Listener's Social Identity Matters in Personalised Response Generation

Guanyi Chen[♠], **Yinhe Zheng**[♣♡], and **Yupei Du**[♠]
[♠]Department of Information and Computing Sciences, Utrecht University
[♣]Samsung Research China - Beijing (SRC-B)
[♡]Department of Computer Science and Technology, Tsinghua University
g.chen@uu.nl, yh.zheng@samsung.com, y.du@uu.nl

Abstract

Personalised response generation enables generating human-like responses by means of assigning the generator a social identity. However, pragmatics theory suggests that human beings adjust the way of speaking based on not only who they are but also whom they are talking to. In other words, when modelling personalised dialogues, it might be favourable if we also take the listener's social identity into consideration. To validate this idea, we use gender as a typical example of a social variable to investigate how the listener's identity influences the language used in Chinese dialogues on social media. Also, we build personalised generators. The experiment results demonstrate that the listener's identity indeed matters in the language use of responses and that the response generator can capture such differences in language use. More interestingly, by additionally modelling the listener's identity, the personalised response generator performs better in its own identity.

1 Introduction

Persona plays an important role in our daily communication since it affects the way we render our dialogues. Social variables, such as gender, age, place of birth or even wealth and social status, account for a large proportion in each individual's persona. Numerous previous studies have suggested that these variables strongly affect each speaker's word preference in dialogues. A growing body of works has been carried out to implicitly or explicitly model these variables in dialogues (Li et al., 2016b; Qian et al., 2017; Kottur et al., 2017; Zhang et al., 2018; Zheng et al., 2019, 2020b).

Despite the reported success, most previous studies for personalised dialogue modelling consider only the persona of speakers. [1] Nevertheless, the pragmatics theory suggests that the speaking style will be adjusted not only by who the speaker is, but also whom the speaker is talking to (Wish et al., 1976; Hovy, 1987). In the computational linguistics community, Dinan et al. (2020) investigates this issue by measuring and mitigating gender bias in dialogue dataset utilising a gender classifier. From the aspect of personalised dialogue generation, Zhang et al. (2018) and Zheng et al. (2020b) tried to attach the listener persona to the encoder of their generator, but interestingly, they obtained very different results, namely, the performance of Zhang et al. (2018) went down while that of Zheng et al. (2020b) went up.

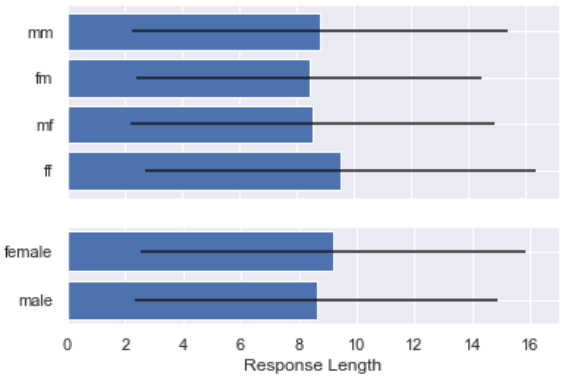

Figure 1: Average response length of each style.

Nonetheless, no systematic studies have been conducted to investigate what role does the listener's identity play in personalised response generation. Research questions that we wish to answer by the proposal put forward in this paper are:

1. How the listener's social identity impacts the responder's language use;

2. Can a response generator capture this impact,

[1] For using the terminology consistently, we use "speaker" referring to the person who produces the response (who is also a personalised dialogue system heading to model) and "listener" referring to the one who utters the post.

Proceedings of The 13th International Conference on Natural Language Generation, pages 205–215,
Dublin, Ireland, 15-18 December, 2020.

if yes, in which way?

To this end, we apply analysis and build a response generator on a Chinese personalised dialog dataset: PERSONALDIALOG, a corpus extracted from Weibo[2]. There are two reasons to use this dataset: one is that the PERSONALDIALOG dataset origins from the real conversations on social media Weibo, in which speakers' social variables play an important role; the other is that this dataset provides a massive amount of dialogue data (over 20M sessions) between a large population of speakers (over 8M speakers). It is of sufficient size to capture a variety of linguistic phenomena that are associated with social variables. Each speaker/listener in PERSONALDIALOG comes up with 4 social variables: gender, age, location, and interests. For simplicity and for conducting controlled analysis and experiments, we only focus on gender in this paper.

As for the first research question, we postulate that a speaker behave differently when s/he speaks to people with different gender stylistically. This yields four possible speaking styles: `ff`, `mf`, `fm`, and `mm`[3]. We, therefore, build a classifier to separate these styles defining on "gender-pairs". Previous analysis on blogging data (Schler et al., 2006; Goswami et al., 2009; Nguyen et al., 2011; Bamman et al., 2014) has identified that one of the key features for distinguishing contents produced by a female from those by a male is the sentence length, i.e., females tend to utter longer sentences. As shown in Figure 1, the same phenomenon is found in PERSONALDIALOG: females' responses are generally longer than males'. Further statistics on the response length falling the above four styles suggest that gender-pairs are also separable, perhaps excepting `mf` and `fm` at first glance. To validate this and understand why, we build a gender-pair classifier and conduct so-called pivot word analysis. We find out which word contributes the most for helping the classifier make decisions. Experiment results show that these styles are separable, but `mf` and `fm` are often confused with each other.

