Persona-aware Dialogue Generation with Enriched Profile

Yinhe Zheng[1,2], **Guanyi Chen**[3], **Minlie Huang**[2], **Song Liu**[1], **Xuan Zhu**[1]
[1] Samsung Research China – Beijing (SRC-B),
[2] Tsinghua University, [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

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

A: You would rather be fashionable than comfortable. (in cold winter) (真是要风度不要温度的)
B: Nope! I am a tomboy who prefer comfort than fashion. (才没有！我是个要温度不要风度的女汉子)
A: As your elder brother, I only have one such faerie like you. You have to take care of yourself for me. (哥哥我就这么一个小仙女，你要替我照顾好自己)
B: You are also in Shenzhen right? (你不是也在深圳)
A: Yeah, I have been in Shenzhen for several years. (对啊在深圳几年了)

Personal profile of A:
{ "Age": "24", "Gender": "Male", "Location": "Guangdong" }
Personal profile of B:
{ "Age": "23", "Gender": "Female", "Location": "Guangdong"}

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

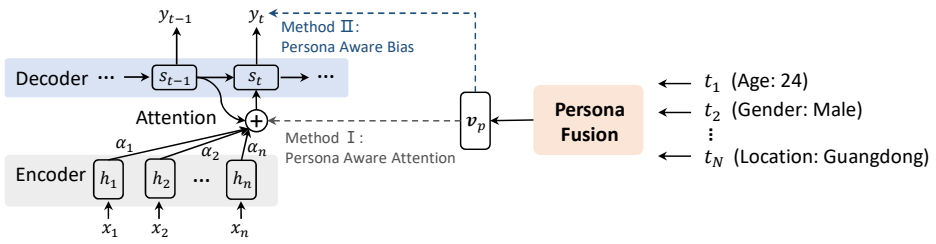

Figure 2: Overview of our persona-aware dialogue model. To obtain the persona representation $\boldsymbol{v}_p$, different attributes are integrated by the persona fusion module. $\boldsymbol{v}_p$ is then used to generate persona-aware attention weights for computing the context vector, or to produce a persona-aware bias for computing the generation distribution.

a model that represents each speaker with a persona vector and fed it to each decoding step to capture speaker-specific styles. Kottur et al. (2017) extended this idea to multi-turn dialogues. In these models, the persona is implicitly represented by a single real-valued vector.

An initial attempt to incorporate explicitly represented persona in open-domain dialogues is proposed by Qian et al. (2018), in which a chatbot is endowed with a structured profile. Their model ensures that a selected profile value must appear in a generated response and it also requires manually-labeled data. Other explicit models use unstructured personas that are described by several sentences in natural language (Wolf et al., 2018; Mazaré et al., 2018; Song et al., 2019). However, these dialogue data are collected via crowdsourcing services and it is expensive to expand such datasets (e.g., PERSONACHAT (Zhang et al., 2018)) to a large scale.

Another interesting approach to generate personalized dialogues is to use transfer learning. These models are usually pre-trained on a large corpus and then transferred to each speaker using a small amount of personalized dialogue data (Yang et al., 2017; Wolf et al., 2018; Madotto et al., 2019; Golovanov et al., 2019). However, most of these approaches focus on unstructured personas (i.e., described in texts), whereas structured personas (i.e., given in key-value pairs) are seldom modeled.

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 $t_i = \langle k_i, v_i \rangle$. 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 $\boldsymbol{v}_p \in \mathbb{R}^{d_p}$. Three approaches are proposed in this fusion module, namely attention, average, and concatenation. Two methods are devised to incorporate $\boldsymbol{v}_p$ into the decoding process: the first method introduces a persona-aware attention, namely, using $\boldsymbol{v}_p$ to generate the attention weights at each decoding position such that the context vector computed at each position is conditioned on $\boldsymbol{v}_p$; 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 $(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n)$, $\boldsymbol{h}_i \in \mathbb{R}^{d_s}$. 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 $\boldsymbol{c}_t \in \mathbb{R}^{d_s}$ at each decoding step t by attending to the encoder's outputs, at the

