# Dual-factor Generation Model for Conversation

RUQING ZHANG, JIAFENG GUO, YIXING FAN, YANYAN LAN, and XUEQI CHENG,
University of Chinese Academy of Sciences, Beijing, China; CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

The conversation task is usually formulated as a conditional generation problem, i.e., to generate a natural and meaningful response given the input utterance. Generally speaking, this formulation is apparently based on an oversimplified assumption that the response is solely dependent on the input utterance. It ignores the subjective factor of the responder, e.g., his/her emotion or knowledge state, which is a major factor that affects the response in practice. Without explicitly differentiating such subjective factor behind the response, existing generation models can only learn the general shape of conversations, leading to the blandness problem of the response. Moreover, there is no intervention mechanism within the existing generation process, since the response is fully decided by the input utterance. In this work, we propose to view the conversation task as a dual-factor generation problem, including an objective factor denoting the input utterance and a subjective factor denoting the responder state. We extend the existing neural sequence-to-sequence (Seq2Seq) model to accommodate the responder state modeling. We introduce two types of responder state, i.e., discrete and continuous state, to model emotion state and topic preference state, respectively. We show that with our dual-factor generation model, we can not only better fit the conversation data, but also actively control the generation of the response with respect to sentiment or topic specificity.

CCS Concepts: • **Computing methodologies → Discourse, dialogue and pragmatics**;

Additional Key Words and Phrases: Conversation, dual-factor generation, responder state modeling

**ACM Reference format:**
Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Dual-factor Generation Model for Conversation. *ACM Trans. Inf. Syst.* 38, 3, Article 31 (June 2020), 31 pages.
https://doi.org/10.1145/3394052



This article is an extension of Reference [55]. The previous conference version aims to handle different mapping mechanisms between utterance-response pairs with respect to their specificity relation. Compared with the previous work, we generalize it into wider scenarios. We view the conversation task as a dual-factor generation problem and introduce to consider the subjective factor of the responder, for better fitting the conversation data and controlling the generation of the response with respect to various responder states. It also includes an extensive experimental assessment of the new model and compares the performance with the state-of-the-art baselines.

This work was supported by Beijing Academy of Artificial Intelligence (BAAI) under Grants No. BAAI2019ZD0306, and funded by the National Natural Science Foundation of China (NSFC) under Grants No. 61722211, 61773362, 61872338, and 61902381, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102, the National Key RD Program of China under Grants No. 2016QY02D0405, the Lenovo-CAS Joint Lab Youth Scientist Project, and the Foundation and Frontier Research Key Program of Chongqing Science and Technology Commission (No. cstc2017jcyjBX0059).

Authors' addresses: R. Zhang, J. Guo (corresponding author), Y. Fan, Y. Lan, and X. Cheng, University of Chinese Academy of Sciences, Beijing, China; CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, NO. 6 Kexueyuan South Road, Haidian District, Beijing, China, 100190; emails: {zhangruqing, guojiafeng, fanyixing, lanyanyan, cxq}@ict.ac.cn.


1046-8188/2020/06-ART31 $15.00
https://doi.org/10.1145/3394052

## 1 INTRODUCTION

Automatic human-machine conversation is believed to be one of the most challenging problems in Artificial Intelligence (AI), where users have a conversation with a computer program using natural languages or voice commands or both in open domain [30, 39]. Generally speaking, having conversations in natural languages is one of the basic ways to communicate. For years, chatbot-style human-machine conversation systems (e.g., XiaoIce[1] and ALIME[2]) as well as academia have paid much attention to improve conversational AI using natural languages, due to its entertainment and functional roles or commercial values in real-world applications.

Early works in open-domain conversation focused on the rules-based methods, which usually rely on manual effort in designing rules and thus hardly apply to non-task-specific and chat-style conversation. Recently, along with the increasing popularity of social media (e.g., Twitter[3] and Weibo[4]), community question-answering platforms (e.g., Baidu Zhidao[5] and Yahoo! Answers[6]) and other web resources, a massive collection of natural human-human conversations are available on the public web. The big data era accelerates fast progress of conversational research, and studies begin to develop data-driven approaches, which can be categorized into two folds, i.e, retrieval-based models [14, 51, 52] and generation-based models [37, 39, 42]. When receiving a response request, the retrieval-based models try to find an existing response from a pre-constructed conversational history repository. Although the retrieved responses are fluent and informative, the performance is limited by the capacity of the response repository. The retrieval-based models lack the flexibility, because the set of responses of a retrieval system is fixed once the historical response repository is constructed in advance.

To make a response highly coherent with respect to the utterance, a better way is to develop the generation-based models for conversation from end to end. Generation-based models formulate the conversation task as a conditional generation problem, i.e., to generate a highly appropriate new sentence as the response given an input utterance. To model conversation as conditional generation, a widely adopted approach is motivated by the previous studies in statistic machine translation (SMT) [40], which translate from one language to another. Specifically, a Sequence-to-Sequence (Seq2Seq) model in which two Recurrent Neural Networks (RNNs) are used as the encoder and the decoder, respectively, is learned to "translate" an input utterance into a response. Generation-based models are developing fast and bring the results of good flexibility and quality.

Generally speaking, the formulation of the generation-based methods is based on an oversimplified assumption that the response is solely dependent on the input utterance (i.e., an objective factor). In human-human conversation, however, the subjective factor of the responder, e.g., his/her emotion or knowledge state, usually plays an important role in a conversation session. The phenomenon is supported by our observation on everyday conversation between humans. We show a motivation case to understand that responders often actively control the responses depending on their own response purpose (which might be affected by a variety of underlying subjective factors such as their current mood, knowledge state, and so on) in Figure 1. Given an utterance, "Do you know a good eating place for Australian special food?," we see that the responder may provide a positive response (i.e., "I know several wonderful places in the downtown") if he/she likes the conversation, or a negative response (i.e., "I do not like Australian food") if he/she dislikes the

[1] http://www.msxiaoice.com/.
[2] https://www.alixiaomi.com.
[3] https://twitter.com/.
[4] http://www.weibo.com/.
[5] https://zhidao.baidu.com.
[6] https://answers.yahoo.com.

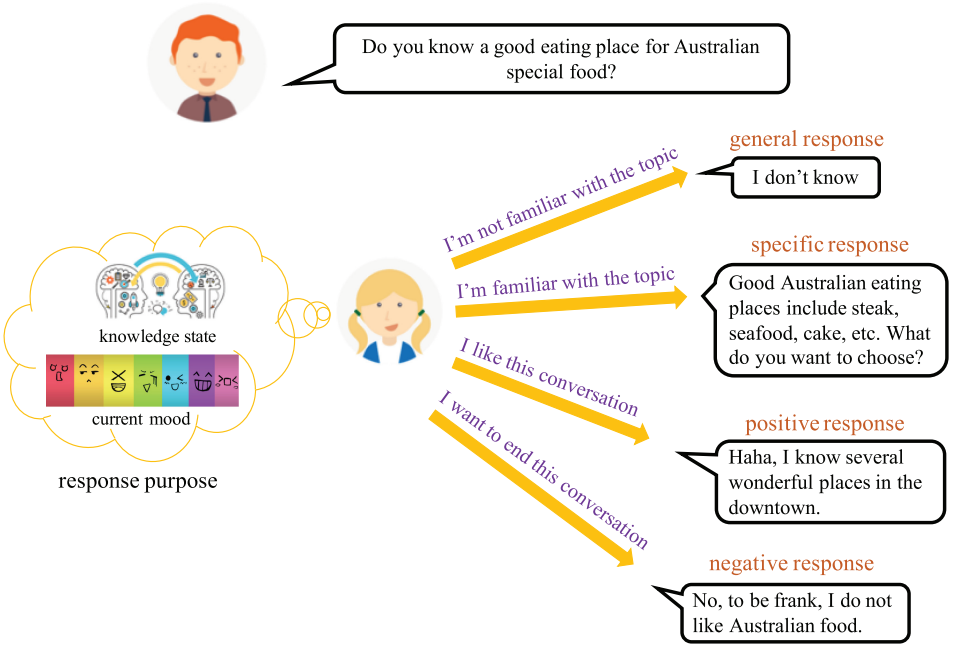



Fig. 1. A motivation example for responder state modeling from human conversation process.

topic. Alternatively, the responder may provide a very specific response (i.e., "Good Australian eating places include steak, seafood, cake, etc.") if he/she is familiar with the topic of "Australian food," or just some general response (i.e., "I don't know") if not. In summary, given an input utterance, different responder states (e.g, the emotion or knowledge state of the responder) may lead to quite different responses from the responder. It means that a one-to-many relationship between an utterance to its responses actually exists in human-human conversation. According to the observation data collected from the real world, the responses are indeed affected by two factors, i.e, the input utterance and the responder state.

