Personalized Dialogue Generation with Diversified Traits

Yinhe Zheng[*]
Samsung R&D Institute of China -
Beijing (SRC-B)
Beijing, China
zhengyinhe1@163.com

Guanyi Chen
Utrecht University
Utrecht, Netherlands
g.chen@uu.nl

Minlie Huang[†]
Tsinghua University
Beijing, China
aihuang@mail.tsinghua.edu.cn

Song Liu
Samsung R&D Institute of China -
Beijing (SRC-B)
Beijing, China
s0101.liu@samsung.com

Xuan Zhu
Samsung R&D Institute of China -
Beijing (SRC-B)
Beijing, China
xuan.zhu@samsung.com

ABSTRACT

Endowing a dialogue system with particular personality traits is essential to deliver more human-like conversations. However, due to the challenge of embodying personality via language expression and the lack of large-scale persona-labeled dialogue data, this research problem is still far from well-studied. In this paper, we investigate the problem of incorporating explicit personality traits in dialogue generation to deliver personalized dialogues.

To this end, **firstly**, we construct PersonalDialog, a large-scale multi-turn dialogue dataset containing various traits from a large number of speakers. The dataset consists of 20.83M sessions and 56.25M utterances from 8.47M speakers. Each utterance is associated with a speaker who is marked with traits like Age, Gender, Location, Interest Tags, etc. Several anonymization schemes are designed to protect the privacy of each speaker. This large-scale dataset will facilitate not only the study of personalized dialogue generation, but also other researches on sociolinguistics or social science.

Secondly, to study how personality traits can be captured and addressed in dialogue generation, we propose persona-aware dialogue generation models within the sequence to sequence learning framework. Explicit personality traits (structured by key-value pairs) are embedded using a trait fusion module. During the decoding process, two techniques, namely *persona-aware attention* and *persona-aware bias*, are devised to capture and address trait-related information. Experiments demonstrate that our model is able to address proper traits in different contexts. Case studies also show interesting results for this challenging research problem.

KEYWORDS

Dialogue System, Dialogue Dataset, Personalization

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, so that the system can gain the user's confidence and trust [39]. Personality settings include

---

[*]Please contact zhengyinhe1@163.com for the PersonalDialog dataset
[†]Corresponding author

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 comfortable than fashionable.
(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. What about you?
(对啊在深圳几年了，你呢)
B: I just came to Shenzhen this year.
(今年刚来深圳)
A: No wonder, we would be a couple if we live closer before.
(怪不得，要是近一点说不定我们都在一起了)

Personality traits of A:

```
{ "age": "24",
  "gender": "Male",
  "location": "Guangdong"}
```

Personality traits of B:

```
{ "age": "23",
  "gender": "Female",
  "location": "Guangdong" }
```

Figure 1: An example dialogue session (translated) in our dataset. Several personality traits are given for each speaker. Words in response are in the same color with the corresponding traits.

age, gender, language, speaking style, level of knowledge, areas of expertise, or even a proper accent. The ability of exhibiting certain personality with diversified traits is essential for conversational systems to well interact with users in a more natural and coherent way [20, 23, 32].

Prior studies have demonstrated promising results for imitating a certain personal style in dialogue systems. Initial efforts focus on modeling characters in movie [3, 8]. Further developments propose to use a speaker embedding vector in neural models to capture the *implicit* speaking style of an individual speaker [20, 23, 31, 46, 48], or the style of a group of speakers [45]. Other approaches also attempt to endow dialogue models with personae which are described by natural language sentences [26, 47].

Recent studies on personalized neural conversational models can be broadly classified into two types: one is *implicit personalization* and the other is *explicit personalization*. In *implicit personalization models* [20, 23, 48], each speaker is represented by a user vector, and the vector is then fed into the decoder to capture the speaking style of the speaker implicitly. In spite of the simplicity and success of this technique, it is unclear how personality is captured and how it can be interpreted because all the information regarding to a user is encoded in a real-valued vector. Moreover, these methods also suffer from the data sparsity issue: each dialogue should be tagged with a speaker identifier and there should be a sufficient amount of dialogues from each speaker to train a reliable user-specific model. In *explicit personalization models*, the generated responses are conditioned either on a given personal profile [32], or on a text-described persona [47]. In these models, personality is presented specifically via key value pairs or natural language descriptions about age, gender, hobbies, etc. However, these methods are limited to either manually-labeled data or crowdsourced dialogues, thereby not scalable to large-scale dialogue datasets.

It is a matter of fact that the persona of a speaker can be viewed as a composite of diversified personality traits. During conversations, people may reveal their personality traits, consciously or unconsciously. For example, for the dialogue shown in 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 the third turns of this session, we can easily infer that both speaker *A* and *B* are living in Shenzhen (a city in Guangdong province, China). As exemplified, a personalized conversational agent should be equipped with diversified traits and be able to decide which personality traits to express in different contexts.

To address above issues, we propose a novel task and construct a large-scale dialogue corpus to study personalized dialogue generation. The task and corpus are unique in several aspects:

- **First**, the persona of each speaker in the corpus is presented by a number of personality traits, which are given explicitly in key-value pairs (as exemplified in Figure 1). Unlike implicit personalization models, such structured personae are more explicit, straight-forward, and interpretable. Moreover, since speakers with the same trait value (e.g., all females) can share their trait representations, the dialogue data across speakers can be shared to train a generation model, thereby the data sparsity issue is alleviated.
- **Second**, although the persona is represented explicitly, the use of such persona information can be captured implicitly by data-driven methods that are scalable to large-scale corpora. This differs from prior explicit personalization models [32] which

require that the given persona values must appear in a generated response and demand for manually-labeled data.
- **Third**, it is interesting to study how personality traits are expressed in dialogues and revealed via language expressions. In fact, the expression of persona via language is usually subtle and implicit[2]. For instance, a female speaker may not necessarily use the word "female" directly in every utterance she responds with, instead, she may consciously or unconsciously use related words that can reveal her gender in particular contexts. Therefore, it is worthy to build a personalized conversational system with the ability to exhibit specific traits in different contexts.

In this paper, we employ the sequence to sequence learning framework [41, 43] and devise a trait fusion module to capture the persona of each speaker in the response generation process. Specifically, each trait of a speaker is encoded as an embedding vector and different traits 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 an persona aware attention mechanism where the persona representation is used to generate the attention weights to obtain the context vector at each decoding position, and the second approach applies an persona-aware bias to estimate the word generation distribution. Automatic and manual evaluation indicate that our proposed models can incorporate proper, diversified traits when generating responses in different contexts.

Since there is no existing corpus to facilitate the aforementioned research task, we construct a large-scale dialogue dataset which contains various personality traits for a large number of speakers. Our dataset is collected from Weibo and contains about 20.83 million dialogue sessions (in Chinese) from about 8.47 million speakers. These dialogues cover a wide range of topics about our daily lives and consist of more than 3.43 million multi-turn sessions (each containing no less than 4 utterances). Various of personality traits are collected for each speaker and three of which are approached and evaluated in our model, namely Gender, Age, and Location. The proposed dataset will be useful not only for the study of dialogue systems, but also for other research topics such as pragmatics or sociolinguistics.

Our main contributions can be summarized as follows:

(1) We propose a new task to incorporate explicit personality traits into conversation generation. This task aims to study how explicit personality traits can be used to train a personalized dialogue model with large-scale, real social conversations.