As for the second research question, we build a personalised response generator conditioning on these styles. The outcomes suggest that the generator could capture the difference between those styles and, in addition, modelling the listener's identity helps the generator to express its own identity. Moreover, based on previous analyses, we have also tried to merge the style of `mf` and `fm` into a single integrated style `mf/fm`. However, the final results of the response generator suggests that it is hard to model utterances with this integrated style.

[2]Weibo is the largest Chinese social media.

[3]We use `fm` to represent the style used by a male speaker when talking to a female listener. Similar definition applies to `mf`, `mm`, and `ff`.

Model	2-way	3-way	4-way
fastText	0.85	0.75	0.68
TextCNN	0.85	0.73	0.63
LSTM	0.85	0.75	0.63
BOW Classifier	0.85	0.74	0.64

Table 1: F1 score of the gender-pair classifiers.

2 Gender-Pair Classification

To approach the first research question, we build a gender-pair classifier to simultaneously recognise the speaker's and listener's social identity based on the dialogue utterances. Concretely, as aforementioned in section 1, we assume the present task as a style classification task and design four labels for each input dialogue utterance: `mm` (male talking to male), `mf` (female talking to male), `fm` (male talking to female), and `ff` (female talking to female).

However, in light of the Linguistic Style Matching theory (Niederhoffer and Pennebaker, 2002), speakers will imitate the linguistic style of their conversation companion to pursue higher engagement. In other words, when two different gendered speakers communicate with each other, their speaking style may assimilate to each other as the conversation proceed. On the top of this observation, one may say that dissociating `fm` and `mf` is hard, and, therefore, it would be favourable if we merge `fm` and `mf` into a single category, namely `mf/fm`.

2.1 Build Gender-Pair Classifiers

Building on what has been discussed, to further get insight from conventional gender classification, we consider the following three classification tasks basing on three speaking style categorisation schemes: 1) two-way classification: classifying only speakers' gender, in which two labels are used: `male` and `female`; 2) three-way classification: classifying the conversational texts based on a merged labelling scheme, i.e., three labels are considered `mm`, `fm/mf`, and `ff`; and 3) four-way classification: the gender-pair classification which classifies

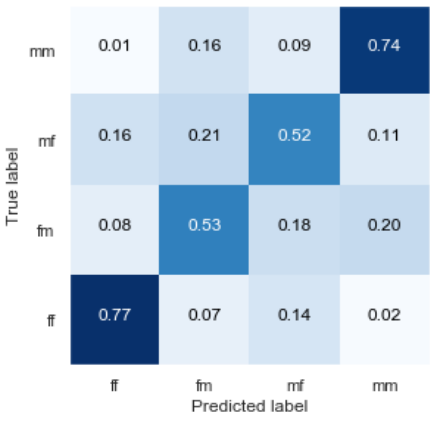

Figure 2: Confusion matrix of the 4-way gender-pair classification using fastText.

the conversational texts into `mm`, `fm`, `mf`, and `ff`.

2.1.1 Classification Models

We test a number of text classification algorithms, including fastText[4] (Joulin et al., 2017), TextCNN (Kim, 2014) and LSTM (Hochreiter and Schmidhuber, 1997) (in which the hidden states of all the tokens are max pooled before being feed into the final Softmax layer). In order to conduct interpretable analysis, we train a Bag-of-Word (BOW) classifier: a logistic regression with only unigram features.

2.1.2 Experimental Settings

Building on the fact that classifying the social variables based on the social media data is hard (Nguyen et al., 2013, 2014), and the exhibition of speakers' social identities is sparse in social media text (Zheng et al., 2020b), we adopt the classification strategy used by Zheng et al. (2019). Specifically, each classifier input is a concatenations of N randomly sampled responses with the same style. In this study, we use $N = 20$. We train and test the classifiers on PERSONALDIALOG, where the dataset has been divided into training and testing sets without overlapping. The training data are down-sampled to balance the corpus. 10% of the training set is held out for tuning parameters, and the final models are trained on the whole training set. The classifiers are evaluated using F1 scores.

2.1.3 Experimental Results

Table 1 depicts the performances of these classifiers. FastText performs remarkably well. It outperforms both TextCNN and LSTM, which are models having much higher complexity and capacity. It is surprising that the simplest BOW classifier also achieves comparably good performance, which suggests that the word usage is the most important feature for distinguishing speakers' social identity (at least for the gender). Further comparison of the fastText and BOW classifier embodies that the unigram features are sufficient for conducting gender classification in the coarse 2-way classification setting, while higher-ordered N-gram features (used by the fastText) are useful in more fine-grained 3-way and 4-way classification settings.

