Group Linguistic Bias Aware Neural Response Generation

Jianan Wang[1], **Xin Wang**[2], **Fang Li**[1], **Zhen Xu**[2],
Zhuoran Wang[3] and **Baoxun Wang**[3]
[1]Shanghai Jiao Tong University, Shanghai, China
[2]Harbin Institute of Technology, Harbin, China
[3]Tricorn (Beijing) Technology Co., Ltd, Beijing, China
[1]{wangjianan,fli}@sjtu.edu.cn
[2]{xwang,zxu}@insun.hit.edu.cn
[3]{wangzhuoran, wangbaoxun}@trio.ai

Abstract

For practical chatbots, one of the essential factor for improving user experience is the capability of customizing the talking style of the agents, that is, to make chatbots provide responses meeting users' preference on language styles, topics, etc. To address this issue, this paper proposes to incorporate linguistic biases, which implicitly involved in the conversation corpora generated by human groups in the Social Network Services (SNS), into the encoder-decoder based response generator. By attaching a specially designed neural component to dynamically control the impact of linguistic biases in response generation, a Group Linguistic Bias Aware Neural Response Generation (GLBA-NRG) model is eventually presented. The experimental results on the dataset from the Chinese SNS show that the proposed architecture outperforms the current response generating models by producing both meaningful and vivid responses with customized styles.

1 Introduction

Automated Chat Agents (a.k.a chatbots) have drawn great attention in Natural Language Processing research in recent years (Shang et al., 2015; Li et al., 2016; Wu et al., 2016; Xing et al., 2017), and the springing up of the practical chatbots (e.g., Duer[1], XiaoIce[2], etc.) indicates the great potential of such systems for naturally connecting human beings with various online services.

[1]http://duer.baidu.com/
[2]http://www.msxiaoice.com/

The core functionality of chatbots is to interact with users for the purpose of general conversation. This requires chatbots to generate responses not only relevant to users' queries but also in accordance with users' preferred talking styles (Allwood et al., 1992). State-of-art practical chatbots are capable of providing basic chatting functionality with necessary task-oriented abilities. However, they all lack the capability of adapting the generated responses to meet users' preferences. To improve the user experience, it is necessary to add such talking style customization function in these chat agents. In previous studies, **inter-group linguistic biases** are observed and found in daily conversations among people from different communities (Maass et al., 1989). In this paper we generalize such biases into those among groups of people based on their profiles or social attributes. People from different groups may have different talking styles, including syntactic (sentence structure), semantic (choice of words) or even attitudinal differences. Then such challenge of chatbots could be defined as: how to express such differences in the generated responses for different user preferences.

Benefiting from the nature of the Deep Neural Networks (DNN) based encoder-decoder framework (Sutskever et al., 2014), previous studies tried to jointly learn the representation of each **individual's talking habits** in the training procedure of word embedding, and take such habits as a part of the input to generate personalized responses (Li et al., 2016; Alrfou et al., 2016). It is found that, given a large amount of high-quality utterance data of each user, this methodology can obtain promising results of personalized response generation. For practical chat agents, however, it is not trivial to collect such high-quality utterance from users. As a consequence, it is observed that the expected responses to similar queries tend to

Proceedings of the 9th SIGHAN Workshop on Chinese Language Processing, pages 1–10,
Taipei, Taiwan, December 1, 2017.

be uncharacteristic, that is, the effect of individual-level personalized response generator is not significant. Therefore, it is more reasonable to generate personalized responses by modeling the feature of a group of users, rather than an individual (Hu et al., 2014).

Query	What are you doing?
Group A	I am watching a **basketball game**.
Group B	I am **shopping** on Amazon!
Query	I uninstalled your game just now.
Group A	How could you do that!
Group B	You are so dead.

Table 1: Responses from two user groups (A & B) categorized by user gender. The examples are selected from the real Chinese Social-Network-Service (SNS) dataset and translated into English.