same time conditioned on the previous state of the decoder $\boldsymbol{s}_{t-1} \in \mathbb{R}^{d_s}$. Formally, we have:

$$\begin{aligned} \boldsymbol{c}_t &= \sum_{i=1}^{n} \alpha_i \boldsymbol{h}_i, \quad \alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \\ e_i &= V^T \cdot \tanh(W_\alpha^1 \boldsymbol{s}_{t-1} + W_\alpha^2 \boldsymbol{h}_i) \end{aligned} \tag{1}$$

where V, W_α^1 and W_α^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:

$$P_t = \text{softmax}(W_o^1 \boldsymbol{s}_t + b_{out}), \quad \boldsymbol{s}_t = \text{RNN}(\boldsymbol{s}_{t-1}, \boldsymbol{c}_t, w_{t-1}). \tag{2}$$

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

3.2 Persona Fusion

The persona fusion module is used to compute an integrated persona representation $\boldsymbol{v}_p$. We first map each attribute t_i in T to an embedding representation $\boldsymbol{v}_{t_i}$ using its corresponding attribute encoder. In this study, attribute encoders are implemented using look-up tables since each attribute considered (i.e., `Age`, `Gender` and `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 $\{\boldsymbol{v}_{t_1}, \boldsymbol{v}_{t_2}, ..., \boldsymbol{v}_{t_N}\}$, we can merge them using a *persona fusion function* to obtain $\boldsymbol{v}_p$. In this paper, three different fusion methods are investigated:

Attributes Attention. Merge $\boldsymbol{v}_{t_i}$ $(i = 1, 2, ..., N)$ based on an attention mechanism. Specifically, an attention weight α_i' is computed for $\boldsymbol{v}_{t_i}$ based on the state from the previous decoding step $\boldsymbol{s}_{t-1}$, and $\boldsymbol{v}_p$ is obtained as a weighted sum of all attribute representations:

$$\begin{aligned} \boldsymbol{v}_p &= \sum_{i=1}^{N} \alpha_i' \boldsymbol{v}_{t_i}, \quad \alpha_i' = \frac{\exp(e_i')}{\sum_{j}^{N} \exp(e_j')} \\ e_i' &= \overline{V}^T \cdot \tanh(\overline{W_\alpha^1} \boldsymbol{s}_{t-1} + \overline{W_\alpha^2} \boldsymbol{v}_{t_i}) \end{aligned} \tag{3}$$

where $\overline{V}$, $\overline{W_\alpha^1}$ and $\overline{W_\alpha^2}$ are trainable parameters. The weight α_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: $\boldsymbol{v}_p = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_{t_i}$, 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 $\boldsymbol{v}_p$. In this case, the length of $\boldsymbol{v}_p$ (i.e., d_p) is divisible by N.

3.3 Decoding with Persona Representation

In order to incorporate the persona representation $\boldsymbol{v}_p$ 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 $\boldsymbol{v}_p$, namely,

$$e_i = V^T \cdot \tanh(W_\alpha^1 \boldsymbol{s}_{t-1} + W_\alpha^2 \boldsymbol{h}_i + W_\alpha^3 \boldsymbol{v}_p) \tag{4}$$

where V, W_α^1, W_α^2, and W_α^3 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 $\boldsymbol{v}_p$, i.e., being persona-aware.

Persona-Aware Bias (PAB). The second method tries to incorporate $\boldsymbol{v}_p$ 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:

$$P_t = \text{softmax}(g_t \cdot W_o^1 \boldsymbol{s}_t + (1 - g_t) \cdot W_o^2 \boldsymbol{v}_p + b_{out}), \quad g_t = \sigma(V_o^T \cdot \boldsymbol{s}_t). \tag{5}$$

Table 1: Basic statistics of PERSONALDIALOG.

Total number of dialogues	20.83 M
Total number of speakers	8.47M
Total number of utterances	56.25 M
Dialogues with more than 4 utterances	3.43 M
Average utterances per dialogue	2.70
Average tokens per utterance	9.35

where W_o^1, W_o^2, 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: PERSONALDIALOG, 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 PERSONALDIALOG 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 PERSONALDIALOG 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 PERSONALDIALOG and balanced the classes in each dataset with oversampling, and random split each dataset into train, valiation 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.).