(2) We construct a large-scale dialogue dataset that contains various traits of each speaker (such as Age, Gender, Location, Interest Tags etc.). To the best of our knowledge, this is the first dialogue corpus that contains real social conversations and diversified personality traits for each speaker. The proposed dataset will facilitate not only the study of personalized dialogue generation, but also other researches such as sociolinguistics.

(3) We propose persona-aware models which apply a trait fusion module in the encoder-decoder framework to capture and address personality traits in dialogue generation. We devise a persona-aware attention mechanism and persona-aware

bias to incorporate the persona information in the decoding process. Experiments demonstrate that our model is able to address proper traits in different contexts.

2 RELATED WORK

It has been demonstrated that personality is vital for building a human-like dialogue system [15, 39] which can exhibit a consistent persona. Personality settings such as age, gender, level of knowledge, and personal interests can be implicitly or explicitly expressed during the conversations [39]. In order to deliver more intelligent conversations, it is thus necessary to model these personality traits properly in a personalized conversational system.

There have been various prior studies for personalized dialogue generation. Traditional models are proposed to build personalized dialogue systems by modeling the "*Big Five*" [14]. This concept has been well defined in psychology [30], and is proved to be a stable personality evaluation metric [7]. Some personalized dialogues systems were built upon the basis of "*Big Five*", such as PERSONAGE [24, 25] and the work of Gill et al. [12]. However, such personality metric is extremely implicit and subtle in language expression, and thus challenging to be captured in dialogue generation [32]. Moreover, the dialogue data with "*Big Five*" annotation are extremely complex and expensive to collect. Therefore, "*Big Five*" is not suitable for building large-scale personalized dialogue systems, particularly with data-driven neural models.

Recently, the availability of large-scale dialogue corpora has significantly advanced the research of data-driven personalized dialogue models [20]. Some early studies focused on modeling characters in movie dialogues [3, 8], in which the presented "Character Style" usually depends on the scenes and plots of each movie. Further development of personalized dialogue generation models is inspired by the successful application of social media data [34, 35] and the sequence to sequence learning framework [36, 37, 40, 41, 43]. Specifically, Li et al. [23] represented each speaker with a persona vector and fed the vector to the decoder at each decoding step. The persona embedding is supposed to capture speaker-specified styles. Kottur et al. [20] extended this idea to multi-turn dialogues. In these models, the persona is implicitly represented by a single real-valued vector, which lacks interpretability.

In spite of the success of user embedding in above models, training these models requires abundant dialogue data from each speaker. When there are no such data available, it is unlikely to train a reliable model. A possible attempt to deal with this issue is to train personalized models with the gender attribute [45]. This approach helps to alleviate the data sparsity issue since the dialogue data within a group of same gendered speakers can be shared.

Note that personality traits in these embedding-based approaches are modeled implicitly. An initial attempt to incorporate explicitly represented persona is proposed by Qian et al. [32], in which a chatbot is endowed with a persona defined by a key-value table. A pair of forward and backward decoders are used to generate a response starting from a selected profile value (e.g., *female*), which ensures that a selected value must appear in a generated response. This approach requires manually-labeled data, and it may not be scalable to the large dialogue dataset as proposed in this paper. It is also

expensive to collect large-scale dialogue data via crowdsourcing services [47].

The construction of large-scale dialogue datasets is another important topic for recent research on dialogue systems. Serban et al. [35] presents a comprehensive summary of available dialogue datasets that can be used to construct a data-driven dialogue model. However, most of the existing corpora are not suitable for the study of personalized dialogue generation. Initial efforts collect dialogues from movie scripts [8, 44], with the annotations of *Character Styles*. Zhang et al. [47] crowd-sourced a dataset by asking randomly paired crowd workers to chat based on some given personae, however, this dataset is limited in its small size. The dataset introduced by Qian et al. [32] has a similar personality trait format (i.e., key-value pairs) with ours, and consists of manually-labeled data but only covers a small amount of patterns for a few traits, which is thus not scalable to large datasets. The dataset proposed by Joshi et al. [18] is constructed using limited templates and thus not suitable for dialogue generation tasks. We believe the dataset presented in this paper will offer new possibilities for studying personalized dialogue models with large-scale, real social conversation data.

3 MODEL

In order to capture diversified personality traits in the response generation process, we equip the general sequence to sequence model with a personality trait fusion module, which produces a persona representation v_p that can be incorporated into the decoder. In this study, two methods are proposed to utilize v_p in the decoding process, one is a persona-aware attention mechanism, and the other is a persona-aware bias. We will present the details in this section.

3.1 Task Definition and Overview

Our task can be formulated as follows: Given a post $X = x_1, x_2, ..., x_n$ and a set of traits $T = \{t_1, t_2, ..., t_N\}$ for the responder, the system should generate a response $Y = y_1, y_2, ..., y_m$ that embodies the personality traits in T:

$$Y^* = \arg\max_Y P(Y|X, T) \tag{1}$$

Where $x_i, (i = 1, 2, ..., n)$ and $y_i, (i = 1, 2, ..., m)$ are words. Note that each trait $t_i \in T$ is given as a key-value pair $t_i = \langle k_i, v_i \rangle$, and the exact values of $v_i(i = 1, 2, ..., N)$ are not required to appear in Y. Moreover, although the personality traits of the speaker who makes the post are also provided in our dataset, they are not modeled in our task. We leave this as future works.

An overview of our personalized dialogue generation model is shown in Figure 2. Given a trait set T, a personality trait fusion module is used to merge the traits in T into a persona representation $v_p \in \mathbb{R}^{d_p}$. Three approaches are proposed to fuse the personality traits. A sequence encoder is used to encode the post into a series of real-valued vectors $(h_1, h_2, ..., h_n)$, where $h_i \in \mathbb{R}^{d_s}$. Two methods are proposed 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.
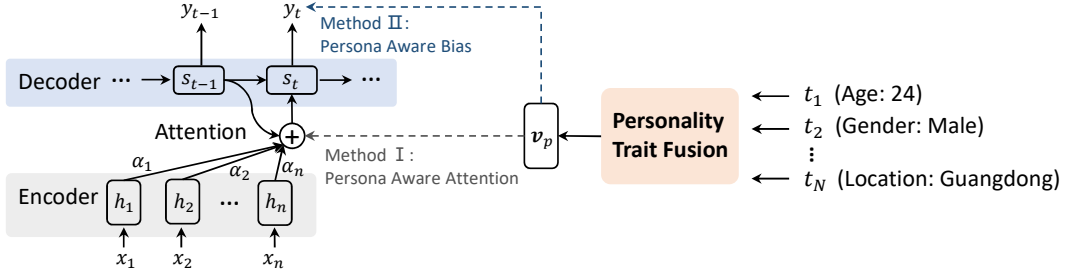
Figure 2: Overview of personalized dialogue generation model. To obtain the persona representation $\boldsymbol{v}_p$, different traits are integrated by the personality trait fusion component. $\boldsymbol{v}_p$ is then used to generate persona-aware attention weights for computing the context vector, or to produce a persona-aware bias for computing the generation distribution.