In real world data, it is observed that the distinct features of generated responses are generally reflected by keywords and sentence structures, as shown in Table 1, which matches previous findings. Therefore, in order to leverage the group linguistic biases in an encoder-decoder based model, we need to apply such biases in the process of word generation. The model should be capable of controlling the distribution of such biases, rather than assigning equal intensities on each word in the response. This could make the generated responses distinguishable in groups of people in the distinctiveness of keywords and sentence structures while still guaranteeing the validity of the sentence both on semantic and syntactic level. This could thus prevent the generated responses similar on structure and use of words.

In this paper, we propose an encoder-decoder based architecture, Group Linguistic Bias Aware Neural Response Generation (GLBA-NRG) model, which incorporates **linguistic biases of human groups** into an encoder-decoder based response generator, in order to tackle the talking style customization problem of practical chatbots. We attempt to learn the representations of the linguistic biases from a gender-split corpora. Such representations are then used to bias the word selection in the response generation process. More importantly, we present a specially designed neural network component, as a soft-switch to conduct the dynamic controlling of the impact of linguistic biases on each generation step. With the adoption of the linguistic bias impact controlling mechanism, our model is able to generate responses highly corresponding to the specified talking style, while the semantic relevance between queries and responses is well maintained.

The rest of this paper is organized as follows: Section 2 surveys the related work. Our proposed model is detailed in Section 3. Section 4 describes the experimental setups and analyzes the results. Finally, our work is concluded in Section 5.

2 Related Work

Along with the development of Neural Machine Translation(NMT), many recent studies show that the basic neural-based encoder-decoder framework (Sutskever et al., 2014; Bahdanau et al., 2014) can also be successfully applied in conversation modeling (Vinyals and Le, 2015; Yao et al., 2015; Zhou et al., 2016; Iulian et al., 2017), which generates a response on the basis of a given query. Based on the work of Vinyals and Le (2015) that directly applies sequence-to-sequence (Seq2Seq) architecture for response generation, Shang et al. (2015) introduce the global and local scheme with attention signal into the generation of response, while Sordoni et al. (2015) take contextual information into account to generate context-sensitive responses.

Besides the query and contextual information, several explicit and implicit factors (e.g. topic, emotion) play great roles in response generation. To utilize such factors for generating informative, diverse and interesting responses, several works incorporate topics, external knowledge, emotional content, and responding mechanism into conversation models. Xing et al. (2017) and Xu et al. (2016) extract related topics or knowledge from the query and context respectively, then add these info into conversational models, so as guiding them to generate informative and interesting responses. Ghosh et al. (2017) and Zhou et al. (2017b) explore the influence of the affective information in response generation with Affect-LM and emotional memory separately. Taking explicit and implicit factors as the high-level semantic content of the response, Serban et al. (2017) and Zhou et al. (2017a) propose latent variable and responding mechanism respectively to enrich the capability of conversation models to generate diverse responses.

Moreover, the personality is of great importance for chatbots to respond coherently, as argued by Vinyals and Le (2015). The very first attempt to model persona is from Li et al. (2016),

