responses (see §4.2) and $r \in S_j(r)$, i.e., assuming we were building gender-biased set and $r$ was issued by a female, then $S_j(r)$ was a concatenation of $r$ and 19 female-issued responses randomly sampled from $D$. Then a confidence score for $r$ was computed by:

$$c(r) = \frac{1}{M} \sum_{j=1}^{M} \delta[S_j(r)]P[S_j(r)]$$  \hspace{1cm} (6)

where $P[S_j(r)] \in [0, 1]$ was the confidence score produced by our classifier for $S_j(r)$, $\delta[S_j(r)]$ was 1 if the label of $S_j(r)$ was correctly predicated, and -1 otherwise. We chose top 10k scored responses and used the associated post-response pairs as the biased test set. The intuition is that higher confidence score is brought by richer persona-related features. Therefore a response $r$ with higher $c(r)$ contains richer persona-related features, thereby being a biased response.

Classifier inputs were constructed using these selected biased responses (cf. §4.2) and fed into our classifiers. The resulting accuracy increased with the number of input $M$. In this study, we used $M = 2,000$ because the accuracy increase brought by larger $M$ was less than 0.1%. The accuracy scores associated with final biased test sets are shown in the last row of Table[3]. These high scores verified that the selected biased responses indeed carry rich persona-related features.

4.4 Implementation Details

The encoder and decoder were 2 layer GRUs with 512 hidden units for each. The vocabulary size was 40K and the dimension of word vectors was 100. Word embeddings were updated during training and shared by the encoder and decoder. The embedding size of the persona representation $v_p$ was 90. The Adam optimizer was used to train our model with a batch size of 120 and a learning rate of 0.001. Each model was trained about a week on a Titan X GPU machine.
Table 3: Automatic evaluation on biased test sets with attribute accuracy (acc.).

<table>
<thead>
<tr>
<th>Model</th>
<th>Gen.</th>
<th>Age</th>
<th>Loc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>85.3</td>
<td>79.8</td>
<td>27.2</td>
</tr>
<tr>
<td>Gen GLBA</td>
<td><strong>95.5</strong></td>
<td><strong>81.6</strong></td>
<td>31.8</td>
</tr>
<tr>
<td>Age GLBA</td>
<td>86.8</td>
<td><strong>92.1</strong></td>
<td>32.0</td>
</tr>
<tr>
<td>Loc GLBA</td>
<td>87.3</td>
<td>78.2</td>
<td><strong>48.2</strong></td>
</tr>
<tr>
<td>Avg. + PAA</td>
<td>91.1</td>
<td>88.5</td>
<td>43.3</td>
</tr>
<tr>
<td>Con. + PAA</td>
<td>91.7</td>
<td>88.9</td>
<td>44.5</td>
</tr>
<tr>
<td>Att. + PAA</td>
<td>94.0</td>
<td>88.3</td>
<td>42.5</td>
</tr>
<tr>
<td>Avg. + PAB</td>
<td>94.8</td>
<td>91.9</td>
<td>48.5</td>
</tr>
<tr>
<td>Con. + PAB</td>
<td>95.0</td>
<td>91.6</td>
<td>48.9</td>
</tr>
<tr>
<td>Att. + PAB</td>
<td><strong>96.0</strong></td>
<td><strong>92.5</strong></td>
<td><strong>50.3</strong></td>
</tr>
</tbody>
</table>

Golden responses\(^1\) 100.0 99.8 90.8

\(^1\) Human-generated responses.

Table 4: Manual evaluation with Fluency and Appropriateness.

<table>
<thead>
<tr>
<th>Model</th>
<th>Fluency</th>
<th>Appropriateness</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seq2Seq</td>
<td>4.685</td>
<td>3.889</td>
</tr>
<tr>
<td>Gen GLBA</td>
<td>4.732</td>
<td>3.850</td>
</tr>
<tr>
<td>Age GLBA</td>
<td>4.792</td>
<td>3.898</td>
</tr>
<tr>
<td>Loc GLBA</td>
<td>4.730</td>
<td>3.707</td>
</tr>
<tr>
<td>Att. + PAB</td>
<td><strong>4.822</strong></td>
<td><strong>3.971</strong></td>
</tr>
</tbody>
</table>

4.5 Baselines

We chose two different types of baselines: 1) a Seq2Seq model, which does not use persona features; 2) Three Group Linguistic Bias Aware (GLBA) models \cite{Wang2017}, which respectively incorporates three individual attributes, i.e., Gender, Age, and Location.

Our model was tested with different combinations of persona fusion methods and decoding schemes. Three persona fusion methods were attribute attention (Att.), average (Avg.) and concatenation (Con.). Two decoding schemes were Persona-Aware Attention (PAA) and Persona-Aware Bias (PAB). Thus six variants were tested.

Note that we did not adopt the speaker model \cite{Li2016} as our baseline because it requires a large amount of dialogues for each speaker. In fact, GLBA models were stronger baselines because it was a variation of the speaker model by modeling single persona attributes.