However, without explicitly differentiating such subjective factor and objective factor, there exist two issues in existing generation-based models. First, existing generation-based models employ a one-fits-all model to capture the one-to-many relationship between an utterance and its responses in open-domain conversation. Thus, these models tend to learn the general shape of the conversations and will inevitably favor common and meaningless responses with high frequency—namely, the blandness problem. Although these responses are safe for replying different utterances, they may quickly lead to an end of the conversation, severely hurting the user experience of a chatbot. Second, there is no intervention mechanism or variable as restricted by a human within existing generation process—namely, the intervention problem. The reason is that the response is completely decided by the input utterance without other factors. Hence, it is unable to actively control the generation of the response, which can not well meet responders' needs and improve responders' satisfaction.

There have been a few efforts attempting to address these two issues in literature. For the blandness issue, Reference [17] proposed to use the Maximum Mutual Information (MMI) as the objective to penalize general responses. It could be viewed as a post-processing approach, which did not solve the generation of trivial responses fundamentally. Reference [47] proposed a joint attention mechanism model that modified the generation probability by adding the pre-defined topic

keywords "likelihood" to the "maximum likelihood." However, it is difficult to ensure that the topics learned from the external corpus are consistent with that in the conversation corpus, leading to the introduction of additional noises. For the intervention issue, the work in Reference [57] resolved it to some extent. Reference [57] assumed that there exist some latent responding mechanisms for an input utterance and introduced latent responding factors to model multiple latent responding mechanisms. However, the response is intrinsically dependent on the input utterance, because the probability of latent responding mechanism is only conditioned on the input. Also, these latent factors are usually difficult in interpretation and it is hard to decide the number of the latent factors. Overall, these works cannot fundamentally solve the blandness and intervention issue, since they fail to explicitly model the subjective factor of the responder.

In this article, we formulate the conversation task as a dual-factor generation problem and propose a novel Dual-Factor Generation (DFGen) model to solve this problem. Given an input utterance, we aim to produce the response based on two factors, i.e., an objective factor denoting the input utterance and a subjective factor denoting the responder state. We employ a neural sequence-to-sequence framework and further introduce an explicit state variable to accommodate the responder state modeling. Specifically, we introduce two types of responder state, namely, discrete state and continuous state, to model emotion state and topic preference state, respectively. Meanwhile, we assume that each word, beyond the semantic representation that relates to its meaning, also has another representation that relates to the usage preference under different responder states. We name this representation as the usage representation of words. To mimic different responder states, we employ two types of intervention mechanism, i.e., discrete-based intervention mechanism and continuous-based intervention mechanism. In this way, we can not only better fit the intrinsic shapes of conversation data, but also actively control the response generation with respect to sentiment or topic specificity by varying the state variable. To the best of our knowledge, so far there have been no works to simultaneously consider the objective factor and subjective factor that affect the response generation in a general way.

We conduct an empirical study on two large public conversation datasets and compare our model with several state-of-the-art response generation methods. Empirical results show that our model is capable of responding to utterances within sentiment and topic specificity intervention as restricted by a human and significantly outperform existing methods under both automatic and human evaluations. We also provide detailed analysis on DFGen model and conduct case studies to verify the patterns captured by our model over different responder states.

## 2 RELATED WORK

In this section, we briefly review the related work on the traditional conversation models, diversity-enhanced conversation models, persona-based conversation models, and emotion-aware conversation models.

### 2.1 Traditional Conversational Models

Automatic conversation has attracted increasing attention over the past few years. At the very beginning, people started the research using hand-crafted rules and templates [11, 43, 45]. These approaches required little data for training but huge manual effort to create enough handcraft rules or templates, which is very time-consuming. For now, with the prosperity of social media, forums, and other web resources, people begin to pay more attention to data-driven methods instead of human-driven methods. Most modern conversational models fall into two major categories: retrieval-based and generation-based.

*2.1.1 Retrieval-based Conversational Models.* Retrieval-based conversational models search the most suitable response from candidate responses using different schemas [14, 15, 44, 46, 48–50, 52, 59]. Most retrieval-based methods could be decomposed into two steps: (1) retrieve a set of candidate responses using basic retrieval models, e.g., BM25 [34]; and (2) re-rank the candidate responses with neural ranking models to find the best matching response. Reference [44] proposed a retrieval-based response model for short-text-based conversation to leverage the huge samples collected from social media. Reference [48] proposed a retrieval-based conversation system with the deep learning-to-respond schema through a deep neural network framework driven by web data. Reference [52] considered to leverage external knowledge into the matching process of dialogue context and candidate responses for response ranking. Although the retrieved responses are fluent with great diversity, these approaches lack the flexibility, since the pool of possible responses is constructed in advance (e.g., pre-existing human responses). Thus, retrieval systems may fail to return appropriate responses for those unseen input utterances [9] and are difficult to be extended to open domains.

*2.1.2 Generation-based Conversational Models.* In recent years, generation-based conversation systems have greatly advanced with the help of deep learning and reinforcement learning techniques [4, 7, 10, 20, 32, 33, 35, 37, 39, 41, 42]. The basic neural encoder-decoder framework for generation-based conversation models is actually developed with a Statistic Machine Translation (SMT) framework [3, 6, 40]. Reference [40] proposed the original Seq2Seq framework, which used multi-layered LSTM as the encoder and another deep LSTM as the decoder for machine translation. Reference [3] introduced the attention mechanism into the neural network to improve the performance of SMT for long input sentences. Along the way of neural SMT, many recent studies showed that these models can also be successfully used in conversation modeling. Reference [39] presented a novel response generation model that can be trained end to end on large unstructured Twitter conversations. Reference [37] explored using encoder-decoder-based neural network with local and global attentions to generate replies for short-text conversation. Together with the contemporaneous work [42], these papers first proposed the neural approaches to fully end-to-end conversation modeling. Later, Reference [35] built an end-to-end dialogue system using a hierarchical recurrent neural network generative model. Reference [10] introduced copynet into the sequence-to-sequence learning framework to simulate the repeating behavior of humans in conversation. Reference [20] introduced a reinforcement learning framework for neural response generation by integrating the strengths of neural Seq2Seq systems and reinforcement learning for conversation. Gradually, researchers introduce various elements into conversation generation, such as diversity, persona, and emotion.

## 2.2 Diversity-enhanced Conversation Models

Although the current Seq2Seq model has the ability to generate fluent responses, one serious problem is that the generated responses are usually general (e.g., "I don't know" or "I'm OK"). Some recent studies began to focus on improving the generation quality, i.e., generating less bland and more specific responses. It is also called a diversity problem, since if each response is more specific, it would be more diverse between responses of different utterances.

As an early work, Reference [17] used Maximum Mutual Information (MMI) as the objective function for conversation response generation. It is not a unified training model. Instead, it still trained a maximum likelihood model and used the MMI criterion only for testing to rerank the top-n candidates generated by Seq2Seq. Reference [19] proposed a data distillation method, which trains a series of generative models at different levels of specificity and uses a reinforcement

learning model to choose the model best suited for decoding, depending on the conversation context. These methods circumvented the general response issue by using either a post-processing approach or a data selection approach. Reference [28] proposed a forward-backward keyword method that used a pointwise mutual information to predict a noun as a keyword and then used two Seq2Seq models to generate the forward sentence and the backward sentence. Reference [53] attempted to improve the specificity with the reinforcement learning framework by using the averaged IDF score of the words in the response as a reward. Reference [38] presented a conditional variational framework for generating specific responses based on specific attributes. Also, some works (e.g., such as seqGAN [54] and Adver-REGS [21]) try to use Generative Adversarial Networks (GAN) for generation, where the discriminator scores are used as rewards for reinforcement learning. However, the meaning of this reward function is not clear. Reference [36] presented a latent Variable Hierarchical Recurrent Encoder-Decoder (VHRED) model to explicitly model generative processes with multiple levels of variability. Moreover, Reference [57] assumed that there exists some latent responding mechanisms and introduced latent responding factors to the Seq2Seq model to avoid generating safe responses. However, these latent factors are usually difficult in interpretation and hard to decide the number. Reference [47] incorporated the topic information from an external corpus into the Seq2Seq framework to guide the generation. It has to train an extra LDA model from an extra corpus to generate the topic keyword candidates. However, external dataset may not always be available or consistent with the conversation dataset in topics.