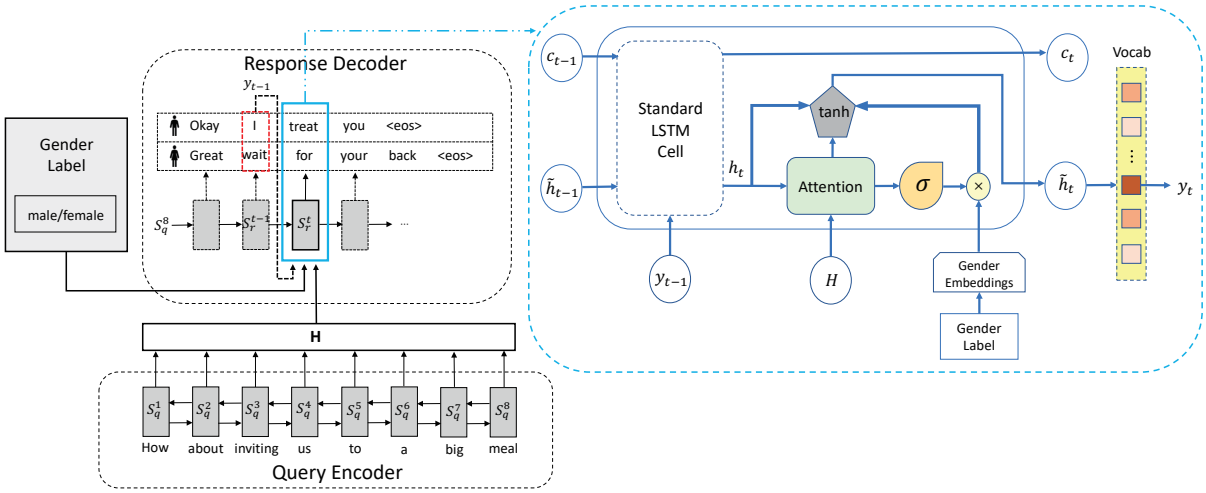

Figure 1: Architecture of our Group Linguistic Bias aware Neural Response Generator (GLBA-NRG).

who propose to encode speaker-specific information and conversation style with user embeddings to influence each generation step. In contrast to launch persona in response generation from individual view (Li et al., 2016), this paper explores to endow chatbots with language styles from the human group aspect.

3 Learning to Generate Linguistic Biased Responses

In this section, we will first formalize the problem, then present the model overview, and finally describe the encoder and decoder architecture of GLBA-NRG.

3.1 Problem Formalization

Our goal is to train an encoder-decoder based model M to generate the response $r = \{y_1, y_2, \cdots, y_j\}$ conditioned on an input query $q = \{x_1, x_2, \cdots, x_n\}$ and the pre-defined **user group label** gl, that is, the training target is to maximize the conditional probability $p(r|q, gl)$. Here, the group label indicates the linguistic biases in the user generated contents.

3.2 Model Overview

Inspired by the research work of Schwartz et al. (2013), who point out that the difference of word distribution of distinct groups is revealed by a few words usage, this paper aims at exploring a mechanism for introducing the linguistic bias into response generating models, and meanwhile controlling the impact of this factor in the generation of each word. They also point out that gender is the most distinguishable feature to split human groups, and thus we take responses from males and females respectively as the corpus for demonstrating our idea. According to the work of Li et al. (2016), it is reasonable to represent users with special embeddings in Seq2Seq based models. This paper follows this set-up by learning the gender embedding g_v and integrating it into our framework.

In our model, a standard Bi-LSTM is taken as the encoder to represent the query q. In this process, the output of the Bi-LSTM is fed into the decoder for response generation. Unlike the decoding procedure in classic Seq2Seq model, we introduce a specially designed neural component to attach to the decoder. This neural component works as a soft-switch gate, converting the attention results based on the hidden layer outputs into a scalar ranging from 0 to 1. Taking the scaled gender embedding and the attention output as parts of inputs, the decoder conducts general steps to generate responses. Figure 1 illustrates the architecture of the proposed response generator.

Our model enjoys several advantages comparing with current response generators. On one hand, through the newly introduced neural component, the linguistic bias information could be adopted into the encoding-decoding process and thus the linguistic biases of different human groups are integrated to the response generation process effectively. On another, our model is able to pick the keywords that reflect different language

styles by dynamically control the impact of the linguistic bias in each generating step. Due to such advantages, our model is expected to generate vivid responses to queries.