3.2 Sequence to Sequence Framework

The backbone of our model is the sequence to sequence (Seq2Seq) learning framework [41, 43], which is commonly used in language generation tasks such as machine translation and dialogue generation. A typical Seq2Seq model usually consists of two components: an encoder and a decoder. For dialogue generation tasks, the encoder takes the post $X = x_1, x_2, ..., x_n$ as input and encodes X into a sequence of vectors $(\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n)$, $\boldsymbol{h}_i \in \mathbb{R}^{d_s}$. 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. [1], 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 $\boldsymbol{h}_1, \boldsymbol{h}_2, ..., \boldsymbol{h}_n$, at the same time conditioned on the preceding state of the decoder $\boldsymbol{s}_{t-1} \in \mathbb{R}^{d_s}$. Formally, we have:

$$\boldsymbol{c}_t = \sum_{i=1}^{n} \alpha_i \boldsymbol{h}_i,$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{n} \exp(e_j)} \tag{2}$$

$$e_i = \text{MLP}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i)$$

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

where $V \in \mathbb{R}^{d_s}$, $W_\alpha^1 \in \mathbb{R}^{d_s \times d_s}$ and $W_\alpha^2 \in \mathbb{R}^{d_s \times d_s}$ are parameters for the attention mechanism.

In general Seq2Seq models, the output probability y_t at step t of the decoder is produced by a softmax function:

$$y_t = \text{softmax}(W_o^1 \boldsymbol{s}_t + b_{out}),$$

$$\boldsymbol{s}_t = \text{RNN}(\boldsymbol{s}_{t-1}, \boldsymbol{c}_t, w_{t-1}). \tag{3}$$

where w_{t-1} is the word vector of the decoded word from previous time step. $W_o^1 \in \mathbb{R}^{|V| \times d_s}$ and $b_{out} \in \mathbb{R}^{|V|}$ are parameters for the decoder ($|V|$ is the vocabulary size).

In this study, the encoder we use is a two-layer bi-directional RNN with gated recurrent units (GRU) [6], and the decoder is also a two-layer GRU.

3.3 Personality Trait Fusion

In our personalized dialogue model, we first compute an integrated persona representation $\boldsymbol{v}_p$ and then use $\boldsymbol{v}_p$ to affect the decoding process. The construction of $\boldsymbol{v}_p$ starts with mapping each traits t_i in T to an embedding representation $\boldsymbol{v}_{t_i}$ using its corresponding trait encoder. Note that traits considered in this study (i.e., Age, Gender and Location) are all single-valued, i.e., each trait only has one unique value for each speaker. Therefore these trait encoders can be implemented using look-up tables. Actually, other categories of traits can also be modeled if a proper encoder is provided. For instance, an LSTM encoder can be applied to represent a one-sentence self-description of a speaker.

After encoding all the traits in T into a set of trait representations $\{\boldsymbol{v}_{t_1}, \boldsymbol{v}_{t_2}, ..., \boldsymbol{v}_{t_N}\}$, we can merge them using a *personality trait fusion function* to obtain the persona representation $\boldsymbol{v}_p$. In this paper, three different fusion methods are investigated.

*3.3.1 **Traits Attention**.* Merge all the trait representations in T based on an attention mechanism. Specifically, given the hidden state from the previous decoding step $\boldsymbol{s}_{t-1}$, an attention weight α_i' is computed for each trait. Then, $\boldsymbol{v}_p$ is obtained as a weighted sum of all the trait representations:

$$\boldsymbol{v}_p = \sum_{i=1}^{N} \alpha_i' \boldsymbol{v}_{t_i}$$

$$\alpha_i' = \frac{\exp(e_i')}{\sum_{j}^{N} \exp(e_j')} \tag{4}$$

$$e_i' = \text{MLP}(\boldsymbol{s}_{t-1}, \boldsymbol{v}_{t_i})$$

$$= \overline{V}^T \cdot \tanh(\overline{W_\alpha^1} \boldsymbol{s}_{t-1} + \overline{W_\alpha^2} \boldsymbol{v}_{t_i})$$

where $\overline{V} \in \mathbb{R}^{d_s}$, $\overline{W_\alpha^1} \in \mathbb{R}^{d_s \times d_s}$ and $\overline{W_\alpha^2} \in \mathbb{R}^{d_s \times d_p}$ are parameters for the trait attention mechanism. The calculated weight α_i' indicates how much the current context favors trait t_i. The trait attention mechanism here allows us to make proper combination of personality traits with respect to the contexts.

*3.3.2 **Traits Average**.* Average all the trait representations in T:

$$\boldsymbol{v}_p = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{v}_{t_i} \tag{5}$$

This is a special case of **Traits Attention**, where the traits in T are weighted equally.

3.3.3 **Traits Concatenation**. Concatenate all the trait representations in T to produce $\boldsymbol{v}_p$. Note that in this case the length of $\boldsymbol{v}_p$, i.e., d_p should be divisible by N and the length of each trait representation vector $\boldsymbol{v}_{t_i}$ should be d_p/N.

3.4 Decoding with Persona Representation

In order to incorporate the persona representation $\boldsymbol{v}_p$ in our decoder, we develop the following methods:

3.4.1 **Persona-Aware Attention (PAA).** : The first method extends the computation of attention weights (Equation 2) used in the decoder. The attention weight is now dependent on not only the decoder's state, but also the persona representation $\boldsymbol{v}_p$, namely,

$$
\begin{aligned}
e_i &= \text{MLP}(\boldsymbol{s}_{t-1}, \boldsymbol{h}_i, \boldsymbol{v}_p) \\
&= V \cdot \tanh(W_\alpha^1 \boldsymbol{s}_{t-1} + W_\alpha^2 \boldsymbol{h}_i + W_\alpha^3 \boldsymbol{v}_p)
\end{aligned}
\tag{6}
$$

where $V \in \mathbb{R}^{d_s}$, $W_\alpha^1 \in \mathbb{R}^{d_s \times d_s}$, $W_\alpha^2 \in \mathbb{R}^{d_s \times d_s}$, and $W_\alpha^3 \in \mathbb{R}^{d_s \times d_p}$ are learnable parameters. The score e_i is the input to the softmax function for computing the attention weight. This approach can help our decoder to attend to different contexts based on the persona representation, which is termed persona-aware attention mechanism.

3.4.2 **Persona-Aware Bias (PAB).** : The second method tries to incorporate $\boldsymbol{v}_p$ in the output layer of the decoder. Specifically, we extend Equation 3 to include a persona bias to obtain the generation distribution. A gate is devised to balance the original term and the persona bias term, as follows:

$$
\begin{aligned}
y_t &= \text{softmax}(a_t \cdot W_o^1 \boldsymbol{s}_t + (1 - a_t) \cdot W_o^2 \boldsymbol{v}_p + b_{out}) \\
a_t &= \sigma(V_o^T \cdot \boldsymbol{s}_t)
\end{aligned}
\tag{7}
$$

where $W_o^1 \in \mathbb{R}^{|V| \times d_s}$, $W_o^2 \in \mathbb{R}^{|V| \times d_p}$, $V_o \in \mathbb{R}^{d_s}$ and $b_{out} \in \mathbb{R}^{|V|}$ are learnable parameters. Note that although the bias brought by $\boldsymbol{v}_p$ seems to be context independent (i.e., it may select words independently at each decoding step), the computed scalar variable $a_t \in [0, 1]$ works as a gate to control how much persona related features should be incorporated at each time step t. It can decide whether to use trait related word or semantic related word, and thus makes the response generation process more consistent.

As can be seen, the persona-aware bias is assumed to be more direct in influencing the generation distribution, which is verified by experiment results shown in §5: PAB works generally better than PAA. Similar model structures have also been used in the work of Jaech and Ostendorf [17] and Zhou et al. [49], and promising results have achieved.