### 2.3 Persona-based Conversation Models

In fact, conversational datasets characterize multiple speakers, which often have different or conflicting personas and backgrounds. There have also been some recent studies on persona-based conversation models to be capable of adapting to different kinds of users. Reference [18] tried to build a persona-based conversation engine, including a single-speaker speaker model and a dyadic speaker-addressee model to generate personalized responses. Reference [1] personalized the prediction of responses depending on each user's personal history across all the conversations in which he or she participated in. Reference [24] described an interesting approach that uses multitask learning approach to train neural conversation models. It leverages both conversation data across speakers and other types of data pertaining to the speaker. Reference [31] added intermediate supervision to detect whether a profile should be used when responding to an utterance. Reference [25] presented a personalized end-to-end model for goal-oriented dialogues by incorporating the profile vector and using conversation context from users. There are some works attempting to introduce personalized information to dialogues by transfer learning [27, 56]. Different from prior persona-based works using external personal profiles that are difficult in interpretation and controllability, our proposed model introduces an explicit state variable that can not only interpret the responder state but also control the response generation.

### 2.4 Emotion-aware Conversation Models

Generating diverse emotional responses is becoming increasingly popular as a new research frontier in Natural Language Processing (NLP). To model the emotion influence in large-scale conversation generation and enrich communication, many emotion-aware conversation models have recently been proposed. Reference [58] used embedding emotion categories, together with capturing the change of implicit internal emotion states and explicit emotion expressions with external emotion vocabulary to generate responses that are both contextually and emotionally appropriate. Reference [60] constructed a novel corpus using Twitter conversations with emojis in the response and employed several conditional variational autoencoders to use emojis to control the emotion of

the generated text. However, most conversation datasets lack the fine-grained emojis, which can be seen as the natural label for the emotion of the response, and thus this model can not be applied to most conversation datasets. Reference [12] concatenated the desired emotion with the source input during the learning, or pushed the emotion in the decoder. It seems that the model is not able to capture the precise information carried by the additional emotion category, since it is treated as just one additional word to concatenate with the input utterance. Reference [2] advanced the emotional development of affectively neural encoder-decoder dialogue systems by three affective strategies, namely, affective word embeddings, affect-based objective functions, and affectively diverse beam search. Reference [23] encoded the emotion state of the conversation as distributed embedding into the process of response generation and then introduced a re-rank function to select the appropriate response. However, these two methods that simply copy and use the emotion of the input utterance are unable to generate responses of different emotions for the same utterance. Specifically, the model proposed by Reference [2] depends heavily on linguistic resources and needs manual parameter adjustments. Recently, Reference [5] proposed to first determine the appropriate emotion to be included in a respons and then generate the responses with the given emotion. Nevertheless, the probability of each emotion is only conditioned on the input utterance, and thus the response is still fully decided by the input utterance, which is similar with the problem in Reference [57]. Compared with existing emotion-aware conversation methods, our method can explicitly model the emotion information and actively control the response generation in an end-to-end way.

To the best of our knowledge, existing open-domain chatbot-like conversation models usually generate the response based on the objective factor and ignore the explicit modeling of the subjective factor of the responder, which is also a key factor that affects the generation of the response. Unlike these existing methods, we propose to view the conversation task as a dual-factor generation problem and introduce responder state modeling into the existing Seq2Seq model.

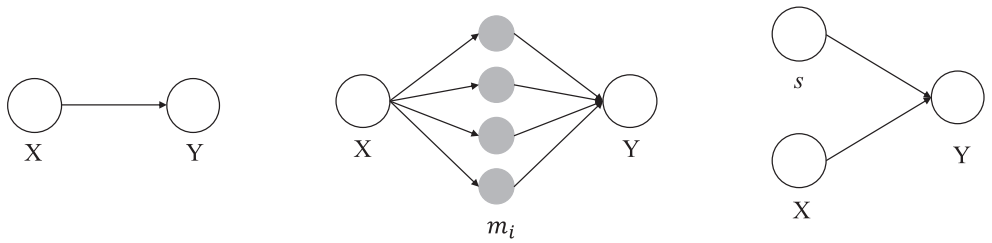
## 3 DUAL-FACTOR GENERATION MODEL

In this section, we present the Dual-Factor Generation (DFGen) model, a novel Seq2Seq-based model to accommodate the responder state modeling for the conversation task. We first give a discussion about our model compared with previous works and an overview of the model architecture. We then describe each component of our model as well as the learning and generation procedure specifically.

### 3.1 Model Discussion

The basic idea of a generative conversational model is to learn the mapping from an input utterance to its response, typically using an encoder-decoder framework. Formally, given an input utterance sequence $\mathbf{X} = (x_1, x_2, \ldots, x_T)$ and a target response sequence $\mathbf{Y} = (y_1, y_2, \ldots, y_{T'})$, a neural Seq2Seq model is employed to learn $p(\mathbf{Y}|\mathbf{X})$ based on the training corpus $\mathcal{D} = \{(\mathbf{X}, \mathbf{Y})|\mathbf{Y}$ is the response of $\mathbf{X}\}$ (as shown in Figure 2(a)). Without explicitly differentiating the underlying factors of the responder that might affect the response, the learned Seq2Seq model cannot capture the 1-to-n relationship between all the utterance-response pairs and will inevitably favor those general responses with high frequency.

Later, Reference [57] considers that the modeling of $P(\mathbf{Y}|\mathbf{X})$ for conversation should be complex enough to represent all the suitable responses. Hence, as shown in Figure 2(b), they assume that there exists some latent responding mechanisms $m_i$ and propose to explicitly model the multiplicity of responding mechanisms by learning $p(\mathbf{Y}|\mathbf{X}, m_i)p(m_i|\mathbf{X})$. However, the probability of the mechanism $p(m_i|\mathbf{X})$ is conditioned on $\mathbf{X}$, and thus the response is still fully decided by the input utterance.




(a) (b) (c)

Fig. 2. (a) A graphical model representation of traditional Seq2Seq model. $X$ is the input utterance and $Y$ is the generated response. (b) A graphical model representation of mechanism-aware Seq2Seq model. $m_i$ is the latent responding mechanism. (c) A graphical model representation of our dual-factor generation model. $s$ is the subjective factor of the responder.

As compared with previous works, we assume that the response depends on not only the input utterance, but also the subjective factor of the responder. Rather than involving some latent responding mechanisms as in Reference [57], we propose to introduce an explicit state variable $s$ into a Seq2Seq model to represent the responder state. As shown in Figure 2(c), the goal of our model becomes to learn $p(\mathbf{Y}|\mathbf{X}, s)$ over the corpus $\mathcal{D}$, where we need labels for $s$ for learning. By doing so, we hope that (1) $s$ would have explicit meaning on the responder state (e.g., emotion or knowledge state) and (2) $s$ could not only interpret but also actively control the generation of the response $\mathbf{Y}$ given the input utterance $\mathbf{X}$.

### 3.2 Model Overview

Formally, given an input utterance $\mathbf{X} = (x_1, x_2, \ldots, x_T)$ with $T$ words and the responder state $s$, DFGen aims to generate an appropriate response $\mathbf{Y} = (y_1, y_2, \ldots, y_{T'})$.

Basically, the DFGen employs the encoder-decoder framework for the task. In the encoding phase, we employ the unified encoder framework to obtain the representation of the input utterance. In the decoding phase, we introduce two types of intervention mechanism, i.e., discrete-based intervention mechanism and continuous-based intervention mechanism, to guide the model to generate responses under different emotion states and topic preference states, respectively. Moreover, DFGen employs a similar attention mechanism as traditional Seq2Seq approaches [3] to help the response generation. We will detail our model as follows.