3.3 Encoder

As is shown in Figure 1, a query $q = \{x_1, x_2, \cdots, x_n\}$ is fed into the encoder of our model, and projected to a representation vector $H = [h_1^q, \cdots, h_i^q, \cdots, h_n^q]$, where

$$h_i^q = \begin{bmatrix} \overrightarrow{h_i^q} \\ \overleftarrow{h_i^q} \end{bmatrix} \tag{1}$$

The encoding process of the query by bidirectional LSTM (Schuster and Paliwal, 1997) is detailed as follows. First, the forward states $\left(\overrightarrow{h_1^q}, \cdots, \overrightarrow{h_n^q}\right)$ are computed:

$$\begin{bmatrix} i_t \\ f_t \\ o_t \\ l_t \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{bmatrix} \overrightarrow{W} \begin{bmatrix} \overrightarrow{h_{t-1}^q} \\ e_t \end{bmatrix} \tag{2}$$

$$\overrightarrow{c_t} = f_t \odot \overrightarrow{c_{t-1}} + i_t \odot l_t \tag{3}$$

$$\overrightarrow{h_t^q} = o_t \odot \tanh(\overrightarrow{c_t}) \tag{4}$$

where i_t, f_t and o_t indicate the input gate, memory gate and output gate respectively, $e_t \in \mathbb{R}^{1\times m}$ denotes the word embedding for an individual word at time step t, $\overrightarrow{h_t^q} \in \mathbb{R}^{|q|}$ denotes the vector computed by the LSTM model at time t, $\sigma(\cdot)$ is the logistic sigmoid function, and $\overrightarrow{W} \in \mathbb{R}^{4|q|\times(m+|q|)} = \left[\overrightarrow{W_i}, \overrightarrow{W_f}, \overrightarrow{W_o}, \overrightarrow{W_l}\right]$. $\odot$ denotes the element-wise multiplication.

The backward states $\left(\overleftarrow{h_1^q}, \cdots, \overleftarrow{h_n^q}\right)$ are computed similarly. We share the word embedding between the forward and backward LSTMs.

3.4 Group Linguistic Bias Aware Decoder

Basically, with the query representation H inherited from the encoder, our proposed methodology aims at building a decoding mechanism $f(\cdot)$ that is able to systematically adopt both the query semantics and the Group Linguistic Bias (GLB) to generate responses. Formally, this decoding mechanism can be described by the following Equation:

$$\tilde{h_t} = f(h_t, H, e^g) \tag{5}$$

where e^g denotes the dense vector (a.k.a, embedding) representing the human group g and this embedding is designed to imply the group linguistic bias. h_t is the hidden state of the decoder at time step t, and $\tilde{h_t}$ can be taken as an updated hidden state integrated with group linguistic bias. Based on $\tilde{h_t}$, the decoding process follows:

$$p(y_t = w|q, g) = softmax(W_v^\top \tilde{h_t}) \tag{6}$$

where $p(y_t = w|q, g)$ indicates the output word distribution at time step t, and W_v is the weight matrix of the output layer.

The major motivation for proposing the GLBA decoding mechanism is to make the impact of such linguistic bias controllable in the generation of each word in responses. As stated in Schwartz et al. (2013), different human groups differs in use of words in general and thus we can take advantage of such distinct and specified words. Therefore, a model that is capable of highlighting such distinct words for different groups could express group differences effectively.

In this part, we define the specified decoding mechanism $f(\cdot)$ as

$$f(h_t, H, e^g) = W_f[h_t, a_t, e^g \odot g_t] + b_f \tag{7}$$

where W_f and b_f denote the NN related weights and biases respectively. Especially g_t indicates a neural gate transferring the attention outputs upon H into a scalar, so as to control the impact of e^g by performing the element-wise multiplication represented by $\odot$. The operations within the gate g_t can be described by:

$$g_t = \sigma(W_g a_t + b_g) \tag{8}$$

where W_g and b_g denote the weight and bias respectively.

Noticing that Equation 7 and 8 have taken the attention result denoted by a_t, the attention mechanism is formalized as follows:

$$a_t = \sum_{j=1}^{T} \alpha_{tj} h_j \tag{9}$$

$$\alpha_{tj} = \frac{\exp(e_{tj})}{\sum_{k=1}^{T_x} \exp(e_{tk})} \tag{10}$$

$$e_{tj} = W_a[h_t, h_j] + b_a \tag{11}$$

The reason for introducing attention model into the gate is that, intuitively, the impact of the linguistic bias (represented by the group embedding)

on the response words is determined by the semantic of the corresponding query, that is, based on the content of a query, our model is expected to locate the essential words in the response to apply a stronger impact of linguistic bias.