4 PERSONALDIALOG DATASET

The dialogue dataset that we construct for the proposed task, named as `PersonalDialog`, involves a large amount of speakers with a wide variety of personality traits. The data in `PersonalDialog` are collected from Weibo[1], one of the largest Chinese social media. In fact, according to the theories in sociolinguistics, people tend to perform specific personae when they use language to socialize [13,

[1]www.weibo.com

38]. Therefore, social media becomes an ideal source to collect large-scale dialogues with diversified personality traits. The features and statistics of `PersonalDialog` are detailed in this section together with a brief introduction to the data collection process.

4.1 Features and Statistics

Dialogues in our dataset are composed of Weibo posts and their comments. Specifically, when a user post a Weibo message, other users may comment on it, which may receive further comments. It forms a tree structure which is rooted at the original Weibo post. We regard an original post and one branch of its comments as a dialogue session. These dialogues are collected along with the publicly available personality traits of each speaker. Some attractive features of `PersonalDialog` are presented in this section.

4.1.1 *Personality Traits.* The most important and appealing property of `PersonalDialog` is the personality traits collected for each speaker, which are provided by speakers themselves on Weibo. Various interesting tasks can be investigated with the help of these information, such as personalized dialogue generation, text style transfer or text-based personality analysis.
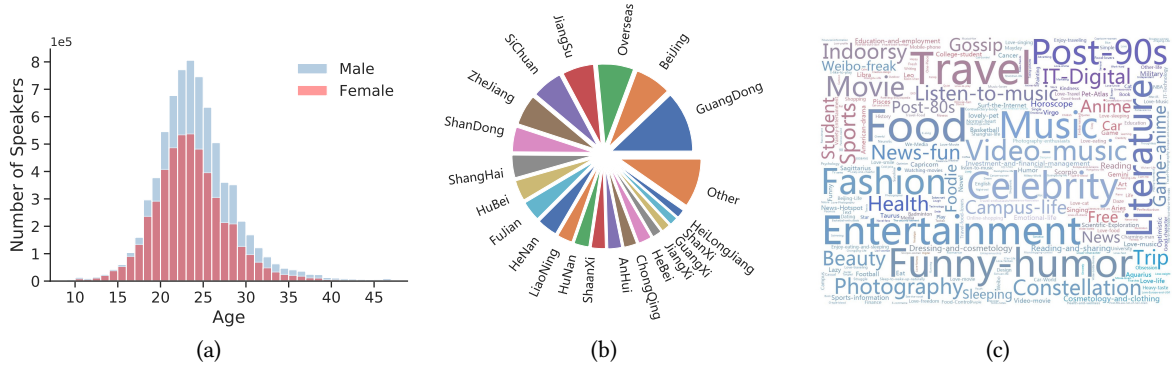
Table 1: Statistics of personality traits in `PersonalDialog`.

Total number of speakers	8.47M
Total number of interest tags	39.6K
Number of interest tags per speaker	2.187
Average speaker age	25.23
Average length for self descriptions	10.09

Each speaker presented in our dataset has five personality traits: Gender, Age, Location, Interest Tags, and Self Description. Specifically, Gender is a binary-valued trait, i.e., the gender of a speaker can be either "Male" or "Female"; Age is represented by an integer ranging from 8 to 48. Our observation indicates that Age values out of this range are very likely to be "fake", i.e., some Weibo users prefer to not to reveal their true ages by providing unreasonable birthdays. Therefore, a speaker with an age out of this range is reserved in our dataset but is given an empty Age value. Location is the province or urban district indicating where the speaker comes from. This trait has 35 different values that cover all the areas of China. Interest Tags is a set of keywords indicating the speaker's hobbies and interests. Each speaker may provides several different tags. In order to reduce the noise of the collected dataset, tags that are shared by less than 10 speakers are ignored in `PersonalDialog`; Self Description contains some self-provided description utterances of each speaker. It could be his/her quotations or biography; Basic statistics of these personality traits are shown in Table 1 and Figure 3.

Note that our data collection process strictly follows the privacy setting of Weibo. All these five personality traits collected in our dataset are publicly available on Weibo. We believe these traits are closely related to speakers' personae in dialogues and speakers contained in our dataset cannot be traced based on the trait information.

Figure 3: Statistics of personality traits. (a) Distributions of `Age` and `Gender` traits. The red and blue bars correspond to female and male speakers, respectively; (b) Distributions of top 21 frequent `Locations` (provinces); (c) Word cloud visualization of top 250 frequent `Interest Tags` (translated). The top 10 frequent tags are "Travel", "Food", "Entertainment", "Funny-humor", "Celebrity", "Music", "Fashion", "Literature", "Video-music" and "Post-90s".

4.1.2 Corpus Size. In addition to rich personality traits, another appealing feature of `PersonalDialog` is its large size. Table 2 presents a basic statistic of dialogues in `PersonalDialog` where there are 20.83M dialogues and 56.25M utterances.

Table 2: Statistics of dialogues in `PersonalDialog`.

Total dialogues	20.83 M
Total 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

Another advantage of `PersonalDialog` involves the length of dialogue session. A considerable amount of dialogues (3.43M sessions) in `PersonalDialog` have multiple turns of conversations. These dialogues can facilitate the research on multi-turn open-domain dialogue systems. To the best of our knowledge, there is still no such publicly available corpus.

4.1.3 One-to-Many in Dialogue Generation. Different from machine translation where two sentences from different languages are equivalent in semantic, dialogue generation is essentially a one-to-many mapping problem: for a same post, there are many possible responses dependent on the context, scene, emotional mood, and many other factors. `PersonalDialog` offers an opportunity to study this challenging research problem since most of existing open-domain dialogue corpora do not contain multiple responses to a post or are of limited scale. Actually, more than 2M posts have at least two replies in our dataset. We believe `PersonalDialog` will facilitate further studies on developing conversational agents that are able to generate diversified responses.

4.1.4 Sociolinguistics Phenomena. `PersonalDialog` presents a large amount of informal dialogue contents generated in *computer mediated communications* [16]. Together with diversified personality traits, our dataset can facilitate the study of language usages in *computational sociolinguistics* and help to build key components

for such research [28]. In addition, comparing to crowd-sourced corpora, conventional corpora that are collected from social media carry rich social meanings corresponding to each speaker, where the large size of our dataset makes it more feasible to be used in such research. Therefore, `PersonalDialog` might become a good choice for computational sociolinguistics research.

In fact, in this work, we have explored a preliminary application of `PersonalDialog` on computational sociolinguistics: the detection of social identities [28]. Specifically, trait classifiers are devised (introduced in §4.3) to predict the gender, age and location of social media users based on users' Weibo posts. Our classifiers achieve reasonable performance and the corpus facilitates further studies in this direction. In addition, our dataset can also facilitate the modeling of dialectal variations [9] as well as syntactic and pragmatic variations with respected to `Age`, `Gender`, `Location`, or a mixture of these traits.

It is also worth noting that dialogue datasets used in traditional sociolinguistics researches are usually collected in a way that each speaker explicitly indicates his/her audiences. However, `PersonalDialog` provides a very different settings because a Weibo post usually do not specify a particular audience, which provides us a chance to validate the findings of prior sociolinguistics studies on new units of analysis [42].

4.2 Data Collection and Filtering

Our data collection process was separated into two stages to have a smooth initiation and avoid collecting posts from spammers. The first stage collected seed users who commented under some manually chosen Weibo accounts that were specialized at posting news and maintained by dedicated staffs from mass media. The collected seed users were further filtered based on some user statistics such as the number of followers, posts, and followees. About 300k seed users were resulted. The second stage involved collecting Weibo messages posted by these seed users, together with the received comments and personality traits of each commenting user. Note that a tree structure can be constructed based on the *reply-to* relations between these collected comments, and a dialogue session

can be obtained by traversing a path from a root comment to each leaf comment. Finally, about 60 million sessions of raw dialogues and 12 million speakers were obtained.