### 3.3 Encoder

The encoder is to map the input utterance $\mathbf{X}$ into a compact vector that can capture its essential topics. Specifically, we use a bi-directional GRU [6] as the utterance encoder, and each word $x_i$ is first represented by its semantic representation $\mathbf{e}_i$ mapped by semantic embedding matrix $\mathbf{E}$ as the input of the encoder. The specific implementation of GRU is parameterized as:

$$\begin{aligned}
\mathbf{z}_t &= \sigma(\mathbf{W}_z \mathbf{x}_t + \mathbf{U}_z \mathbf{h}_{t-1} + \mathbf{b}_z), \\
\mathbf{r}_t &= \sigma(\mathbf{W}_r \mathbf{x}_t + \mathbf{U}_r \mathbf{h}_{t-1} + \mathbf{b}_r), \\
\widetilde{\mathbf{h}}_t &= \tanh(\mathbf{W}_h \mathbf{x}_t + \mathbf{U}_h (\mathbf{h}_{t-1} \circ \mathbf{r}_t) + \mathbf{b}_h), \\
\mathbf{h}_t &= (1 - \mathbf{z}_t) \circ \widetilde{\mathbf{h}}_t + \mathbf{z}_t \circ \mathbf{h}_{t-1},
\end{aligned} \tag{1}$$

where $\mathbf{x}_t$, $\mathbf{h}_t$, $\mathbf{z}_t$, and $\mathbf{r}_t$ are the input vector, output vector, update gate vector, and reset gate vector, respectively. $\mathbf{W}_z$, $\mathbf{W}_r$, $\mathbf{W}_h$, $\mathbf{U}_z$, $\mathbf{U}_r$, $\mathbf{U}_h$, $\mathbf{b}_z$, $\mathbf{b}_r$, and $\mathbf{b}_h$ are parameter matrices and vectors.

The input to the utterance encoder is the word sequence of a sentence and the utterance encoder sequentially updates its hidden state after receiving each word. The forward GRU in word encoder reads the words in the $t$th word $x_t$ in the left-to-right direction, resulting in a sequence of hidden states $(\overrightarrow{\mathbf{h}_1}, \ldots, \overrightarrow{\mathbf{h}_T})$. The backward GRU reads $x_t$ in the reversed direction and outputs $(\overleftarrow{\mathbf{h}_1}, \ldots, \overleftarrow{\mathbf{h}_T})$. Then, we concatenate the last hidden states of the forward and backward passes as the representation of the word $x_t$, denoted as $\mathbf{h}_t = [\overrightarrow{\mathbf{h}_t} || \overleftarrow{\mathbf{h}_t}]$. The encoder represents the utterance $\mathbf{X}$ as a series of hidden vectors $\{\mathbf{h}_t\}_{t=1}^{T}$ modeling the sequence from both forward and backward directions. Finally, we use $\tanh(\overleftarrow{\mathbf{h}_1})$ as the initial hidden state of the decoder, where $\overleftarrow{\mathbf{h}_1}$ is the final backward hidden state.

### 3.4 Decoder

The decoderx is to generate a response $\mathbf{Y}$ given the hidden representations of the input utterance $\mathbf{X}$ under the responder state denoted by the state variable $s$. Here, we introduce two types of responder state, i.e., discrete and continuous state, to model emotion state and topic preference state, respectively. Specifically, at step $t$, we define the probability of generating any target word $y_t$ by a "mixture" of probabilities:

$$p(y_t) = \beta p_M(y_t) + \gamma p_S(y_t), \tag{2}$$

where $\beta$ and $\gamma$ are the coefficients. $p_M(y_t)$ denotes the semantic-based generation probability affected by the subjective factor, and it decides what to say next given the input utterance. $p_S(y_t)$ denotes the state-based generation probability affected by the objective factor, and it decides how to reply under the responder state. We now describe the specific generation probability as follows:

*3.4.1 Semantic-based Generation Probability.* Specifically, the semantic-based generation probability $p_M(y_t)$ is defined the same as that in the traditional Seq2Seq model [40]:

$$p_M(y_t = w) = \mathbf{w}^{\mathrm{T}}(\mathbf{W}_M^h \cdot \mathbf{h}_{y_t} + \mathbf{W}_M^e \cdot \mathbf{e}_{t-1} + \mathbf{b}_M), \tag{3}$$

where $\mathbf{w}$ is a one-hot indicator vector of the word $w$ and $\mathbf{e}_{t-1}$ is the semantic representation of the $t-1$-th generated word in decoder. $\mathbf{W}_M^h$, $\mathbf{W}_M^e$, and $\mathbf{b}_M$ are learned parameters. $\mathbf{h}_{y_t}$ is the $t$th hidden state in the decoder, which is computed by:

$$\mathbf{h}_{y_t} = f(y_{t-1}, \mathbf{h}_{y_{t-1}}, \mathbf{c}_t), \tag{4}$$

where $f$ is a GRU unit and $\mathbf{c}_t$ is the context vector [3] to allow the decoder to pay different attention to different parts of input at different steps. A natural option is to represent $\mathbf{c}_t$ as a weighted sum of the source hidden states $\{\mathbf{h}_1, \ldots, \mathbf{h}_T\}$, i.e.,

$$\mathbf{c}_t = \sum_{i=1}^{T} \alpha_{ti} \mathbf{h}_i, \tag{5}$$

where $\alpha_{ti}$ indicates how much the $i$th word $x_i$ from the input utterance contributes to generating the $t$th word of the response, and is usually computed as:

$$\alpha_{ti} = \frac{\exp(\mathbf{h}_i \cdot \mathbf{h}_{y_{t-1}})}{\sum_{j=1}^{T} \exp(\mathbf{h}_j \cdot \mathbf{h}_{y_{t-1}})}, \tag{6}$$

where $\mathbf{h}_{y_{t-1}}$ represents the RNN hidden state (just before emitting $y_t$) of the decoder.

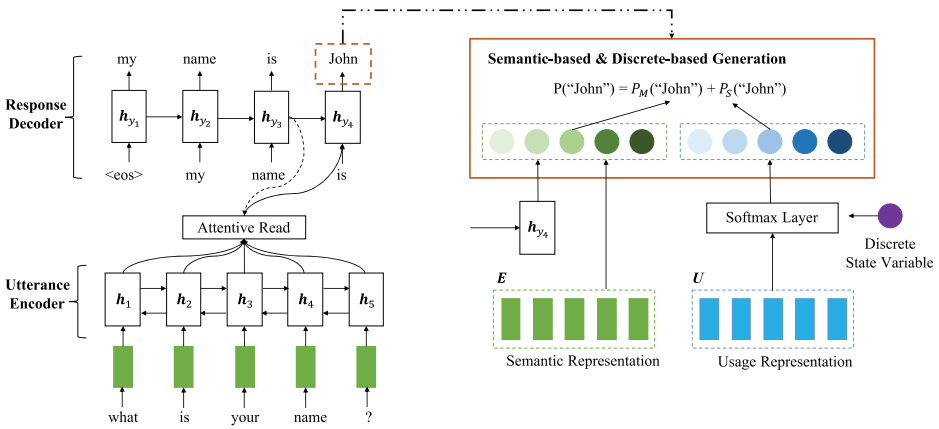



Fig. 3. The overall architecture of our DFGen model with the discrete-based intervention mechanism.

*3.4.2 State-based Generation Probability.* In this work, we introduce two types of responder state, namely, discrete state and continuous state, to model emotion state and topic preference state, respectively. To mimic different responder states, we introduce two types of intervention mechanisms based on the state variable $s$, namely, discrete-based intervention mechanism and continuous-based intervention mechanism. Thus, we achieve two types of state-based generation probability $p_S(y_t)$ in the decoder, namely, discrete-based generation probability and continuous-based generation probability, to generate the responses with different emotion categories or at different topic specificity levels.

- **Discrete-based intervention mechanism.** For the discrete-based intervention mechanism, $p_S(y_t)$ denotes the discrete-based generation probability, i.e., the probability of the target word with respect to the given emotion state. Without loss of generality, the emotion categories are isolated from each other in a certain sense and thus can be quantified by a finite number of values. Hence, $s$ denotes the discrete state variable for modeling the emotion categories over discrete space.