For decoding, the N-best lists are generated using the decoder with beam size B=30. We set a maximum length of 30 for the generated candidates. At each time step of the decoding process, we first examine all $B \times B$ possible next-word candidates, and add these next-word probabilities up to the corresponding hypothesis' joint probability, which contain all the previous words' probabilities in a certain hypothesis. After that, the candidate words are sorted by their joint probabilities and pick out the new top-B unfinished hypotheses and move to the next word position. If any hypothesis meets an *EOS* token, this hypothesis will be added to the result set as one of finished response.

4 Experiments

4.1 Data Preparation

For validating the capability to integrate group linguistic bias into response generation of our model, the dataset should possess group attributes. Therefore, we collected data from one of Chinese real-name social network sites (SNS), in which some utterances have explicit group attributes (e.g. gender, age, etc). We obtained about 240,000 sessions with multi-turn conversations and user profiles from the SNSs. We filtered out potential advertisements, forwards and non-original utterances (including queries and responses), and only kept Chinese words, English letters and digits in each utterance. After the above preprocessing, there are about 5 million query-response pairs remained. The number of words (sequence length) of query or response ranges from 1 to 30.

Each query-response pair has 4 parts of basic information $\langle q, q_u, r, r_u \rangle$, where q is a query, r is the response corresponding to q, and q_u or r_u indicates user ID who posted the query or response. If the profile of r_u is accessible from his or her home page, we tagged the query-response pair with *group linguistic label* obtained from r_u's profile. According to the *group linguistic label*, the 5 million query-response pairs were split into two subsets: one subset includes 4 million pairs without *group linguistic label*, the other is composed of about 1 million labeled pairs. We take the 4 million pairs to pre-train the models, and the details will be given in Subsection 4.3. For the 1 million labeled pairs that used in group linguistic bias experiments, we firstly sampled 2,000 labeled pairs as testing data, and then sampled training and validation data from the remaining pairs. The sizes of training and validation sets are illustrated in Table 2.

Train	male	483,228
	female	482,915
Valid	male	8,052
	female	8,091

Table 2: Data Description

Notice that there is no overlap among pairs in training, validation, and testing sets.

4.2 Baselines

We consider the following baselines in our experiments.

S2S: the standard Seq2Seq model (Sutskever et al., 2014; Bahdanau et al., 2014).

GLBA-Static: to verify the effectiveness of the gate in GLBA-NRG, we keep the attention module but remove the gate which is specially designed to weight gender embeddings. Thus, the attention module only contributes to the output, with no effect on the gender embeddings. The gender embeddings are injected into the decoder as their weights equal 1.0. This baseline is an variant of our GLBA-NRG model. It should be noted that, the GLBA-Static model is **equivalent** to the speaker model proposed by (Li et al., 2016).

For the sake of comparison, we rename our GLBA-NRG model as GLBA-Dyna in order to distinguish from GLBA-Static.

4.3 Training Protocols

Pre-Training: The 4 million query-response pairs **without group label** were utilized to pre-train the basic Seq2Seq model, to initialize the LSTM parameters including baselines and our approach. This is based on the following considerations:

- To obtain better word embeddings benefiting from a bigger dataset, which is trained from random initialization;
- To accelerate convergence in the following experiments since parameters are initialized by a raw Seq2Seq conversation system (Erhan et al., 2010);

The S2S, GLBA-Static and GLBA-Dyna all follow the training protocols below: 1) the encoder is a 2-layer Bi-LSTM network with 1,000 hidden cells for each layer; 2) the decoder is a 1-layer unidirectional LSTM network with 1,000 hidden cells; 3) the batch size is set to 128; 4) use Adam optimizer with a fixed learning rate 0.0001; 5) parameters are initialized by sampling from the uniform distribution [-0.1, 0.1]; 6) gradients are clipped to avoid gradient explosion with a threshold of 1; 7) the vocabulary size is limited to 100,000; 8) the dimensions of word and gender embeddings are both 500.