Several pre-processing steps were used to clean these raw dialogues. We first eliminated the dialogues that contained abusive utterances based on a pre-defined abusive word list (containing 3,089 abusive words). A session was discarded if it contained an abusive utterance. Then, all the utterances were tokenized using jieba[2]. The sessions containing utterances that were too short (less than 3 tokens), too long (more than 40 tokens) or with only stop words were discarded. We also applied some rules to further reduce the noise, such as removing consecutive punctuations and emojis, and truncating dialogues at the utterances that contained only emojis, punctuations, Latin characters, or external links.
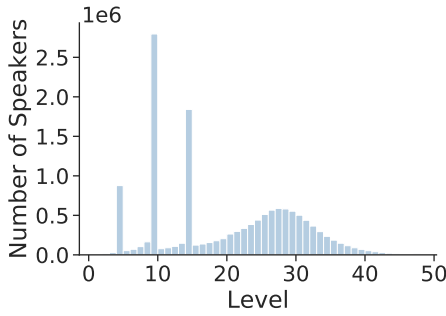
Figure 4: Distribution of the activeness level of collected Weibo users.

Another pre-processing step of our data was to filter spammers. In fact, spammer detection is a quite challenging task in social media analysis. It is not our target in this paper to discuss how to accurately detect spammer utilizing the information we have collected. However, what we do care is to ensure the data presented in `PersonalDialog` are produced by normal human users. Fortunately, the distribution of users' activeness level (shown in Figure 4) on Weibo sheds a light on this task. The level of a user on Weibo is an indicator of his/her activeness, i.e., a user must be active enough to obtain a high level. [3]. It is interesting to note that there are anomaly peaks at level 4, 9 and 14 in Figure 4. This may dues to the strict "upgrade" rule introduced by Weibo. Specifically, a user has to meet extra requirements (e.g. follows or being followed by a specific number of users) to upgrade in these levels. We argue that it is hard for most spammers to bypass these levels because it will increase the cost of spamming sharply. We further argue that most spammers are located under level 15, and users with levels higher than 15 are more likely to be regular users. Therefore, dialogues that came from speakers whose activeness level were under 15 were abandoned in `PersonalDialog`.

4.3 Personality Trait Classifiers

In order to take full advantages of personality traits in personalized dialogue generation, we need to ensure that dialogues collected in `PersonalDialog` indeed carry trait-related features. A natural

[2]http://github.com/fxsjy/jieba
[3]A detailed explanation for the level system of Weibo can be found here: http://level.account.weibo.com/level/levelexplain

idea to demonstrate this is to predict the value of personality traits based on dialogues. In particular, to build trait classifiers that take in dialogue texts and predict the value of each trait associated with each speaker. The constructed trait classifiers can also be used to evaluate our generation models. Namely, we can determine whether the generated responses reveal certain personality traits using these classifiers.

To this end, three trait classifiers were built, i.e. classifiers for `Gender`, `Age`, and `Location`, respectively. Ideally, these classifiers should be able to identify speakers' `Gender`, `Age`, and `Location` based on the dialogues they issued. In the following sections, we will present details of the data preparation process and classification models.

4.3.1 Data Preparation. A naive approach to construct the trait classifier is to predict trait values based on each individual Weibo post (utterance). However, according to a series of crowd-sourcing experiments done by Nguyen et al. [29], speakers may not reveal their persona in each single utterance. Therefore, training trait classifiers using a single utterance as an input generally produces suboptimal performance.

In order to alleviate above issues, we argue that the value of each trait carried in dialogue texts should be judged based on a set of utterances rather than a single utterance. In this study, we used a concatenation of n utterances as an input to the trait classifier. Specifically, for a given trait, we concatenate every n utterances issued by speakers with a same trait label and use the concatenated texts as inputs to that trait classifier. Obviously, the concatenated texts contain richer information about that trait. This is a commonly used strategy in trait perception tasks performed on social media data [11, 27].

More specifically, taking the gender classifier as an example: since there are only two labels for `Gender` ("Male" and "Female"), we first construct two sets of utterances that are issued by males and females respectively, and then concatenate every n utterances in each set as an input to the gender classifier. We randomly sample 50K such inputs for validation and test, respectively, and use the rest of these inputs for training. Data for other trait classifiers are similarly processed.

Note that the performance of the constructed classifier will be affected by the value of n. In fact, the more utterances contained in an input (if n is large), the more evidences will be provided to the classifier to make a decision, which generally leads to higher performance. We have tested different choose of n and found that $n = 20$ yields plausible performance in most cases. The accuracy increase will be less than 1% if $n > 20$ for all trait classifiers. Therefore we used $n = 20$ for all our classifiers.

Table 3: Statistics of the datasets (with balanced labels) used to build trait classifiers.

Trait Type	Train	Validation	Test	Class Num.
Age	2.0M	150K	150K	4
Gender	5.1M	65K	65K	2
Location	2.9M	75K	75K	10

Moreover, labels used for each classifier deserve further discussion. For Gender classifier, only two labels were used: "Male" and "Female". Users who did not provide their gender on Weibo were omitted for constructing Gender classifier. For Age classifier, four labels were used, i.e., we grouped the values of Age into four ranges: "post-70s" (born within 1970-1079), "post-80s"(1980-1989), "post-90s" (1990-1999), and "post-00s" (born after 2000). This simplification was made because previous studies indicate that it was impractical to predict the exact Age of a speaker with only text from social media [10]. Users with Age values that were out of this range were not used to produce our Age classifier. Similar strategy was used for Location classifier, where ten Location labels were chosen based on the theory of geolinguistics about the dialect area distribution of Chinese [5]. Instead of predicting exact province, we assigned identical labels for provinces (or districts) corresponding to similar dialect areas.

Table 4: Accuracy of trait classifiers.

Model	Gender	Age	Location
CNN	89.71	76.84	61.02
LSTM	90.23	78.02	61.69
RCNN	**90.61**	**78.32**	**62.04**

Note that datasets constructed following above settings may suffer from label imbalance issue, i.e. some labels may have remarkably more instances than others. In order to be consistent with the model evaluation scheme presented in §5.4, The datasets used to train, validate and test each trait classifier were balanced using the random minority oversampling approach [4], i.e., instances from minor labels were repeatedly up-sampled. Statistics of the datasets that were used to train each classifier were shown in Table 3.

4.3.2 Classification Models. We trained several classifiers using the constructed datasets, including CNN [19], LSTM (the states of every time steps of the LSTM layer are averaged before feeding them into a fully connected layer), and RCNN [21] (the outputs of LSTM states are feed into a CNN layer). Results in Table 4 show that RCNN achieves the best performance. Therefore, it is used as the classification model in subsequent automatic evaluation of dialogue generation.

The filter sizes used in CNN and RCNN are 2, 3, 4. The feature size for each filter is 128. The hidden size of LSTM and RCNN is 265. Word embeding size is 100 and these models are trained with a dropout rate of 0.8. Note that performances of these classifiers are not sensitive to the choose of hyper-parameters.

5 EXPERIMENTS

5.1 Data Preparation for Experiments

To evaluate the influence of personality traits on dialogue generation, we performed single-turn dialogue generation[4]. To this end, we used 10M sessions of single-turn dialogues (post-response pairs) extracted from PersonalDialog. Three personality traits were considered in our models: Gender, Age, and Location. Similar to the

data preparation process presented in §4.3.1, we only considered coarse labels for Age and Location, i.e., 4 labels for Age and 10 labels for Location. We randomly sampled 20,000 dialogue sessions for validation.