  As shown in Figure 3, we introduce a Softmax layer to define this probability. Specifically, we assume that each word, beyond its semantic representation $\mathbf{e}$, also has a usage representation $\mathbf{u}$ mapped by usage embedding matrix $\mathbf{U}$. The usage representation of a word in the discrete-based intervention mechanism denotes its usage preference in different emotional expressions. Assuming that there exists $K$ emotion categories in the responses, the discrete state variable $s$ then interacts with the usage representations through the Softmax layer to produce the discrete-based generation probability $p_S(y_t)$:

$$\begin{aligned} p_S(y_t = w) &= softmax(\Phi_D(\mathbf{U}, \mathbf{w}))_s, \\ \Phi_D(\mathbf{U}, \mathbf{w}) &= \mathbf{w}^{\mathrm{T}}(\mathbf{U} \cdot \mathbf{W}_U^d + \mathbf{b}_U^d), \end{aligned} \tag{7}$$

  where $\Phi_D$ maps the word usage representation into $K$-dimensionality vector. $\mathbf{W}_U^d \in \mathcal{R}^{M \times K}$ and $\mathbf{b}_U^d$ are parameters to be learned, where $M$ is the dimension of each usage representation $\mathbf{u}$.

  $K$ elements in the vector $\Phi_D(\mathbf{U}, \mathbf{w})$ denote the probabilities of a word belonging to $K$ different emotion categories, respectively. The softmax function is used to normalize all the elements in the vector as a probability distribution. Then, the discrete state variable is used to select the $s$-th element from the vector as the predicted discrete-based generation

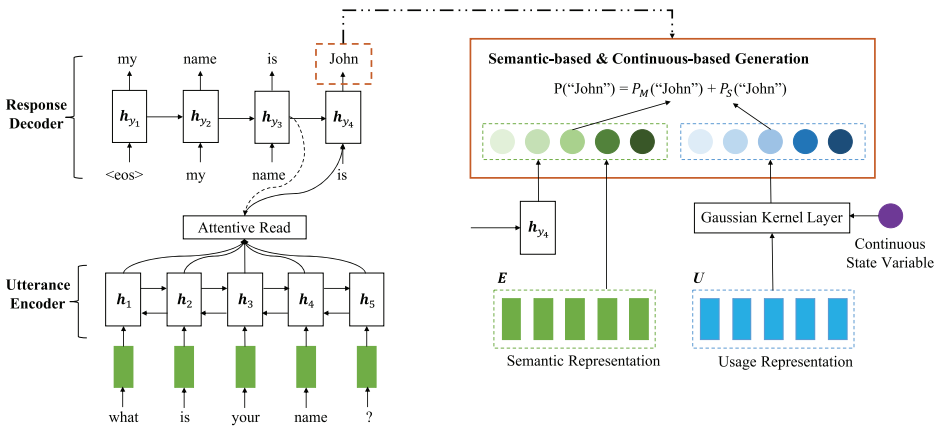



Fig. 4. The overall architecture of our DFGen model with the continuous-based intervention mechanism.

probability belonging to the desired emotion. Note here that, in general, we can use any multi-classification function to define the $\Phi_D(\mathbf{U}, \mathbf{w})$.

- **Continuous-based intervention mechanism.** For the continuous-based intervention mechanism, $p_S(y_t)$ denotes continuous-based generation probability, i.e., the probability of the target word with respect to the topic preference state. The topic specificity level changes from general to specific gradually and thus can be quantified by an infinite number of values between any two values. Therefore, $s$ denotes the continuous state variable for modeling the topic specificity levels over continuous space.
  As shown in Figure 4, here, we introduce a Gaussian Kernel layer to define this probability. Similar with the discrete-based intervention mechanism, each word in the continuous-based intervention mechanism also has a usage representation $\mathbf{u}$, which denotes its usage preference at different topic specificity levels. Given the continuous state variable $s$, we want the word usage representation to regress to the topic specificity of the response through certain mapping function. Hence, the continuous state variable $s$ interacts with the usage representations through the Gaussian kernel layer to produce the continuous-based generation probability $p_S(y_t)$:

$$p_S(y_t = w) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(\Psi_C(\mathbf{U}, \mathbf{w}) - s)^2}{2\sigma^2}\right),$$
$$\Psi_C(\mathbf{U}, \mathbf{w}) = \sigma(\mathbf{w}^{\mathrm{T}}(\mathbf{U} \cdot \mathbf{W}_U^s + \mathbf{b}_U^s)), \tag{8}$$

  where $\sigma^2$ is the variance, and $\Psi_C(\cdot)$ maps the word usage representation into a real value with the continuous state variable $s$ as the mean of the Gaussian distribution. $\mathbf{W}_U^s$ and $\mathbf{b}_U^s$ are parameters to be learned.

  Note here that, in general, we can use any real-value function to define $\Psi_C(\mathbf{U}, \mathbf{w})$. In this work, we use the sigmoid function $\sigma(\cdot)$ for $\Psi_C(\mathbf{U}, \mathbf{w})$, since we want to define $s$ within the range $[0, 1]$ so each end has very clear meaning on the topic specificity, i.e., 0 denotes the most general response, while 1 denotes the most specific response. In the next, we will also keep this definition of each end when we define the distant label for the continuous state variable.
- **Intervention mechanism discussion.** Here, we give a detailed discussion about the discrete-based and continuous-based intervention mechanism to further understand how our model works.

— The common purpose of these two intervention mechanisms is to explicitly accommodate the responder state modeling and control the generation of the response as restricted by a human. This can provide the user with a better overall conversation experience and, in turn, boost the response performance and satisfaction of a conversation system.
— The key difference between the discrete-based and continuous-based intervention mechanism is to model different responder states. In a conceptual way, the discrete and continuous state could cover a majority of responder states. First, other than the topic preference state used in this work, there could be many other continuous states, e.g., humor state, curious state, and imagination state. What they have in common is that we can quantify them on an uncountable set of values between the maximum and minimum. Thus, the basic idea of the continuous-based intervention mechanism is to let the word usage representation regress to the continuous state variable. When we do this under a probabilistic framework, this is equivalent to define a Gaussian kernel layer as in Equation (8). Second, the discrete state could model many other states in addition to the emotion state, such as expression pattern state (e.g., question, imperative, and exclamatory). Different from the continuous state, the discrete state is a kind of statistics quantified on discrete specific values. Therefore, we define a discrete probabilistic framework as in Equation (7) to classify each word into the discrete state variable, i.e., one pre-defined target emotion category. Finally, the regression problem in the continuous-based intervention mechanism is fundamentally different from the classification problem in the discrete-based intervention mechanism. Specifically, we could convert continuous state into a discrete state variable by dividing continuous state ranges into certain blocks. However, it is hard to decide the number of the blocks, and the differences between blocks are not statistically significant. This may result in the limitation of precisely describing different blocks (i.e., categories) and worse performance of the response generation. In addition, the discrete state is not continuous and thus cannot be well fitted by a continuous function. Hence, given a responder state, we should select the corresponding intervention mechanism to mimic it.

### 3.5 Model Learning

We train our DFGen model by maximizing the log likelihood of generating responses over the training set $\mathcal{D}$:

$$\mathcal{L} = \sum_{(\mathbf{X},\mathbf{Y})\in\mathcal{D}} \log P(\mathbf{Y}|\mathbf{X}, s; \theta), \tag{9}$$

where $\theta$ denotes all the model parameters. Note here, since $s$ is an explicit state variable in our model, we need the triples $(\mathbf{X}, \mathbf{Y}, s)$, i.e., the input utterance, the ground-truth target response, and the ground-truth responder state, for training. Thus, we need to acquire the labels for discrete and continuous state variable $s$, respectively, to learn our model.

*3.5.1 Labels for Discrete State Variable.* For the discrete variable, the emotion category of the response is directly available in the public conversation corpus [58]. And thus, we directly use the conversation dataset annotated with emotions, where the emotion category (e.g., *Like, Sadness, Disgust, Anger,* and *Happiness*) is labeled with each utterance and response.