4.6 Automatic Evaluation

**Metrics:** Perplexity was used to evaluate our model at the content level. Smaller perplexity scores mean the generated responses are more grammatical and fluent. We also used **Distinct** \cite{Li2016} to measure the diversity of generated responses. To evaluate how well our models can capture personas, we defined attribute accuracy as the agreement between the expected attribute labels (as input to the model) and labels predicted by attribute classifiers based on generated responses. A higher attribute accuracy indicates stronger abilities to incorporate that attribute in responses. Similar metric was also used in other works \cite{Zhou2018}.

**Results:** The performance on the unbiased test set is shown in Table 2. The attribute accuracy shown in this table was obtained by assigning different values to the target attribute. For example, for the Gender attribute, we generated two sets of responses to the same posts with “Female” and “Male” label, respectively. We then constructed inputs to the attribute classifier as introduced in §4.2. We also tested our models on these biased test set (Table 3). In this case, a single response is generated for each post with a personal profile same as the responder in the biased test set.

Results in these tables show that: 1) GLBA models only performed well on single attributes. For example, the Gender GLBA model only achieved high attribute accuracy regarding to Gender, whereas it degraded remarkably on Age and Location. In comparison, our models achieved higher attribute accuracy with respect to all the attributes. This verifies that the persona fusion module was necessary to model personas with various attributes in different contexts; 2) Models equipped with PAB generally outperformed models with PAA. This may be due to the fact that PAB can influence the decoding process more directly; 3) Our model Att. + PAB obtained the best performance in terms of almost all the metrics, particularly on the biased test sets. This indicated that the attribute attention facilitates the modeling of personas with various attributes and it also helps to choose proper attributes in different contexts.
Table 5: Sample responses generated by baselines and our model (Att. + PAB). Words in response are in the same color with the corresponding persona attributes.

<table>
<thead>
<tr>
<th>Post</th>
<th>Seq2Seq</th>
<th>Personal Profile:</th>
<th>Age GLBA</th>
<th>Gen GLBA</th>
<th>Att. + PAB</th>
<th>Personal Profile:</th>
</tr>
</thead>
<tbody>
<tr>
<td>You should firstly have a boyfriend</td>
<td>I think so too</td>
<td>Female, Post-90s, Sichuan</td>
<td>I agree</td>
<td>You are my boyfriend</td>
<td>This princess has a boyfriend</td>
<td>Male, Post-90s, Sichuan</td>
</tr>
</tbody>
</table>

4.7 Manual Evaluation

We also performed manual evaluation. Given a post and the personal profile of a responder, responses were generated using all the baseline models and our best performing model (Att. + PAB). These responses were presented to three human annotators along with these posts and personal profiles.

Metrics. Annotators were asked to score a response in two aspects with a five-point scale, i.e., from 1 (strongly disagree) to 5 (strongly agree): 1) Fluency: The overall quality of the utterance is good in terms of its grammatical correctness and fluency. 2) Appropriateness: The usage of personal profiles in the generated response is reasonable and the response is coherent to the dialogue context.

Results. 100 posts were sampled from each of these four test sets (i.e., 400 posts in total), and totally 2K responses were generated using five models. The inter-rater consistency of the annotation results was measured using the Fleiss’ kappa $\kappa$ [Randolph 2005]. In particular, the $\kappa$ value for Fluency and Appropriateness was 0.82 and 0.53, respectively, indicating fairly good agreements between these annotation results. As shown in Table 4, our model outperformed all the baselines significantly ($t$-test, $p < .05$) in both metrics. This indicates that our model can learn to incorporate proper persona in generated responses. It is also interesting to see that the Seq2Seq model outperforms some GLBA models in Appropriateness. We argue that GLBA models tried to emphasize anchored attributes in every utterances they generated, resulting sub-optimal in producing logical and appropriate responses. In fact, different attributes should be embodied in respect of different contexts, and sometimes we do not even need to address any persona related features in the response.

4.8 Case Study

Table 5 shows a sampled case, in which our model could incorporate proper persona attributes in the produced responses based on the context, while the Seq2seq model tends to generate universal responses and GLBA models can only consider single attributes. Specifically, as the given post was talking about Gender related topic, our model chose to incorporate the Gender attribute in the generated response. Note that both our model and Gender GLBA model could produce different responses with different Gender labels, i.e., they can act as either a “Female” (colored in orange) or “Male” (colored in blue), while other models tended to ignore the Gender attribute.

5 Conclusion and Future Work

In this paper, we investigate a novel task to generate persona-aware dialogue responses conditioned on an explicitly represented personal profile with various attributes. A large-scale dialogue dataset is constructed to facilitate such a task. We present a persona-aware model to capture and address rich attributes in the dialogue generation process. A persona fusion module is applied to obtain the persona representation of a speaker. Two approaches are devised in the decoding process: namely persona-aware attention which dynamically generates context vectors conditioned on the persona representation, and persona-aware bias which manipulates the generation distribution directly. Automatic and manual evaluations show that our models can incorporate richer attributes in generated dialogues and can learn to choose proper persona in different contexts. In the future, we can extend our model to multi-turn dialogues and consider personas of the two speaking parties.
References