To test how well our model can utilize diversified personality traits in different contexts, we constructed four test sets: unbiased set, gender-biased, age-biased, and location-biased set, each of which included 10,000 dialogue sessions. Biased test sets provided us different contexts under which human speakers tend to reveal certain personality trait 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 will be interesting to see whether our model can learn to incorporate these traits in the generated dialogues under these biased contexts. In this study, dialogues in the unbiased set were randomly sampled. Whereas dialogues in the biased sets were deliberately selected to contain biased responses that carry obvious features related to each trait (like the Gender biased response "I am a girl" given in above example). Specifically, the trait label of a biased response should be correctly predicted with high confidence score using associated trait classifier.

The construction process of our biased sets would be straight forward if we had a classifier that toke single response utterances as inputs, i.e., we could feed each individual response to that classifier and selected the correctly predicted responses that had high confidence scores (i.e., maximum value of Softmax outputs) as biased responses. However, as discussed in section 4.3, our trait classifier toke a concatenation of 20 utterances as an input because not all utterances collected on social-network carry trait-related features. It means that we cannot directly calculate the confidence score for each individual response utterance using our classifier. Moreover, if an input (a concatenation of 20 response utterances) was correctly predicted with high confidence score by our classifier, not all response utterances constituting that input carried trait-related features.

In order to solve above issues, we argue that if a response utterance r is biased, then the input that contains r is more likely to be correctly classified with higher confidence score compared to the input that do not contain r. Therefore, we define the confidence score $c(r)$ of each individual response utterance r as the averaged confidence score of all possible inputs that contain r. If $c(r)$ is high, it is more likely for the input that contains r to be correctly classified, i.e., for r to be biased.

Apparently, it is unpractical to compute $c(r)$ precisely using its definition. In this study, we obtained an approximation of $c(r)$. Specifically, for a given trait, e.g. Gender, we randomly sampled N post-response pairs (p_i, r_i), $i = 1, 2, ..., N$, and for each response sentence $r \in \{r_i, i = 1, 2, ..., N\}$, we constructed m classifier inputs $S_j(r)$, $j = 1, 2, ..., m$ containing r. Specifically, each input $S_j(r)$ was a concatenation of 20 response sentences, in which r was contained. Assuming r was issued by a female speaker, then the rest 19 response sentences constituting $S_j(r)$ were randomly sampled female-issued responses from $\{r_i, i = 1, 2, ..., N\}$. We fed $S_j(r)$,

---

[4]Multi-turn dialogue generation can be considered in our framework by encoding additional contexts, and we leave it as the future work.

$j = 1, 2, ..., m$ into our Gender classifier and calculated the approximated confidence score for r as:

$$c'(r) = \frac{1}{m} \sum_{j=1}^{m} \delta(S_j(r))P(S_j(r)) \quad (8)$$

in which $P(S_j(r)) \in [0, 1]$ was the confidence score produced by our Gender classifier when processing $S_j(r)$. $\delta(S_j(r))$ was set to 1 if the label of $S_j(r)$ was correctly predicated, and set to -1 otherwise. In our experiments, we use $N = 50,000$ and $m = 1,000$. The top 10,000 high scored responses were selected as biased responses and the corresponding post-response pairs were used as the biased set.

We tested each biased set by concatenating every 20 response utterances with a same trait label and feeding these concatenations to our classifier. The classification accuracy of these constructed inputs are reported in the last row of Table 6. These high accuracy scores indicate that the responses contained in each biased set indeed carry rich trait related features, i.e., can be correctly classified more easily.

Note that we have also tested higher values for N and m, which are certainly beneficial for obtaining a more accuracy confidence score. However, these accuracy scores shown in the last row of Table 6 are near-perfect. So we decide to use $N = 50,000$ and $m = 1,000$ until we can manage to get a better trait classifier.

5.2 Implementation Details

We implemented our model and tuned all the hyper-parameters on the validation set. Specifically, the encoder and decoder are 2-layer GRUs with 512 hidden units for each layer. We set the word vocabulary size to 40,000 and the dimension of word vectors to 100. The word vectors are updated during the training process and shared by the encoder and decoder. The embedding size of the persona representation $\boldsymbol{v}_p$ is set to 100. The Adam optimizer is used to train our model with a batch size of 120 and a learning rate of 0.001. The training process of each model took about a week on a Titan X GPU machine.

5.3 Baselines

We chose several baselines:

- A Seq2Seq model, which does not use any persona features.
- Three Group Linguistic Bias Aware (GLBA) models [45], which respectively incorporate three individual personality traits, namely Gender, Age, and Location.

We implemented several variants of our proposed model with different combinations of trait fusion methods and decoding schemes. Three trait fusion methods are attention-based (Att. see §3.3.1), average fusion (Avg. see §3.3.2), and concatenation (Concat. see §3.3.3). Two decoding schemes are Persona-Aware Attention (PAA, see §3.4.1) and Persona-Aware Bias (PAB, see §3.4.2). As a result, six variants were tested.

Note that we did not adopt the speaker model introduced by Li et al. [23] as our baseline because it requires a large amount of dialogues for each speaker to train a reliable model. Furthermore, we also did not adopt a variation of the speaker model, where speaker embedding is replaced by trait embedding. In fact, the gated bias approach used in GLBA models generally outperforms using trait

embedding in the speaker model[17]. In other words, GLBA models are stronger baselines for our task.

Table 5: Automatic evaluation on the unbiased test set with *perplexity* (ppx.), *distinct-1* (dist1), *distinct-2* (dist2) and *trait accuracy* (acc.).

Model	ppx.	dist1	dist2	Gender acc.	Age acc.	Loc. acc.
Seq2Seq	84.07	0.0226	0.0599	50.2	25.3	10.2
Gender 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
Location 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
Concat. + 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
Concat. + 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**

Table 6: Automatic evaluation on biased test sets with *trait accuracy* (acc.). The *trait accuracy* is obtained using the corresponding biased test set with the originally provided persona traits.

Model	Gender acc.	Age acc.	Location acc.
Seq2Seq	85.3	79.8	27.2
Gender GLBA	95.5	81.6	31.8
Age GLBA	86.8	92.1	32.0
Location GLBA	87.3	78.2	48.2
Avg. + PAA	91.1	88.5	43.3
Concat. + PAA	91.7	88.9	44.5
Att. + PAA	94.0	88.3	42.5
Avg. + PAB	94.8	91.9	48.5
Concat. + 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] This score shows the trait accuracy obtained using golden (human-generated) responses in the biased test sets.

5.4 Automatic Evaluation

We performed automatic evaluation to verify whether our model can incorporate diversified personality traits in dialogue generation.

5.4.1 Metrics. Perplexity was used to evaluate our model at the content level. Smaller perplexity scores indicate that the model can generate more grammatical and fluent responses. We also used *Distinct*[22] to evaluate the diversity of the generated responses. To evaluate how well our models can capture personality traits, we defined *trait accuracy* as the agreement between the expected trait values (i.e., inputs to the personality trait fusion model) and the trait labels predicted by the trait classifiers. A higher *trait accuracy*

indicates a stronger ability to incorporate that trait. For example, for the Gender trait, if a set of responses were generated with a "female" label, we were expecting these responses can be easily classified to "female" by our Gender classifier. Therefore, in order to calculate the Gender accuracy, we first generated responses with different Gender values, and then we followed the process introduced in section 4.3.1 to construct classifier inputs using these generated responses. The classification accuracy of these inputs using our Gender classifier was obtained as the Gender accuracy. Note that in this process, the values of other traits were identical to those of the responder in the test set.