*3.5.2 Distant Supervision for Continuous State Variable.* For the continuous variable used in the continuous-based generation, it is difficult to directly obtain the topic specificity of the responses, and thus, we propose to employ distant supervision to train our model. Specifically, we introduce two ways of distant supervision, namely, Normalized Inverse Response Frequency (NIRF) and Normalized Inverse Word Frequency (NIWF).

- **Normalized Inverse Response Frequency.** Normalized Inverse Response Frequency (NIRF) is based on the assumption that a response is more general if it corresponds to more input utterances in the corpus. Therefore, we use the inverse frequency of a response in a conversation corpus to indicate its topic specificity level. Specifically, we first build the response collection $\mathcal{R}$ by extracting all the responses from $\mathcal{D}$. For a response $\mathbf{Y} \in \mathcal{R}$, let $f_{\mathbf{Y}}$ denote its corpus frequency in $\mathcal{R}$. We compute its Inverse Response Frequency (IRF) as:

$$\text{IRF}_{\mathbf{Y}} = \log(1 + |\mathcal{R}|)/f_{\mathbf{Y}}, \tag{10}$$

  where $|\mathcal{R}|$ denotes the size of the response collection $\mathcal{R}$. Next, we use the min-max normalization method [13] to obtain the NIRF value. Namely,

$$\text{NIRF}_{\mathbf{Y}} = \frac{\text{IRF}_{\mathbf{Y}} - \min_{\mathbf{Y}' \in \mathcal{R}}(\text{IRF}_{\mathbf{Y}'})}{\max_{\mathbf{Y}' \in \mathcal{R}}(\text{IRF}_{\mathbf{Y}'}) - \min_{\mathbf{Y}' \in \mathcal{R}}(\text{IRF}_{\mathbf{Y}'})}, \tag{11}$$

  where $\max(\text{IRF}_{\mathcal{R}})$ and $\min(\text{IRF}_{\mathcal{R}})$ denote the maximal and minimum IRF value in $\mathcal{R}$, respectively. The NIRF value is then used as the distant label of $s$ in training. Note here that by using normalized values, we aim to constrain the continuous state variable $s$ to be within the pre-defined continuous value range [0,1].
- **Normalized Inverse Word Frequency.** Normalized Inverse Word Frequency (NIWF) is based on the assumption that the topic specificity level of a response depends on the collection of words it contains, and the sentence is more specific if it contains more specific words. Hence, we can use the inverse corpus frequency of the words to indicate the topic specificity level of a response. Specifically, for a word $y$ in the response $\mathbf{Y}$, we first obtain its Inverse Word Frequency (IWF) by:

$$\text{IWF}_y = \log(1 + |\mathcal{R}|)/f_y, \tag{12}$$

  where $f_y$ denotes the number of responses in $\mathcal{R}$ containing the word $y$. Since a response usually contains a collection of words, there would be multiple ways to define the response-level IWF value, e.g., sum, average, minimum or maximum of the IWF values of all the words. In our work, we find that the best performance can be achieved by using the maximum of the IWF of all the words in $\mathbf{Y}$ to represent the response-level IWF by

$$\text{IWF}_{\mathbf{Y}} = \max_{y \in \mathbf{Y}}(\text{IWF}_y). \tag{13}$$

  This is reasonable, since a response is specific as long as it contains some specific words. We do not require all the words in a response to be specific, thus sum, average, and minimum would not be appropriate operators for computing the response-level IWF.

  Again, we also use the min-max normalization method to obtain the NIWF value for the response $\mathbf{Y}$, and the NIWF value is then used as the distant label of $s$:

$$\text{NIWF}_y = \frac{\text{IWF}_y - \min_{\mathbf{Y}' \in \mathcal{R}}(\text{IWF}_{\mathbf{Y}'})}{\max(\text{IWF}_{\mathcal{R}}) - \min_{\mathbf{Y}' \in \mathcal{R}}(\text{IWF}_{\mathbf{Y}'})}, \tag{14}$$

  where $\max(\text{IWF}_{\mathcal{R}})$ and $\min(\text{IWF}_{\mathcal{R}})$ denote the maximal and minimum IWF value in $\mathcal{R}$, respectively.

### 3.6 Controlled Response Generation

In the response generation step, our model provides the intervention mechanism to actively control the generation of the response for a new input utterance. Specifically, we propose to flexibly vary the state variable $s$ to generate different responses with respect to sentiment or topic, or employ the ground-truth state variable to fit the conversation data.

*3.6.1 Sentiment Controlled Response Generation.* Given an input utterance, there are multiple emotion categories that are suitable for its response. Hence, we can employ the learned DFGen model with the discrete-based intervention mechanism to generate emotional responses by varying the discrete state variable $s$, such as "happiness," "sadness," and "anger." In this way, our work makes it possible to generate a response to an arbitrary emotion by conditioning the generation on the responder's current mood. Moreover, we provide our DFGen model with the ground-truth emotion category of the response for prediction and comparison.

*3.6.2 Topic Specificity Controlled Response Generation.* Given a new input utterance, we can employ the learned DFGen model with the continuous-based intervention mechanism to generate responses at different topic specificity levels by varying the cause variable $s$. In this way, we can simulate human conversations where one can actively control the response specificity, depending on his/her current knowledge state. When we apply our model to a chatbot, there might be different ways to use the continuous state variable for conversation in practice. If we want the agent to always generate informative responses, we can set $s$ to 1 or some value close to 1. If we want the agent to be more dynamic, we can sample $s$ within the range [0,1] to enrich the styles. Since the inconsistency of response distributions between the training response collection and testing response collection, we may not be able to obtain the ground-truth continuous state variable for the response generation process.

We may further employ some reinforcement learning technique to estimate a model of which state the responder is likely to express, and then adjust the state variable depending on the responder' feedbacks. This would make the agent even more vivid, and we leave this as our future work.

## 4 EXPERIMENT

In this section, we conduct experiments to demonstrate the effectiveness of our proposed model on benchmark collections.

### 4.1 Dataset Description

To evaluate the performance of our model, we conducted experiments on two conversation datasets: (1) Emotional Conversation (EC) Dataset[7] for the response generation with respect to sentiment, (2) Short Text Conversation (STC) dataset[8] for the response generation with respect to topic specificity.

*4.1.1 EC Dataset.* The Emotional Conversation (EC) dataset is released in NLPCC-2017, which is used for the Emotional Conversation Generation challenge task. EC is constructed from Weibo posts and comments, and contains more than 1M Weibo post-comment pairs that could be used to simulate the utterance-response pairs in conversation. The dataset also includes emotion labels of each utterance and response. These sentences are rated on a scale of $0-5$, i.e., 0: *Other*, 1: *Like*, 2: *Sadness*, 3: *Disgust*, 4: *Anger*, 5: *Happiness*. Since the test dataset only includes utterances without the reference responses, we use the provided complete training dataset for our experiments. We employ the Jieba Chinese word segmenter[9] to tokenize the utterances and responses into sequences of Chinese words. The statistics of the EC Dataset are shown in Table 1. We randomly selected two subsets as the development and test dataset, each containing 10K pairs. The remaining pairs are used for training.

[7] http://tcci.ccf.org.cn/conference/2017/cfpt.php.
[8] http://ntcirstc.noahlab.com.hk/STC2/stc-cn.htm.
[9] https://pypi.python.org/pypi/jieba.

Table 1. Emotional Conversation (EC) Data Statistics: #w Denotes the Number of Chinese Words

| | | |
|---|---|---|
| Utterance-response pairs | | 1,119,207 |
| Utterance vocabulary #w | | 78,254 |
| Response vocabulary #w | | 78,175 |
| Utterance max #w | | 33 |
| Utterance avg #w | | 8 |
| Response max #w | | 33 |
| Response avg #w | | 10 |
| Response | Like | 200,001 |
| | Sadness | 181,252 |
| | Disgust | 200,001 |
| | Anger | 139,883 |
| | Happiness | 200,001 |
| | Other | 198,069 |

Table 2. Short Text Conversation (STC) Data Statistics: #w Denotes the Number of Chinese Words

| | |
|---|---|
| Utterance-response pairs | 3,788,571 |
| Utterance vocabulary #w | 120,930 |
| Response vocabulary #w | 524,791 |
| Utterance max #w | 38 |
| Utterance avg #w | 13 |
| Response max #w | 74 |
| Response avg #w | 10 |

*4.1.2 STC Dataset.* The public Short Text Conversation (STC) dataset is released in NTCIR-13. STC maintains a large repository of post-comment pairs from the Sina Weibo, which is one of the popular Chinese social media sites. STC dataset contains roughly 3.8M post-comment pairs. We also employ the Jieba Chinese word segmenter to tokenize the utterances and responses into sequences of Chinese words, and the detailed dataset statistics are shown in Table 2. We randomly selected two subsets as the development and test dataset, each containing 10K pairs. The left pairs are used for training.