The F1 score of the 4-way gender-pair classification using fastText reaches 0.68. This means that it is feasible to identify the style of the listener by only considering the utterances issued by the speaker. We print the confusion matrix of this result in Figure 2. The utterances from `ff` and `mm` are rarely confused with each other. This indicates that the language use of both males and females have clear differences when they speak to people with the same gender When they talk to people with different gender, in line with the results of gender classification, they tend to express stylistic characteristics related to their own gender since confusions appear between `fm` and `mm` as well as between `mf` and `ff`. Nonetheless, we also observe equally severe confusion between `fm` and `mf`, which approves that the linguistic style matching hypothesis plays a certain role when people expressing their social identities.

In addition, we also observe a certain level of confusion between `fm` and `ff` as well as between `mf` and `mm`. This said, the classifier sometimes confuse between, for example, an utterance from a male and an utterance from a female when they both speak to male listeners. This, yet again, could be seen as an evidence for the existence of linguistic style matching. Although the utterance from `fm` and `mf` shows a tendency of assimilation, it appears that the speakers still maintain the characteristics of their own gender and, in this sense, there are still certain reasons to disassociate the style of `fm` from `mf`.

2.2 Pivot Word Discovery

To understand how people change their language use with respect to social identities of themselves and of whom they speak to, or, in other words, to understand how the gender-pair classifiers make

[4]The official implementation of fastText from Facebook is used: https://github.com/facebookresearch/fastText.

Algorithm 1 Classifier-based Pivot Word Discovery

Input: Dataset $\mathcal{D}$, Style Set $\mathcal{S}$, BOW Classifier f, Confidence Threshold α, and Word Pivot Frequency Threshold β.
Output: A set of Pivot Words Ω

1: **for** each input sentence x and corresponding label $y = s \in \mathcal{S}$ in $\mathcal{D}$ **do**
2: Predict label $\hat{y}$ and confidence p for x
3: **if** $\hat{y} = y$ **then**
4: **for** each word type t in x **do**
5: Construct $x_{\setminus t}$ by removing all t in x
6: Predict label $\hat{y}'$ and confidence p' for $x_{\setminus t}$
7: **if** $\hat{y}' \neq y$ or $p - p' > \beta$ **then**
8: Add t to Ω_c and add pivot word frequency $p(t, s)$ by 1
9: **end if**
10: **end for**
11: **end if**
12: **end for**
13: **return** All t in Ω_c if $p(t, s) > \beta$ for all $s \in \mathcal{S}$

their decisions, we apply the *Pivot Word Analysis*. Pivot words are words that have substantial influence on the classifier's decision making and have been widely used for interpreting the language use in many language generation tasks such as Style Transfer (Fu et al., 2019) and Table-to-Text Generation (Ma et al., 2019).

2.2.1 Pivot Word Extraction Algorithm.

Since the expression of social identity is sparse in the social media data, the appearance of pivot words in the utterance is also sparse. Therefore, the pivot word discovery algorithms introduced in (Fu et al., 2019) and (Ma et al., 2019) are not applicable in the present task. Instead, we use a simple yet efficient pivot word discovery algorithm coined as *Classifier-based Pivot Word Discovery* for extracting pivot words using the trained BOW classifier.

The algorithm is of finding out which word type in the training data plays a major role in the BOW classifier's decision-making. It is sketched in Algorithm 1. As can be seen from lines 2-5, this algorithm only considers samples that have been correctly classified. For each word type t in a sample x, it compares the classification results and confidences when including and excluding t in x (lines 2-8). Specifically, if the classifier's predicted result is changed or the prediction confidence's change exceeds a certain threshold of β, we extract it as a pivot word candidate (line 10). If the same word type has been extracted as a candidate for more than α times under a single category, the algorithm returns it as a pivot word (line 15). In this work, we set α and β to 10 and 0.5, respectively.

2.2.2 Extracted Pivot Words.

Table 2 lists typical examples of the extracted pivot words in each category for the gender classifier and the gender-pair classifier. As for the gender classification, we observe that the general topics used by males and females have clear differences on Weibo. Specifically, males focus on the topic of digital products, politics, and games while females like talking about starstruck, teleplays, makeup, and shopping. It is worth noting that one reason that Weibo users concentrate on these topics is that most of them are young people according to the statistics in Zheng et al. (2019). These topics might change if use data extracted in more recent years since the PERSONALDIALOG dataset was crawled in 2018.

More interestingly, we also find that differences exist in the use of punctuation and pronouns. Males use punctuation in a more formal way on social media (in which comma and period are frequently used), but females eager to concatenate a sequence of punctuation to express certain emotions or speech acts (e.g., "∼∼", "!!!!"). The first person pronoun was extracted as pivot word for the `female` category, which might suggest that males are more likely to drop pronoun on social media. [5] To say the last word on how the use of zero pronouns is affected by the speaker's social identity needs further research, which is not the focus of this paper.