5.4.2 Results. Table 5 shows the performance of each model on the unbiased test set. The *trait accuracy* shown in this table was obtained by assigning the target trait with different values. For example, for the Gender trait, we generated two sets of responses to the same posts with "Female" and "Male" label, respectively. Moreover, in order to investigate the behaviour of our models in different contexts, we also tested our models on three biased test sets (Table 6). The *trait accuracy* shown in Table 6 was obtained by assigning the same traits to those of the responder in the test set, i.e., when generating the response for a given post, we provide our models with same traits as the responder in the biased set. We also listed the *trait accuracy* calculated using the responses generated by actual human speakers in the last row of Table 6. These scores can be used as the upper-bound of the generation models.

Results in these tables show that:

- The models equipped with PAB generally outperform the models with PAA on all the metrics. This may be attributed to the fact that PAB can influence the decoding process more directly.
- GLBA models only perform well on single trait. For example, the Gender GLBA model only achieves good *trait accuracy* regarding to Gender, whereas it degrades remarkably on Age and Location compared to our models. In comparison, our models achieve higher *trait accuracy* with respect to all the traits. This verifies that the trait fusion module is necessary to model diversified traits in different contexts.
- The model equipped with trait attention (Att.) and PAB obtains the best performance in terms of almost all the metrics, particularly on the biased test sets. This indicates that the trait attention approach facilitates the modeling of diversified traits and it also helps to choose proper traits in different contexts.

5.5 Manual Evaluation

In order to further evaluate the performance of our models, we performed manual evaluation. Given a post and the personality traits of a responder, we generated responses from all the baseline models and our best performing model (Att. + PAB). These responses were presented to three human annotators along with the specified personality traits and the post.

5.5.1 Metrics. Annotators were asked to score a response in terms of two aspects with a five-star scale (1 means not good or not true, while 5 means excellent):

(1) *Fluency*: How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?

Table 7: Manual evaluation with *Fluency* (Flu.) and *Appropriateness* (App.).

Model	Flu.	App.
Seq2Seq	4.685	3.889
Gender GLBA	4.732	3.850
Age GLBA	4.792	3.898
Location GLBA	4.730	3.707
Att. + PAB	**4.822**	**3.971**

(2) *Appropriateness*: Do you think the usage of personality traits in the generated response is logical and meet the common practice of a native speaker in daily communications?

5.5.2 Annotation Statistics. 100 posts were randomly sampled from each of these four test sets (400 posts in total), and 2,000 responses were generated using five models. The inter-rater consistency of the annotation results were measured using the Fleiss' kappa κ [33]. In particular, the κ value for *Fluency* and *Appropriateness* was 0.82 and 0.53, respectively, indicating fairly good agreements between these annotation results.

5.5.3 Results. The results are shown in Table 7. Our model outperforms all our baselines significantly (t-test, $p - value < 0.05$) in both metrics. This indicates that diversified personality traits help to generate more fluent and appropriate responses and our model can learn to incorporate proper personality traits in generated responses.

It is also interesting to see that the Seq2Seq model outperforms some GLBA models in *Appropriateness*. We argue that these GLBA models try to emphasize single personality trait in each utterance they generate, resulting in sub-optimal performance in producing logical and appropriate responses. In fact, different traits should be embodied in different contexts, and sometimes we do not need to address any trait in the response.

5.6 Case Study

Some sampled cases are shown in Table 8. Words in responses are in the same color with the corresponding traits. It can be seen that our model can generate responses incorporating certain traits and can choose proper personality traits for different contexts, whereas the Seq2seq model tends to generate universal responses and the GLBA model can only consider a single trait. For the first post, both our model and the Gender GLBA model incorporate proper trait (i.e., Gender) in the generated responses, i.e., these models can act as either "Male" (colored in blue) or "Female" (colored in orange) when generating responses, while other models can only generate responses with random traits or universal responses. Furthermore, responses to the second post are associated with the Age trait, which is usually expressed in a more implicit way. Our model and the Age GLBA model can incorporate stylistic features related to Age. Specifically, an elder agent ("Post-70s", colored in blue) tends to use rigorous and formal expressions whereas a younger agent ("Post-90s", colored in orange) uses casual and informal phrases.

The visualization of trait attention scores further proves that the trait fusion module helps to model diversified personality traits. As shown in Figure 5, when decoding the first four words (i.e.,

Table 8: Sample responses generated by baselines and our model (Att. + PAB). Words in response are in the same color with the corresponding traits.

Post	Traits for Resp.	Model	Response	Translated Responses
我的理想男友。 Ideal boyfriend for me	Post-90s Male Haiwai	Seq2Seq	给你一个么么哒。	Give you a kiss
		Location GLBA	你是我的小男友。	You are my little boyfriend
		Age GLBA	你的男朋友是我的！	Your boyfriend is mine!
		Gender GLBA	你是我的小情人！	You are my little lover!
		Att. + PAB	你是我的优乐美	You are my beloved
	Post-90s Female Haiwai	Att. + PAB	他不只是你的理想男友，呵呵	He is not just your ideal boyfriend, LOL
		Gender GLBA	他是我的	He is mine
好美的景色。 The view is so beautiful.	Post-70s Male Beijing	Seq2Seq	谢谢喜欢。快乐。	Thanks for your liking. Enjoy.
		Location GLBA	嗯嗯，很美。	Emm, it is beautiful.
		Age GLBA	谢谢好友美评夸奖！	Thanks for my friend's appreciation and praise!
		Gender GLBA	是啊，很美！	Yeah, it is beautiful!
		Att. + PAB	谢谢好友美评，晚上好	Thanks for my friend's appreciation, good evening
	Post-90s Male Beijing	Att. + PAB	有机会来玩呀	Come on, pay a visit when you have a chance
		Age GLBA	谢谢亲爱的肯定	Thank you, my dear

Figure 5: Visualization of trait attention scores for our model (Att. + PAB). The generated response is "来吧来吧，我还在等你" (*Come on come on, I am still waiting for you (in Yunnan)*). "云南"("Yunnan") is a province in China.

"come on come on"), which are commonly used by *females*, the trait attention scores regarding to Gender are higher. When generating contents related to locations (i.e. "still waiting for you (in Yunnan)"), the trait attention scores regarding to Location are higher.

6 PRIVACY PROTECTION

Privacy is an important issue for most datasets that are collected from social media, particularly when personal information is involved. In order to protect the privacy of each speaker in our dataset, we designed several anonymization schemes following a critical principle that *the speaker's identity should not be traceable*:

(1) The IDs for speakers and Weibo posts were masked;

(2) The dialogues that involve explicit references to other users (in particular, using "@") were abandoned;

(3) A de-lexicalization operation was performed by replacing all the numbers with a ⟨NUM⟩ placeholder. It helped to hide more details of each speaker, such as phone numbers or addresses.

The above proposed schemes can effectively anonymize the private information related to each speaker. Moreover, in order to further protect speakers' privacies, we limit the use of PersonalDialog to be strictly constrained to academic researchers without any attempts to de-anonymize the released data. Anyone who want to use the data has to sign a contract to obey these rules strictly.