In our experiments, S2S, GLBA-Static and GLBA-Dyna use the same dataset, which consists of 1 million query-response pairs **with gender labels** (no overlapping with the 4 million query-response pairs for pre-training). The details of train/valid splitting are described in Table 2. Notice that when applying the dataset in S2S, we only use query-response pairs and ignore group labels.

4.4 Evaluation Methods

According to (Liu et al., 2016), the perplexity and BLEU metrics are not suitable for evaluating the response generators, although they are widely used in translation evaluation. Hence, we only use the human judgement in our experiment.

We recruit 3 annotators for human evaluation. The annotators are instructed to judge responses from 2 aspects, response quality and accuracy.

Response Quality: The response quality refers to whether a response is appropriate and attractive to the input query. Three levels are assigned to a response with scores of 0, +1, +2:

- **Attractive (+2)**: the response is evidently a vivid and informative response to the query;
- **Neutral (+1)**: the response is plain and general but suitable to the query;
- **Unsuitable (0)**: it is hard or impossible to find a scenario where the response is suitable.

To make the annotation task operable, the suitability of generated responses is judged from the following four criteria:

(a) **Grammar and Fluency**: Responses should be natural language and free of any fluency or grammatical errors;

(b) **Logic Consistency**: Responses should be logically consistent with the test query;

(c) **Semantic Relevance**: Responses should be semantically relevant to the test query;

(d) **Vividness**: Responses are vivid and information-rich but should not contradict the first three criteria;

If any of the first three criteria (a), (b), and (c) is contradicted, the generated response should be labeled as "Unsuitable". The responses that conform to the first three criteria (a), (b), and (c) but general or flat should be labeled as "Neutral". The responses that completely satisfy the four criteria (a), (b), (c), (d) should be labeled as "Attractive".

Accuracy: Besides the measure of response quality, we also consider whether a response corresponds to its expected group category (as input to the model). For GLBA-NRG models (GLBA-Static and GLBA-Dyna), we ask the annotators to provide a rating score 0, 1 for the judgement.

S2S, GLBA-Static and GLBA-Dyna generate $B = 30$ responses separately using the beam search algorithm described in Section 3.4. Responses generated by different models are pooled and randomly shuffled for each annotator.

4.5 Results & Analysis

Table 3 shows an overall evaluation by calculating the average score of the generated responses. It is clear that GLBA-Dyna outperforms the baseline models, whose average score is up to 1.404, while others' scores are both below 1.0. This means that the responses generated by GLBA-Dyna are grammatical and query-relevant, and also possess vividness which is crucial for making chatbots attractive to users.

Method	Average Score
S2S	0.923
GLBA-Static	0.944
GLBA-Dyna	**1.404**

Table 3: Average Score of Human Evaluation.

Table 4 details the human evaluation scores of generated responses from all approaches in this paper. Compared with S2S, GLBA-NRG models (GLBA-Static and GLBA-Dyna) achieve higher scores on responses labeled as "+2", especially GLBA-Dyna. This phenomenon demonstrates that GLBA-NRG models generate more vivid and informative responses. The improvement on vividness is ascribed to the group linguistic bias of GLBA-NRG models.

Method	Score		
	0	+1	+2
S2S	14.56%	78.56%	6.88%
GLBA-Static	17.56%	70.50%	11.94%
GLBA-Dyna	**8.22%**	**43.11%**	**48.67%**

Table 4: Human annotation results for **responses quality**. The grade evaluation criteria is described in detail in Section 4.4.