As for comparing the extracted pivot words for the gender-pair classifier and the gender classifier, in line with the classification results detailed in section 2.1, we observe more overlaps between `female` and `ff` as well as `male` and `mm` than between `female` and `mf` as well as `male` and `fm`. When comparing the words from different gender-pair categories, we find that people would talk about different topics when they talk to people of the same gender and with a different gender. For

[5] Chinese as a discourse based language, pro-drop (Huang, 1984) is much more common than that in, for example, English, especially when the dropped pronoun referring to one of the speakers in a conversation (Chen et al., 2018).

Model	Example Pivot Words
mm	华为 (Huawei), 苹果 (Apple), 三星 (Samsung), 小米 (Xiaomi), 美国 (America), 日本 (Japan), 中国 (China), 大陆 (Mainland), 台湾 (Taiwan),“，”, ”。”
fm	游戏 (game), 王者 (Honer of Kings), 早安 (good morning), 晚安 (good night), 拍照 (photograph), 读书 (reading), 工作 (working), 我 (I), 你 (you)
mf	大叔 (Uncle), 弟弟 (little Brother), 哥哥 (elder Brother), 上班 (Working), 喝酒 (Drinking), 厦门 (Xiamen), 广东 (Guangdong), 广州 (Guangzhou), 嗯嗯 (Uh-huh), 我 (I), 你 (you), “～～”, “!!!!”, “???”
ff	王俊凯 (a celebrity), 易烊千玺 (a celebrity), 鹿晗 (a celebrity), KPop, 男主 (leading actor), 电视剧 (teleplay), 化妆 (make up), 漂亮 (beauty), 裙子 (skirt), 便宜 (cheap), 淘宝 (Taobao)， 嗯嗯 (Uh-huh), 啊啊啊 (Ah Ah Ah), 我 (I), 你 (you), “～～～～”, ”!!??”, ”!!!!”
male	华为 (Huawei), 苹果 (Apple), 美国 (America), 大陆 (Mainland), 台湾 (Taiwan), 妹子 (girl), 媳妇 (wife), 游戏 (game), “，”, ”。”
female	王俊凯 (a celebrity), 易烊千玺 (a celebrity), 男主 (leading actor), 电视剧 (teleplay), 化妆 (make up), 裙子 (skirt), 面膜 (mask), 刘海 (bang), 我 (I), “～～”, “!!!!”, “～～～～”, hhh, QAQ, mua

Table 2: Lists of extracted pivot words in each categories of the gender classifier and the gender-pair classifier.

	mm	mf	fm	ff	male	female
mm	0.07 (-0.70)	0.96 (+0.19)	0.99 (+0.22)	0.98 (+0.21)	0.12 (-0.65)	0.99 (+0.22)
mf	0.72 (+0.19)	0.00 (-0.53)	0.23 (-0.30)	0.01 (-0.52)	0.41 (-0.12)	0.00 (-0.53)
fm	0.27 (-0.25)	0.31 (-0.21)	0.02 (-0.50)	0.19 (-0.33)	0.04 (-0.48)	0.11 (-0.41)
ff	0.79 (+0.05)	0.10 (-0.64)	0.21 (-0.53)	0.00 (-0.74)	0.94 (+0.20)	0.00 (-0.74)

Table 3: Recall of two pivot free classification experiments. Labels in the first row indicates the source categories the labels in the first column are the target categories. In each cell, a ($\pm b$) means the recall is a and comparing to its original performance the score increases/decreases b.

example, when a female talks to another female, they discuss “idols” they like, shopping, and dressing, which are rarely mentioned when she talks to a male. These observations explain why utterances with style mf (fm) are separable from those with style ff (mm) and suggest that the identities of listeners really matter the way of how speakers speaking.

As for the linguistic matching hypothesis, some evidences have been found. For example, fm and mf shared some topics including travelling, studying, working or gaming. Moreover, first person pronouns are more likely to be used when males speaking to females, but similar matching not appears in the use of punctuation.

2.3 Pivot Free Classification

In order to quantify how the gender-pair influences the language use, we do a *Pivot Free Classification* experiment, where the BOW classifier is evaluated on the test data, in which the pivot words from a certain category are removed. Since we care about, by removing the pivot words, how many samples of a category are mis-classified into other categories, we report the recall scores in Table 3. We test the performance of the gender-pair classifier “attacked” by pivot words extracted by the gender-pair and the gender classifier. We name the category on which we report the performance as the *target category* and the category from which we extract the pivot words as the *source category*.