Model	ppx.	Distinct-1	Distinct-2	`Gen.` acc.	`Age` acc.	`Loc.` acc.
Seq2Seq	84.07	0.0226	0.0599	50.2	25.3	10.2
`Gen` GLBA	79.05	0.0287	0.0764	**73.5**	25.0	10.0
`Age` GLBA	79.21	0.0285	0.0743	50.1	**42.0**	10.0
`Loc` GLBA	80.04	0.0276	0.0689	50.1	25.1	**19.6**
Avg.+PAA	81.47	0.0271	0.0746	63.5	30.2	15.4
Con.+PAA	82.37	0.0272	0.0735	63.4	30.6	15.8
Att.+PAA	82.26	0.0259	0.0707	70.1	29.2	14.3
Avg.+PAB	79.46	0.0287	0.0741	76.7	37.2	20.7
Con.+PAB	81.51	0.0279	0.0779	**77.9**	37.5	20.8
Att.+PAB	**78.44**	**0.0293**	**0.0805**	77.1	**38.9**	**22.2**

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 $\mathcal{D}$. Then for each response $r \in \mathcal{D}$, we constructed M classifier inputs $S_j(r)$, $j = 1, ..., M$ 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 $\mathcal{D}$. Then a confidence score for r was computed by:

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

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 $\boldsymbol{v}_p$ 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.).

Model	Gen.	Age	Loc.
Seq2Seq	85.3	79.8	27.2
Gen GLBA	**95.5**	81.6	31.8
Age GLBA	86.8	**92.1**	32.0
Loc GLBA	87.3	78.2	**48.2**
Avg. + PAA	91.1	88.5	43.3
Con. + PAA	91.7	88.9	44.5
Att. + PAA	94.0	88.3	42.5
Avg. + PAB	94.8	91.9	48.5
Con. + PAB	95.0	91.6	48.9
Att. + PAB	**96.0**	**92.5**	**50.3**
Golden responses[1]	100.0	99.8	90.8

[1] Human-generated responses.

Table 4: Manual evaluation with *Fluency* and *Appropriateness*.

Model	Fluency	Appropriateness
Seq2Seq	4.685	3.889
Gen GLBA	4.732	3.850
Age GLBA	4.792	3.898
Loc GLBA	4.730	3.707
Att. + PAB	**4.822**	**3.971**

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 Wang et al. (2017), 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 Li et al. (2016b) 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* Li et al. (2016a) 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 Zhou et al. (2018).

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.

Post	You should firstly have a boyfriend (首先你要有一个男朋友)
Seq2Seq	I think so too (我也这么觉得)
Personal Profile: Female, Post-90s, Sichuan	
`Loc` GLBA	I agree (要的要的)
`Age` GLBA	I think I am ⟨UNK⟩ (我觉得我是⟨UNK⟩)
`Gen` GLBA	You are my boyfriend (你是我的男朋友)
Att. + PAB	This princess has a boyfriend (本公举有男朋友)
Personal Profile: Male, Post-90s, Sichuan	
`Gen` GLBA	I don't need a boyfriend (我不要男朋友)
Att. + PAB	I am straight, thanks (我是直男，谢谢)

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 κ Randolph (2005). In particular, the κ 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

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Bamman, D., Eisenstein, J., and Schnoebelen, T. (2014). Gender identity and lexical variation in social media. *Journal of Sociolinguistics*, 18(2):135–160.

Banchs, R. E. (2012). Movie-dic: A movie dialogue corpus for research and development. In *ACL*, pages 203–207.

Cao, Z. and Liu, X. (2008). *Linguistic atlas of Chinese dialects*. Beijing: Commercial Press.

Dinan, E., Logacheva, V., Malykh, V., Miller, A., Shuster, K., Urbanek, J., Kiela, D., Szlam, A., Serban, I., Lowe, R., et al. (2019). The second conversational intelligence challenge (convai2). *arXiv preprint arXiv:1902.00098*.