As illustrated in Table 4, in contrast to S2S, GLBA-Dyna increases 41.79% on "+2" responses and reduces 35.45% on "+1" ones, while GLBA-Static achieves ∼5% improvement on "+2" responses. Since both GLBA-Static and GLBA-Dyna introduce the group linguistic features as inputs, this phenomenon is ascribed to the different strategies of controlling group linguistic biases. That means the proposed mechanism biasing the general S2S probability distribution is more effective for incorporating the linguistic features, which renders the responses vivid. On one hand, the proposed mechanism dynamically explore possible positions for keywords that implied gender, on the other hand, it is dynamically aware of which keyword is suited in such a position. Under this mechanism, GLBA-Dyna could select out lively and cute words to make responses vivid.

All models have a proportion of unsuitable responses (labeled as "0") ∼10% but GLBA-Static generates more bad responses (17.56%). After checking its bad responses, we find that GLBA-Static tends to generate swear words for most queries as male, and tends to generate "Uh Hmm" for most queries as female. This observation could be ascribed to the fact that the GLBA-Static model takes the external bias (as the gender embedding) as an input augmentation in every time-step of response generation without dynamic switch. Since only the keywords in one response need such external bias to demonstrate the group distinction, it's unwise to apply external bias in the whole process of response generation. In other words, overweighted group bias excessively intervene in the word distribution when decoding, which leads to less correlation between the response and query. Therefore, the responses from GLBA-Static have no much remarkable variation in the quality compared with S2S.

To validate the capability of our model on generating group linguistic biased responses, this paper evaluates whether the gender inferred from the generated response is consistent with the predefined gender label. Table 5 illustrates the evaluation results.

Method	Accuracy
S2S	-
GLBA-Static	0.340
GLBA-Dyna	0.493

Table 5: Gender Consistency Results.

It can be seen that the proposed model GLBA-Dyna achieves 49.3% on accuracy, which indicates that half of the responses generated by GLBA-Dyna are consistent with the input gender linguistic bias. Comparing with the 34.0% accuracy of GLBA-Static, our model GLBA-Dyna is more effective on controlling the gender linguistic bias in response generation. The reason is that GLBA-Dyna moderates the gender embedding information and dynamically regulates three factors' weights, current hidden state of decoder, query context and gender embeddings, to produce more suitable and reasonable responses. Instead, GLBA-Static deactivates the gate so it can not control the external bias spontaneously.

4.6 Case Study

As is shown in Figure 2, S2S could generate responses relevant to queries but very few of them is appealing to users since basically they are acknowledgements of the queries and do not make any new points to interact with the users. In comparison, the responses generated by both GLBA-Static and GLBA-Dyna are more interesting while keeping the relevance with the queries. We find that GLBA-Dyna generates responses with richer information, compared with GLBA-Static. More importantly, GLBA-Dyna responses exhibit prominent distinction of language styles between males and females, on both semantic and syntactical level. We believe this attributes to the gender distinction aware dynamic switch mechanism in GLBA-Dyna decoder.

In Figure 3, we visualize the gender gate of each time-step in response generation, to illustrate the dynamic controlling process of the gender distinction in GLBA-Dyna model. The different shades of red reflect the degrees of the gate activation. It can be seen that the gender gate is strongly activated in the position where the keywords in-