On the basis of the results in Table 3, we have the following observations: **First**, the performance reduces to almost zero if the source and the target are the same categories, which implies that the extracted pivot words are those which actually bias the decision making of the classifier. **Second**, ff and mm are definitely separable as no impact is found when they “attack” each other. **Third**, in line with the previous findings and the linguistic style matching theory, mf and fm are highly confused with each other, which can be approved from two dimensions: 1) as source categories, they highly reduce each other’s performance; 2) Pivot words from female have remarkably effects on not only mf and ff but also fm. **Fourth**, mf and fm are not exactly the same, since, for instance, the impact of mf on ff is clearly higher than that of fm on ff. **Last**, the style of a conversation for speakers

with a different gender is more similar to the style of how females speak.

3 Personalised Response Generation

For exploring the second research question, that is, can a personalised response generator capture the differences of language use when imitating a speaker talking to listeners with different social identities? We train multiple response generators conditioning on the three style categorising schemes mentioned in section 2. We start by introducing the basic architecture of our generator and the experimental settings. We then describe the evaluation metrics we use, with which we evaluate and analyse the generators.

3.1 The Personalised Response Generator

Since inventing a new state-of-the-art personalised response generator falls out of the scope of this paper, we build the model following a simplified paradigm of Zheng et al. (2020a,b). The architecture of the model we used is sketched in Figure 3.

Concretely, given the dataset, containing N dialogue pairs with each of their style: $\mathcal{D} = \{(x_1, y_1, s_1), ..., (x_N, y_N, s_N)\}$, where x_i is the post, y_i is the response, and s_i is the style label of that response (i.e., in our case, it could be `female` or `fm`). As depicted in Figure 3, each post x is firstly mapped into word embedding space using $\mathbf{e}_w(\cdot)$ and then is encoded via a Transformer (Vaswani et al., 2017) based encoder to a representation $\mathbf{E}_x$.

3.1.1 Encoding Style Information

Following Zheng et al. (2020b), in the decoding phase, we inject the style information by utilising the attention routing mechanism. Specifically, different from the standard Transformer decoder, both multi-head attention (MHA) and masked multi-head attention (MMHA) are deployed. In each decoder block, given the $\mathbf{E}_x$ and the embedded previously decoded response $\mathbf{E}_{y_{pre}} = \mathbf{e}_w(y_{pre})$, they are encoded to:

$$\mathbf{R}_{pre} = \text{MMHA}(\mathbf{E}_{y_{pre}}, \mathbf{E}_{y_{pre}}, \mathbf{E}_{y_{pre}}) \quad (1)$$

$$\mathbf{R}_{post} = \text{MHA}(\mathbf{E}_{y_{pre}}, \mathbf{E}_x, \mathbf{E}_x) \quad (2)$$

Together with the mapped style, $\mathbf{E}_s$ is mapped using the style embedding $\mathbf{e}_s(s)$. These set of representations are merged in the following way to $\mathbf{R}$ before being feed for layer normalisation:

$$\mathbf{R} = (\mathbf{R}_{pre} + \mathbf{R}_{post})/2 + \mathbf{E}_{y_{pre}} + \mathbf{E}_s \quad (3)$$

in which, $\mathbf{R}_{pre}$ and $\mathbf{R}_{post}$ are averaged.

Despite of the simplicity, one major reason of why we do not use the original model of Zheng et al. (2020b) in this study is that they did not encode personae (i.e., gender in our case) of speaker and listener symmetrically. To be more specific, they encode the persona of the listener as a number of style embeddings, which were added to the input together with the positional embeddings, while the speaker's persona was encoded as a sequence of words and was concatenated with the embeded post x. This kind of disassociation makes our experiments less controlled. Instead, in this study, we merge the label for speakers and listeners (i.e., the label such as `mf`) and map it into a single style embedding.

3.1.2 Parameter Sharing and Pre-training.

Encoders and decoders in our model are sharing their parameters. To further increase the quality of the generated responses, akin to many previous research in dialogue modelling (Wolf et al., 2019; Wang et al., 2020a), we initialise the parameters in our model using a pre-trained Chinese GPT model (Radford et al., 2019; Wang et al., 2020b).

3.2 Experiments

3.2.1 Experimental Settings

We train and evaluate the model on the PERSONALDIALOG dataset. For simplicity, in line with Zheng et al. (2019), we only train and test our model using the first turn of each dialogue session in PERSONALDIALOG. For conducting a controlled and fair analysis, we train three models corresponding to the three style categorisation schemes introduced in section 2 (see Table 4). In the following sections, we refer them with their ID, i.e., model 1, 2, or 3.

3.2.2 Evaluation Metrics

Recall that our target is not of defeating state-of-the-art personalised response generator in the sense of generating better responses. Nonetheless, we still report some relevant results using commonly used automatic metrics including: BLUE (Papineni et al., 2002), a metric comparing overlaps of n-grams ($n = 1, 2$) between the reference responses and the generated responses for evaluating the adequacy and fluency; and DIST (Li et al., 2016a), measuring the proportion of distinct n-grams ($n = 1$) for evaluating the diversity of the model outputs.