### 4.2 Baselines Methods

Since we conduct our dual-factor generation on the conversation task, we compare our proposed DFGen model against several state-of-the-art conversation models for comparison:

- **Seq2Seq-att:** the standard Seq2Seq model with the attention mechanism [3];
- **MMI-bidi:** the Seq2Seq model using Maximum Mutual Information (MMI) as the objective function to reorder the generated responses [17];
- **MARM:** the Seq2Seq model with a probabilistic framework to model the latent responding mechanisms [57];

- **Seq2Seq+IDF:** an extension of Seq2Seq-att by optimizing specificity under the reinforcement learning framework, where the reward is calculated as the sentence-level IDF score of the generated response [53];
- **Enc-bef** [12] simply adds the emotion as a token (special "words" in a dictionary) before the utterance sentence as the input. Note the original work contains three methods to directly leverage the desired emotion word and, here, we select one representative method for comparison.
- **ECM:** the Seq2Seq model using emotion category embedding, internal emotion memory, and external memory to generate emotional responses [58].

We refer to our model trained for discrete-based generation and continuous-based generation as $\textbf{DFGen}_d$ and $\textbf{DFGen}_c$, respectively. We refer to the $\textbf{DFGen}_c$ model trained using NIRF and NIWF as $\textbf{DFGen}_c^{\text{NIRF}}$ and $\textbf{DFGen}_c^{\text{NIWF}}$, respectively.

### 4.3 Implementation Details

As suggested in Reference [37], we construct two separate vocabularies for utterances and responses by using 40K most frequent words on each side in the training data of STC dataset and EC dataset, respectively. All the remaining words are replaced by a special token <UNK> symbol. Last, "<eos>" is appended at the end of each sentence to indicate the end of the sentence.

We implemented our model in Tensorflow.[10] We tuned the hyper-parameters on the development set. Specifically, we use one layer of bi-directional GRU for encoder and another uni-directional GRU for decoder, with the GRU hidden unit size set as 300 in both the encoder and decoder. The dimension of semantic word embeddings in both utterances and responses is 300, while the dimension of usage word embeddings in responses is 50. We apply the Adam algorithm [16] for optimization, where the parameters of Adam are set as in Reference [16]. The variance $\sigma^2$ of the Gaussian kernel layer is set as 1, and all other trainable parameters are randomly initialized by uniform distribution within [−0.08,0.08]. The mini-batch size for the update is set as 128. We clip the gradient when its norm exceeds 5. The decoder stops when it generates the "<eos>" token.

Our model is trained on a Tesla K80 GPU card. We run the training for up to 12 epochs for STC and EC dataset, which takes approximately five days and two days, respectively. We select the model that achieves the lowest perplexity on the development dataset, and we report results on the test dataset.

### 4.4 Evaluation Methodologies

For evaluation on the continuous-based generation, we follow the existing work and employ both automatic and human evaluations:

- **Distinct-1 & Distinct-2** [17]**:** We count numbers of distinct unigrams and bigrams in the generated responses and divide the numbers by total number of generated unigrams and bigrams. Distinct metrics (both the numbers and the ratios) measure how specific and diverse the generated responses are and can be used to evaluate the specificity/diversity of the responses.
- **BLEU** [29]**:** BLEU measures the average n-gram precision on a set of reference sentences. BLEU-n is BLEU score that uses up to n-grams for counting co-occurrences. Specifically, BLEU has been proved strongly correlated with human evaluations [22].
- **Average & Extrema** [36]**:** Average and Extrema projects the generated response and the ground-truth response into two separate vectors by taking the mean over the word

[10] https://www.tensorflow.org/.

embeddings or taking the extremum of each dimension, respectively, and then computes the cosine similarity between them.

- **Human evaluation:** Three labelers with rich Weibo experience were recruited to conduct evaluation. Responses from different models are randomly mixed for labeling. Labelers refer to 300 random sampled test utterances and score the quality of the responses with the following criteria: (1) **+2**: the response is not only semantically relevant and grammatical, but also informative and interesting; (2) **+1**: the response is grammatical and can be used as a response to the utterance, but is too trivial (e.g., "I don't know"); (3) **+0**: the response is semantically irrelevant or ungrammatical (e.g., grammatical errors or UNK). Agreements to measure inter-rater consistency among three labelers are calculated with the Fleiss' kappa [8].

For evaluation on the discrete-based generation, we also use the **BLEU** and **Average & Extrema** metric to measure the average n-gram precision on the reference sentences. Furthermore, we follow the existing work in Reference [58] and employ the following metrics to evaluate the emotional responses:

- **Accuracy:** We compute the emotion accuracy as the agreement between the expected emotion category denoted by the input variable $s$ and the predicted emotion category of a generated response by the emotion classifier. Based on the analysis in Reference [58], we use the Bi-LSTM to train an emotion classifier on the NLPCC2013[11] and NLPCC2014,[12] which consists of 23,105 sentences collected from Weibo. The learned classifier is used to predict the emotion of the generated responses.
- **Human evaluation:** Three labelers were asked to score the generated response in terms of *Content* and *Emotion*. *Content* is defined the same as the human evaluation in continuous-based generation with 3-graded scores, which judges whether the generated response is appropriate to the input utterance. *Emotion* is defined as whether the emotion expression of a response agrees with the given emotion category. Following the work in Reference [58], the *Emotion* of the responses is scored with the following criteria: (1) **+1**: the emotion expression of the response agrees with the given emotion category; (2) **+0**: the emotion expression of the response disagrees with the given emotion category. We randomly sampled 300 test utterances from the test dataset. Labelers refer to random sampled test utterances and score the *Content* and *Emotion* of the generated responses.

### 4.5 Evaluation Results for the Discrete-based Generation

In this section, we will compare our proposed DFGen model against existing generative models with respect to sentiment.

*4.5.1 Model Analysis.* We first analyze our $DFGen_d$ model trained with different emotion categories. For each model, given a test utterance, we vary the discrete state variable $s$ by setting it to six different values (i.e., 0, 1, 2, 3, 4, 5) to generate the responses corresponding to six different emotion categories (i.e., other, like, sadness, disgust, anger, and happiness). We first measure the matching between the generated responses with the ground-truth response. The results are shown in Table 3, and we find that: (1) Our $DFGen_d$ model with the discrete state variable $s$ set to 0 (i.e., the other emotion) can achieve the best performances. The results indicate that the emotions of the responses in real data are various, and expressions with the "other" emotion are in the majority.

[11] http://tcci.ccf.org.cn/conference/2013/.
[12] http://tcci.ccf.org.cn/conference/2014/.