Query	Model		Reponse(Original)	Response(Translated)
我这早零下了。 It has been below zero degrees for long.	S2S		我这零下十几度。	It is around minus 10 degrees here.
	GLBA-Static	Female	可怜的孩纸。	Poor child…
		Male	这么冷。	That is so cold.
	GLBA-Dyna	Female	这么冷啊， 我们还没开始穿羽绒服呢。	Wow that is so cold. We do not even need to wear down jackets here.
		Male	这么冷，注意保暖。	That is so cold. Keep warm.
OKOK，俺还想晚上麻将。 OK. I would really like to play Mahjong tonight.	S2S		我也想啊，想你们啊。	I would like to do so. I really miss you.
	GLBA-Static	Female	好的好的，等你回来。	Yeah let us play it once you are back.
		Male	来吧，哥。	Yeah let us do it, dude.
	GLBA-Dyna	Female	好滴好滴， 等你回来， 我们一起去玩。	That is great! We can play it together once you are back.
		Male	来吧，热烈欢迎， 哥招待你。	Yeah let us do it bro. Do you wanna play it at my place?
生病好难受啊， 健康的时候一定要快快乐乐的， 好好珍惜。 It feels so bad to get ill. We should cherish the life when we are healthy.	S2S		嗯哪。	Yeah it is.
	GLBA-Static	Female	嗯嗯，谢谢亲爱的。	Thank you, darling.
		Male	恩，谢谢关心。	Thanks for asking.
	GLBA-Dyna	Female	嗯嗯，知道了， 亲爱的。	Ok I see, my dear.
		Male	恩，谢谢兄弟关心啊。	Thanks for asking, dude.

Figure 2: Cases selected from the testing set.

Query	Gender	Response									
我这早零下了。	Female	这么	冷	啊	我们	还	没	开始	穿	羽绒服	呢
	Male	这么	冷	注意	保暖						
OKOK，俺还想晚上麻将。	Female	好滴	好	滴	等	你	回来	我们	一起	去	玩
	Male	来	吧	热烈欢迎	哥	招待	你				
生病好难受啊，健康的时候一定要快快乐乐的，好好珍惜。	Female	嗯	嗯	知道	了	亲爱	的				
	Male	恩	谢谢	兄弟	关心	啊					

Figure 3: Gate activation of each time step by GLBA-Dyna.

fer gender distinguishable information. The fact that the gate value of the same word varies with gender label corroborates the effectiveness of dynamic gate activation. In other words, our gate is able to use gender information to control the response generation.

5 Conclusion

In this paper, we have presented a group linguistic bias aware neural response generation model, so as to tackle the talking style customization problem in chatbot implementation. The contributions of our work can be summarized as follows.

a) Instead of modeling and adopting the language style of each individual, this paper proposes to learn the linguistic biases of human groups and introduce such biases into the response generator, which makes the style in responses more explicit and reliable;

b) We have designed a special neural component that is able to dynamically control the impact of the introduced group linguistic bias in each generation step, to select the keywords reflecting language styles, rather than rigidly enforcing linguistic bias for each word.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of semantics* 9(1):1–26.

Rami Alrfou, Marc Pickett, Javier Snaider, Yunhsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking .

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

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research* 11(Feb):625–660.

Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-lm: A neural language model for customizable affective text generation. *arXiv preprint arXiv:1704.06851* .

Liang Hu, Jian Cao, Guandong Xu, Longbing Cao, Zhiping Gu, and Wei Cao. 2014. Deep modeling of group preferences for group-based recommendation. In *AAAI*. volume 14, pages 1861–1867.

Vlad Serban Iulian, Klinger Tim, Tesauro Gerald, Talamadupula Kartik, Zhou Bowen, Bengio Yoshua, and C. Courville Aaron. 2017. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*. pages 3288–3294.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and William B. Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023* .

Anne Maass, Daniela Salvi, Luciano Arcuri, and Gün R Semin. 1989. Language use in intergroup contexts: the linguistic intergroup bias. *Journal of personality and social psychology* 57(6):981.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing* 45(11):2673–2681.

H Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Stephanie M Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin EP Seligman, et al. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one* 8(9):e73791.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*. pages 3295–3301.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *Computer Science* .

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. *arXiv preprint arXiv:1506.06714* .

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

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869* .

Bowen Wu, Baoxun Wang, and Hui Xue. 2016. Ranking responses oriented to conversational relevance in chat-bots. In *COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan*. pages 652–662.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2016. Incorporating loose-structured knowledge into conversation modeling via recall-gate lstm .

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with intention for a neural network conversation model .

Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017a. Mechanism-aware neural machine for dialogue response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

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

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. *EMNLP16* .