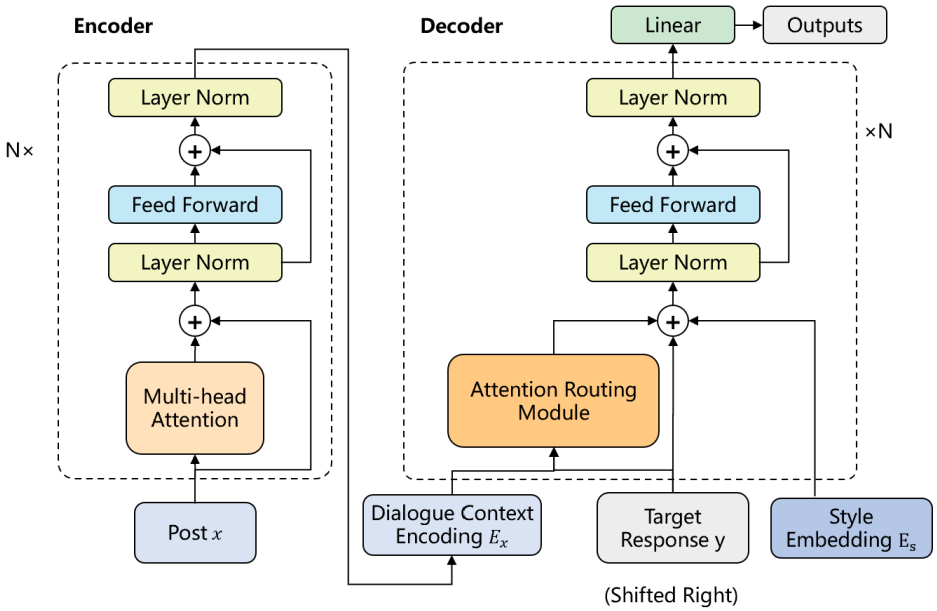

Figure 3: Illustration of our personalised response generator.

To help obtaining insights from the system outputs for the second research question, we design a number of new metrics based on the built classifiers and extracted pivot words from section 2. Specifically, for evaluating a model with n style categories, we propose the following metrics:

1. **ACC.** evaluates whether the generated responses incorporate the target style using the trained n-way classifier. Similar approach is employed to evaluate the outputs of conditional language generators with off-line classifiers (Zhou et al., 2018; Zheng et al., 2019; Li et al., 2020). During evaluation, the system outputs are concatenated in the same way as the train data of those classifiers. Considering the speed and the performance, we use the fastText classifier in the evaluation;

2. **ACC-2.** evaluates whether the generated response reflect gender information using the trained gender classifier. It is worth noting that this metric is not applicable to model 2 since we have merged the `mf` and `fm`, we expect that they are no longer separable;

3. **Pivot Word Precision (PWP).** evaluates to what proportion the generated tokens are pivot words. Suppose the system outputs with style s is $\hat{\mathcal{Y}}_s$ with the vocabulary $\mathcal{V}$ and the pivot words extracted by n-way classifier is Ω_s, the PWP is computed by:

$$\text{PWP}_s = \frac{\sum_{w \in \Omega_s} \#(w, \hat{\mathcal{Y}}_s)}{\sum_{w \in \mathcal{V}} \#(w, \hat{\mathcal{Y}}_s)} \quad (4)$$

where $\#(w, \hat{\mathcal{Y}}_s)$ is the frequency of w in $\hat{\mathcal{Y}}_s$. PWP is calculated for each style and is then micro-averaged;

4. **Pivot Word Recall (PWR).** evaluates how many word types in pivot words has been generated:

$$\text{PWR}_s = \frac{\sum_{w \in \Omega_s} \mathbb{I}(w, \hat{\mathcal{Y}}_s)}{|\Omega_s|} \quad (5)$$

where $\mathbb{I}(w, \hat{\mathcal{Y}}_s)$ equals to one if w appears in $\hat{\mathcal{Y}}_s$, otherwise it equals to 0.

3.2.3 Experimental Results

Table 4 charts the results of all the metrics above. It is not surprising that no significant difference is found in BLEU and DIST score between all three models since all of them have the same model architecture, the same parameter setting and, thus, the same capacity.

Due to the fact that different off-line classifiers have very different performance in their own domain (see Table 1), it is not fair to compare the value of ACC and ACC-2 across different dialogue generation models. However, taking other metrics

ID	Conditioned Styles	BLEU	DIST	ACC	ACC-2	PWP	PWR
1	female, male	**3.94**	**0.092**	**76.47**	76.47	41.44	66.04
2	ff, fm/mf, mm	3.58	0.089	48.00	-	46.93	**69.15**
3	ff, fm, mf, mm	3.60	0.089	63.03	**84.00**	**55.37**	63.05

Table 4: Evaluation results of response generator with different speaking style categorisation scheme by means of metrics introduced in section 3.2.