Eckert, P. (2017). Age as a sociolinguistic variable. *The handbook of sociolinguistics*, pages 151–167.

Flekova, L., Carpenter, J., Giorgi, S., Ungar, L., and Preoţiuc-Pietro, D. (2016). Analyzing biases in human perception of user age and gender from text. In *ACL*, pages 843–854.

Gill, A. J., Brockmann, C., and Oberlander, J. (2012). Perceptions of alignment and personality in generated dialogue. In *INLG*, pages 40–48.

Goldberg, L. R. (1993). The structure of phenotypic personality traits. *American psychologist*, 48(1):26.

Golovanov, S., Kurbanov, R., Nikolenko, S., Truskovskyi, K., Tselousov, A., and Wolf, T. (2019). Large-scale transfer learning for natural language generation. In *ACL*, pages 6053–6058.

Huang, M., Zhu, X., and Gao, J. (2019). Challenges in building intelligent open-domain dialog systems. *arXiv preprint arXiv:1905.05709*.

Jaech, A. and Ostendorf, M. (2018). Low-rank rnn adaptation for context-aware language modeling. *Transactions of the Association of Computational Linguistics*, 6:497–510.

Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv:1408.5882*.

Kottur, S., Wang, X., and Carvalho, V. (2017). Exploring personalized neural conversational models. In *IJCAI*, pages 3728–3734.

Lai, S., Xu, L., Liu, K., and Zhao, J. (2015). Recurrent convolutional neural networks for text classification. In *AAAI*, pages 2267–2273.

Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016a). A diversity-promoting objective function for neural conversation models. In *NAACL*.

Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, B. (2016b). A persona-based neural conversation model. In *ACL*, pages 994–1003.

Luo, L., Huang, W., Zeng, Q., Nie, Z., and Sun, X. (2019). Learning personalized end-to-end goal-oriented dialog. In *AAAI*.

Madotto, A., Lin, Z., Wu, C.-S., and Fung, P. (2019). Personalizing dialogue agents via meta-learning. In *ACL*.

Mairesse, F. and Walker, M. (2007). Personage: Personality generation for dialogue. In *ACL*, pages 496–503.

Mazaré, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training millions of personalized dialogue agents. In *EMNLP*.

Nguyen, D., Trieschnigg, D., Doğruöz, A. S., Gravel, R., Theune, M., Meder, T., and De Jong, F. (2014). Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *ACL*, pages 1950–1961.

Qian, Q., Huang, M., Zhao, H., Xu, J., and Zhu, X. (2018). Assigning personality/identity to a chatting machine for coherent conversation generation. In *IJCAI*.

Randolph, J. J. (2005). Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Ritter, A., Cherry, C., and Dolan, W. B. (2011). Data-driven response generation in social media. In *EMNLP*, pages 583–593.

Serban, I. V., Lowe, R., Charlin, L., and Pineau, J. (2015). A survey of available corpora for building data-driven dialogue systems. *Computer Science*, 33(16):6078–6093.

Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, volume 16, pages 3776–3784.

Shum, H.-y., He, X.-d., and Li, D. (2018). From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Song, H., Zhang, W.-N., Cui, Y., Wang, D., and Liu, T. (2019). Exploiting persona information for diverse generation of conversational responses. In *IJCAI*.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112.

Wang, J., Wang, X., Li, F., Xu, Z., Wang, Z., and Wang, B. (2017). Group linguistic bias aware neural response generation. In *SIGHAN*.

Wolf, T., Sanh, V., Chaumond, J., and Delangue, C. (2018). Transfertransfo: A transfer learning approach for neural network based conversational agents. In *NIPS2018 CAI Workshop*.

Yang, M., Zhao, Z., Zhao, W., Chen, X., Zhu, J., Zhou, L., and Cao, Z. (2017). Personalized response generation via domain adaptation. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1021–1024. ACM.

Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018). Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, pages 2204–2213.

Zhang, W.-N., Zhu, Q., Wang, Y., Zhao, Y., and Liu, T. (2017). Neural personalized response generation as domain adaptation. *World Wide Web*, pages 1–20.

Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.