7 CONCLUSION AND FUTURE WORK

In this paper, we investigate a novel task to generate personalized dialogue responses by considering explicitly represented personality traits. To facilitate such research, we first construct a large-scale (20.83M sessions) multi-turn dialogue dataset, PersonalDialog, from real social conversations. The dataset contains various traits (e.g. Age, Gender, Location and Interest Tags) of a large number of speakers (8.47M speakers). Then, we present personalized dialogue generation models to capture and address personality traits in the dialogue generation process. These models apply a trait fusion module to obtain the persona representation of a speaker, and two approaches to address persona-related features in the decoding process: namely persona-aware attention mechanism which dynamically generates context vectors conditioned on the persona representation, and persona-aware bias which manipulates the final generation distribution directly. Automatic and manual evaluation shows that our models can incorporate richer traits in dialogue generation and can learn to choose proper traits in different contexts.

We demonstrate simple models for personalized dialogue generation, and they can serve as baselines for further studies in this research direction since the topic is still in its infancy. The corpus PersonalDialog will facilitate not only the study of personalized dialogue systems, but also other research areas such as sociolinguistics or social science.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[2] David Bamman, Jacob Eisenstein, and Tyler Schnoebelen. 2014. Gender identity and lexical variation in social media. *Journal of Sociolinguistics* 18, 2 (2014), 135–160.

[3] Rafael E. Banchs. 2012. Movie-DiC: A Movie Dialogue Corpus for Research and Development. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2 (ACL '12)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 203–207. http://dl.acm.org/citation.cfm?id=2390665.2390716

[4] Mateusz Buda, Atsuto Maki, and Maciej A Mazurowski. 2018. A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks* 106 (2018), 249–259.

[5] Zhiyun Cao and Xiaohai Liu. 2008. Linguistic atlas of Chinese dialects.

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase

representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[7] Deborah A Cobb-Clark and Stefanie Schurer. 2012. The stability of big-five personality traits. *Economics Letters* 115, 1 (2012), 11–15.

[8] Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs.. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, ACL 2011.*

[9] Gabriel Doyle. 2014. Mapping dialectal variation by querying social media. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics.* 98–106.

[10] Penelope Eckert. 2017. Age as a sociolinguistic variable. *The handbook of sociolinguistics* (2017), 151–167.

[11] Lucie Flekova, Jordan Carpenter, Salvatore Giorgi, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. Analyzing biases in human perception of user age and gender from text. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1. 843–854.

[12] Alastair J Gill, Carsten Brockmann, and Jon Oberlander. 2012. Perceptions of alignment and personality in generated dialogue. In *Proceedings of the Seventh International Natural Language Generation Conference.* Association for Computational Linguistics, 40–48.

[13] Erving Goffman. 1959. The presentation of self in everyday life. *New York* (1959).

[14] Lewis R Goldberg. 1993. The structure of phenotypic personality traits. *American psychologist* 48, 1 (1993), 26.

[15] Hugo Hernault, Paul Piwek, Helmut Prendinger, and Mitsuru Ishizuka. 2008. Generating dialogues for virtual agents using nested textual coherence relations. In *International Workshop on Intelligent Virtual Agents.* Springer, 139–145.

[16] Susan C Herring. 2007. A faceted classification scheme for computer-mediated discourse. *Language@ internet* 4, 1 (2007).

[17] Aaron Jaech and Mari Ostendorf. 2017. Low-Rank RNN Adaptation for Context-Aware Language Modeling. *arXiv preprint arXiv:1710.02603* (2017).

[18] Chaitanya K. Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in Goal-Oriented Dialog. (2017).

[19] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *Eprint Arxiv* (2014).

[20] Satwik Kottur, Xiaoyu Wang, and Vitor Carvalho. 2017. Exploring Personalized Neural Conversational Models. In *Twenty-Sixth International Joint Conference on Artificial Intelligence.* 3728–3734.

[21] Siwei Lai, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Recurrent Convolutional Neural Networks for Text Classification.. In *AAAI*, Vol. 333. 2267–2273.

[22] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055* (2015).

[23] Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A Persona-Based Neural Conversation Model. (2016), 994–1003.

[24] François Mairesse and Marilyn Walker. 2007. PERSONAGE: Personality generation for dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics.* 496–503.

[25] François Mairesse and Marilyn A Walker. 2008. A Personality-based Framework for Utterance Generation in Dialogue Applications.. In *AAAI Spring Symposium: Emotion, Personality, and Social Behavior.* 80–87.

[26] Pierre-Emmanuel Mazaré, Samuel Humeau, Martin Raison, and Antoine Bordes. 2018. Training Millions of Personalized Dialogue Agents. *arXiv preprint arXiv:1809.01984* (2018).

[27] Dong Nguyen, A Seza Doğruöz, Carolyn P Rosé, and Franciska de Jong. 2016. Computational sociolinguistics: A survey. *Computational linguistics* 42, 3 (2016), 537–593.

[28] Dong Nguyen, A. Seza Doğruöz, Carolyn P. Rosé, and Franciska de Jong. 2016. Computational Sociolinguistics: A Survey. *Comput. Linguist.* 42, 3 (Sept. 2016), 537–593. https://doi.org/10.1162/COLI_a_00258

[29] Dong Nguyen, Dolf Trieschnigg, A Seza Doğruöz, Rilana Gravel, Mariët Theune, Theo Meder, and Franciska De Jong. 2014. Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers.* 1950–1961.

[30] Warren T Norman. 1963. Toward an adequate taxonomy of personality attributes: Replicated factor structure in peer nomination personality ratings. *The Journal of Abnormal and Social Psychology* 66, 6 (1963), 574.

[31] Hiroki Ouchi and Yuta Tsuboi. 2016. Addressee and response selection for multi-party conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* 2133–2143.

[32] Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. (2017).

[33] Justus J Randolph. 2005. Free-Marginal Multirater Kappa (multirater K [free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. *Online submission* (2005).

[34] Alan Ritter, Colin Cherry, and William B Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the conference on empirical methods in natural language processing.* Association for Computational Linguistics.

[35] Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2015. A Survey of Available Corpora for Building Data-Driven Dialogue Systems. *Computer Science* 33, 16 (2015), 6078–6093.

[36] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*, Vol. 16. 3776–3784.

[37] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364* (2015).

[38] David Shulman. 2016. *The Presentation of Self in Contemporary Social Life.* SAGE Publications.

[39] Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From Eliza to XiaoIce: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering* 19, 1 (2018), 10–26.

[40] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.* ACM, 553–562.

[41] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems.* 3104–3112.

[42] Neal W Topp and Bob Pawloski. 2002. Online data collection. *Journal of Science Education and Technology* 11, 2 (2002), 173–178.

[43] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* (2015).

[44] Marilyn A Walker, Grace I Lin, and Jennifer E Sawyer. 2012. An Annotated Corpus of Film Dialogue for Learning and Characterizing Character Style. 1373–1378.

[45] Jianan Wang, Xin Wang, Fang Li, Zhen Xu, Zhuoran Wang, and Baoxun Wang. 2017. Group Linguistic Bias Aware Neural Response Generation. In *The Workshop on Ijcnlp.*

[46] Rui Zhang, Honglak Lee, Lazaros Polymenakos, and Dragomir Radev. 2017. Addressee and response selection in multi-party conversations with speaker interaction rnns. *arXiv preprint arXiv:1709.04005* (2017).

[47] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing Dialogue Agents: I have a dog, do you have pets too?. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers).* Association for Computational Linguistics, 2204–2213. http://aclweb.org/anthology/P18-1205

[48] Wei-Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2017. Neural personalized response generation as domain adaptation. *World Wide Web* (2017), 1–20.

[49] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074* (2017).