	mm	mf	fm	ff	male	female
mm	57.01	51.24	51.82	46.09	46.85	39.07
mf	58.60	59.12	63.13	58.00	46.85	50.91
fm	54.14	55.66	59.22	55.87	44.96	45.90
ff	68.79	70.17	72.63	76.87	57.56	66.97

Table 5: The results of cross-category PWR scores. Same as Table 3, categories in first row means where the pivot words from and categories in first column means where the system outputs from.

into account, we still have some interesting findings. One is that all the ACC results are better than random, which somehow suggest that all of these models have captured the differences of language use under each style. The other is that although model 2 has the highest PWR and moderate level of PWP, but, meanwhile, it has the lowest ACC. In other words, it generates lots of pivot words, but the classifier does not classify them into the correct style. To understand why, we analysed the PWP for each style, and found that it works fine for ff (69.02) and mm (45.96), but collapses at the merged category, i.e., mf/fm. It obtains a PWP at only 25.82 and a PWR at 56.95 (which is not a very bad number). It appears that although the generator has produced fine amount of pivot words for expressing the style of mf/fm, but, the frequency of many of them might not be high. This also suggests that even though we found some evidences from experiments in section 2 supporting the theory of linguistic matching and the merging of mf and fm, but it seems that the generator we use cannot handle this.

More interestingly, we also find that model 3 not only has the highest performance on PWP, which means more than half of the tokens it produces are pivot words of the correct style, but also has the highest score on ACC-2 (i.e., the accuracy of gender classification), which is even better than model 1, a model that originally designed having two styles. This approves that by additionally modelling the social identities of the listeners, it helps the generator to utter more speaker identity related words because it takes the difference on speaking style when talking to listeners with different social identities into account.

3.2.4 Cross-category PWR

To understand how model 3 works, we consider similar experiment to the one in section 2.3 by measuring the cross-category PWR. From Table 5, we observe similar phenomenon as in section 2.3. For example, the pair mm and ff yields the lowest PWR when being as the pivot word source of each other. In contrast, they reach the highest score if they are their own pivot word source. fm and mf have relatively high PWR when being each other's pivot word source. When a male talks to an another male, they say very few words that females always say. Nevertheless, we also observe that sentences produced by ff always have the highest PWR regardless of where the pivot words are coming from. This should be a result of two reasons: most conversations in PERSONALDIALOG dataset are between two females and PRR is a metric that sensitive to the size of test data (i.e., it is very likely that the more sentences are produced the more pivot words are included).

4 Discussion

We investigated the language use on Chinese social media regarding to the social identities of speakers and listeners. Specifically, we aim to explore whether the listener's social identities impact the responder's language use and whether such differences are separable. The primary answers to both

of these questions are "Yes" on the basis of our experiments and, additionally, by conducting pivot word analysis, we also found that `mf` and `fm` are less separable owing to the linguistic matching phenomenon. This raises as open question of which style categorisation scheme (i.e., whether to distinguishing `mf` and `fm` or not) is better for modelling personalised dialogues.

We then trained personalised response generators which take the social identities of listeners into account. To conduct insightful analysis, we design a number of new metrics with the help of the speaking style classifiers and the extracted pivot words. The outcomes show that modelling listener's identity assists the dialogue system to express more of its own identity. However, our system failed to model the style of `mf/fm`, which suggests the necessity of disassociating the style between `mf` and `fm`.

Note that our work focus mainly on the gender, which from our perspective, underlies further studies on investigating the influence of other listener's social variables, such as age or location, or even of listener's persona as a whole. Likewise, since we study only on data from Chinese social media, it is also worth to validate whether our findings still hold in multilingual platforms like Twitter. As for the designing of dialogue systems, we highlighted the importance of modelling listener's persona for the Chatbot to express its own personality, it is also worthwhile to evaluate the built system in other angles, such as relevance and fluency, or to validate whether the resulting chat machine is empathetic (Fung et al., 2018) or not.

Our decision on using single turn dialogue also limits the generalisability of our conclusion to real conversations since the assimilation of each others style may progress in the course of a dialogue. This may result in under-estimating the effect of the linguistic matching between speakers and listeners. In future, we will extend our work into multi-turn dialogue modelling.

5 Ethical Statement

In this paper, we use the gender as an example of social identity to understand how the speaking style of a speaker is influenced. To this end, we build gender classifiers and stylised dialogue systems. In light of the discussion in Larson (2017), gender is notoriously difficult to detect (Buolamwini and Gebru, 2018), and mis-gendering individuals is harmful to users (Keyes, 2018). Therefore, we are not and will not apply or extend the built classifiers and dialogue systems into real applications. We hope our findings could help with further works on mitigating gender bias (Liu et al., 2020) or improving fairness (Liu et al., 2019) in dialogue systems.

Acknowledgements

We thank the anonymous reviewers for their helpful comments. Guanyi Chen is supported by China Scholarship Council (No.201907720022).

References

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

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91, New York, NY, USA. PMLR.

Guanyi Chen, Kees van Deemter, and Chenghua Lin. 2018. Modelling pro-drop with the rational speech acts model. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 159–164, Tilburg University, The Netherlands. Association for Computational Linguistics.

Emily Dinan, Angela Fan, Ledell Wu, Jason Weston, Douwe Kiela, and Adina Williams. 2020. Multi-dimensional gender bias classification. *arXiv preprint arXiv:2005.00614*.

Yao Fu, Hao Zhou, Jiaze Chen, and Lei Li. 2019. Rethinking text attribute transfer: A lexical analysis. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 24–33, Tokyo, Japan. Association for Computational Linguistics.

Pascale Fung, Dario Bertero, Peng Xu, Ji Ho Park, Chien-Sheng Wu, and Andrea Madotto. 2018. Empathetic dialog systems. In *The international conference on language resources and evaluation. European Language Resources Association*.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric analysis of bloggers' age and gender. In *Third international AAAI conference on weblogs and social media*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Eduard Hovy. 1987. Generating natural language under pragmatic constraints. *Journal of Pragmatics*, 11(6):689–719.

C-T James Huang. 1984. On the distribution and reference of empty pronouns. *Linguistic inquiry*, pages 531–574.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431. Association for Computational Linguistics.

Os Keyes. 2018. The misgendering machines: Trans/hci implications of automatic gender recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Satwik Kottur, Xiaoyu Wang, and Vitor Carvalho. 2017. Exploring personalized neural conversational models. In *Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 3728–3734.

Brian Larson. 2017. Gender as a variable in natural-language processing: Ethical considerations. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 1–11, Valencia, Spain. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016b. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Xiao Li, Guanyi Chen, Chenghua Lin, and Ruizhe Li. 2020. DGST: a dual-generator network for text style transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online. Association for Computational Linguistics.

Haochen Liu, Jamell Dacon, Wenqi Fan, Hui Liu, Zitao Liu, and Jiliang Tang. 2019. Does gender matter? towards fairness in dialogue systems. *arXiv preprint arXiv:1910.10486*.

Haochen Liu, Wentao Wang, Yiqi Wang, Hui Liu, Zitao Liu, and Jiliang Tang. 2020. Mitigating gender bias for neural dialogue generation with adversarial learning. *arXiv preprint arXiv:2009.13028*.

Shuming Ma, Pengcheng Yang, Tianyu Liu, Peng Li, Jie Zhou, and Xu Sun. 2019. Key fact as pivot: A two-stage model for low resource table-to-text generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2047–2057, Florence, Italy. Association for Computational Linguistics.

Dong Nguyen, Rilana Gravel, Dolf Trieschnigg, and Theo Meder. 2013. "how old do you think i am?"; a study of language and age in twitter. In *Proceedings of the seventh international AAAI conference on weblogs and social media*. AAAI Press.

Dong Nguyen, Noah A Smith, and Carolyn Rose. 2011. Author age prediction from text using linear regression. In *Proceedings of the 5th ACL-HLT workshop on language technology for cultural heritage, social sciences, and humanities*, pages 115–123.

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*, pages 1950–1961.

Kate G Niederhoffer and James W Pennebaker. 2002. Linguistic style matching in social interaction. *Journal of Language and Social Psychology*, 21(4):337–360.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Qiao Qian, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2017. Assigning personality/identity to a chatting machine for coherent conversation generation. *arXiv preprint arXiv:1706.02861*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9.

Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W Pennebaker. 2006. Effects of age and gender on blogging. In *AAAI spring symposium: Computational approaches to analyzing weblogs*, volume 6, pages 199–205.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Dingmin Wang, Chenghua Lin, Li Zhong, and Kam-Fai Wong. 2020a. Dialogue state tracking with pre-trained encoder for multi-domain trask-oriented dialogue systems. *arXiv preprint arXiv:2004.10663*.

Yida Wang, Pei Ke, Yinhe Zheng, Kaili Huang, Yong Jiang, Xiaoyan Zhu, and Minlie Huang. 2020b. A large-scale chinese short-text conversation dataset. In *NLPCC*.

Myron Wish, Morton Deutsch, and Susan J Kaplan. 1976. Perceived dimensions of interpersonal relations. *Journal of Personality and social Psychology*, 33(4):409.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

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)*, pages 2204–2213. Association for Computational Linguistics.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *CoRR*, abs/1901.09672.

Yinhe Zheng, Zikai Chen, Rongsheng Zhang, Shilei Huang, Xiaoxi Mao, and Minlie Huang. 2020a. Stylized dialogue response generation using stylized unpaired texts. *arXiv preprint arXiv:2009.12719*.

Yinhe Zheng, Rongsheng Zhang, Minlie Huang, and Xiaoxi Mao. 2020b. A pre-training based personalized dialogue generation model with persona-sparse data. In *AAAI*, pages 9693–9700.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *AAAI*.