Table 3. Model Analysis of Our $\mathbf{DFGen}_d$ Model with Respect to Emotion under the Automatic Evaluation

| Models | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|
| $\mathbf{DFGen}_d^{s=0}$ | 5.97 | 2.11 | 25.22 | 13.81 |
| $\mathbf{DFGen}_d^{s=1}$ | 4.83 | 2.00 | 24.36 | 13.21 |
| $\mathbf{DFGen}_d^{s=2}$ | 4.29 | 1.27 | 23.84 | 13.07 |
| $\mathbf{DFGen}_d^{s=3}$ | 4.81 | 1.98 | 24.34 | 13.23 |
| $\mathbf{DFGen}_d^{s=4}$ | 2.34 | 0.94 | 14.52 | 6.71 |
| $\mathbf{DFGen}_d^{s=5}$ | 4.17 | 1.98 | 24.37 | 13.23 |

Table 4. Model Analysis of Our $\mathbf{DFGen}_d$ Model with Respect to Emotion under the Accuracy Evaluation (%)

| | Models | Accuracy |
|---|---|---|
| $\mathbf{DFGen}_d$ | $s = 0$ | 46.1 |
| | $s = 1$ | 69.3 |
| | $s = 2$ | 64.1 |
| | $s = 3$ | 68.9 |
| | $s = 4$ | 32.8 |
| | $s = 5$ | 70.3 |

Table 5. Model Analysis of Our $\mathbf{DFGen}_d$ Model with Respect to Emotion under the Human Evaluation

| | Content | | | | Emotion | | |
|---|---|---|---|---|---|---|---|
| | +2 | +1 | +0 | kappa | +1 | +0 | kappa |
| $\mathbf{DFGen}_d^{s=0}$ | 34.56% | 20.03% | 45.41% | 0.329 | 25.14% | 74.86% | 0.671 |
| $\mathbf{DFGen}_d^{s=1}$ | 31.61% | 15.20% | 53.19% | 0.395 | 51.81% | 48.19% | 0.711 |
| $\mathbf{DFGen}_d^{s=2}$ | 29.07% | 18.52% | 52.41% | 0.371 | 43.43% | 56.57% | 0.730 |
| $\mathbf{DFGen}_d^{s=3}$ | 31.86% | 17.66% | 50.48% | 0.381 | 49.31% | 50.69% | 0.731 |
| $\mathbf{DFGen}_d^{s=4}$ | 21.42% | 24.84% | 53.74% | 0.387 | 18.28% | 81.72% | 0.728 |
| $\mathbf{DFGen}_d^{s=5}$ | 25.39% | 31.21% | 43.40% | 0.383 | 53.32% | 46.68% | 0.725 |

(2) When the discrete state variable $s$ is set to 4 (i.e., the anger emotion), our $DFGen_d$ model performs the worst. This is mainly because the responses with anger emotion is least in the corpus.

Then, we compute the accuracy for each emotion category, which measures whether the emotion expression of the generated response agrees with the given emotion category. The results are shown in Table 4, and we can find that: (1) The overall accuracy is good, and our model can actively control the sentiment of the generated responses. (2) Our $DFGen_d$ model with the discrete state variable $s$ set to 4 (i.e., the anger emotion) performs worst, due to the fact that there are not sufficient training samples in EC dataset for this category. The anger category has only 139,883 responses in the EC dataset, much less than the other categories.

Table 5 shows the human evaluation results for our $DFGen_d$ model given different emotion categories. We can find that: (1) The relative order of $DFGen_d$ model with different $s$ values on the human evaluation is quite consistent with that on the automatic evaluation. (2) The smallest

Table 6. Comparisons between Our **DFGen**$_d$ with Respect to Sentiment and the Baselines under the Automatic Evaluation (%)

| Models | Accuracy | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|
| Seq2Seq-att | 22.42 | 9.37 | 2.34 | 31.15 | 16.76 |
| MMI-bidi | 25.91 | 9.89 | 2.56 | 32.42 | 17.13 |
| MARM | 27.37 | 6.37 | 1.65 | 31.42 | 15.66 |
| Seq2Seq+IDF | 26.52 | 8.01 | 1.98 | 31.03 | 16.75 |
| Enc-bef | 63.49 | 8.04 | 2.03 | 31.04 | 16.56 |
| ECM | 70.07 | 8.11 | 2.32 | 31.11 | 16.62 |
| **DFGen**$_d$ | **76.86** | **11.36** | **4.40** | **34.45** | **19.67** |

kappa value of Emotion evaluation is achieved by $DFGen_d^{s=0}$, which seems reasonable, since it is more difficult to reach an agreement on "other" emotion category than other certain emotion categories. (3) $DFGen_d^{s=5}$ generates the most general responses (labeled as "+0"). It is reasonable, since the sentence generated with the "happiness" emotion usually contains some general phrases, e.g., "哈哈 (haha)," and labelers prefer to consider the sentence as a common response that can reply many other utterances.

*4.5.2 Baseline Comparison.* We conduct the comparison between our $DFGen_d$ model and the baselines. Since *Seq2Seq-att*, *MMI-bidi*, *MARM,* and *Seq2Seq+IDF* are unable to generate responses of different emotions for the same utterance, we leverage the generated response ranked highest for comparison. For our $DFGen_d$ model and the baseline *ECM* and *Enc-bef*, we give the ground-truth emotion category of the response for prediction. The results are shown in Table 6. As can be seen, we can find that: (1) The relative order of different models on the EC dataset is quite consistent with that on the previous STC dataset. These results again demonstrated the effectiveness of our model on the emotion-based generation. All the improvements over the baseline models are statistically significant (p-value $< 0.01$). (2) Without loss of the generality, the most straightforward approach to include additional emotion information in the generation model is to append the emotion category to the input utterance with a separator. As a result, *Enc-bef* can achieve better results than traditional models without the consideration of emotion influence. By further modeling the emotion factor using an internal memory module, *ECM* can significantly outperform the *Enc-bef* (p-value $< 0.01$). (3) Our $DFGen_d$ model achieves the best performance in terms of all automatic metrics, which shows the effectiveness of explicitly incorporating the emotion category of the response. (4) The improvements of our model over the *ECM* baseline show that modeling responder state is more suitable for the conversation task than introducing embedding emotion category.

Table 7 shows the human evaluation results of our $DFGen_d$ model and baselines. From the results on the content measure, we can see that: (1) $DFGen_d$ achieves comparable results, which are significantly better than all the baseline methods. Sign tests demonstrate the improvements of $DFGen_d$ to the baseline models are statistically significant (p-value $<0.01$). (2) To further analyze the human evaluation, we conduct the averaged scores given by different models, i.e., sum of the percentage multiplying the human score. As compared with the best-performing baseline *ECM*, the relative improvement of $DFGen_d$ (1.1581) over *ECM* (0.9933) is about 16.60% in terms of the averaged score. (3) The percentage of the most informative and interesting responses (labeled as "2") of $DFGen_d$ is 52.02%, which is also significantly higher than that of the best baseline *ECM*, i.e., 43.82% (p-value $<0.01$). The improvement shows the ability of $DFGen_d$ to control the balance of sentiment and

Table 7. Results on the Human Evaluation of Our **DFGen**$_d$ with Respect to Sentiment

| | Content | | | | Emotion | | |
|---|---|---|---|---|---|---|---|
| | +2 | +1 | +0 | kappa | +1 | +0 | kappa |
| Seq2Seq-att | 24.94% | 20.62% | 54.44% | 0.452 | 35.24% | 64.76% | 0.727 |
| MMI-bidi | 31.48% | 16.01% | 52.51% | 0.427 | 38.26% | 61.74% | 0.731 |
| MARM | 28.22% | 15.55% | 56.23% | 0.447 | 41.52% | 58.48% | 0.709 |
| Seq2Seq+IDF | 25.40% | 32.56% | 42.04% | 0.451 | 38.97% | 61.03% | 0.711 |
| Enc-bef | 35.66% | 14.52% | 49.82% | 0.452 | 42.18% | 57.82% | 0.729 |
| ECM | 43.82% | 11.69% | 44.49% | 0.441 | 47.33% | 52.67% | 0.739 |
| **DFGen**$_d$ | 52.02% | 11.77% | 36.21% | 0.456 | 51.61% | 48.39% | 0.740 |

appropriate content for the generated response. (4) The kappa value of our model for the content is larger than 0.4, considered as "moderate agreement" regarding quality of responses.

From the results on the emotion measure, we can observe that: (1) *DFGen*$_d$ outperforms the other methods. Sign tests demonstrate the improvements of *DFGen*$_d$ to the baseline models are statistically significant (p-value $< 0.01$). (2) The performance in correct emotion (labeled as "1") is improved from 47.33% of the best baseline *ECM* to 51.61% of *DFGen*$_d$, indicating our model can generate more explicit expressions of emotion compared with using emotion embeddings. (3) The kappa value of our model for the emotion is larger than 0.7, considered as "substantial agreement" regarding emotion of responses. It seems reasonable, since it is easy to reach an agreement on the emotion categories.

*4.5.3 Case Study.* Here, we show some generated responses from our *DFGen*$_d$ for demonstration. First, Table 9 gives four utterances and the top generated responses from the *Seq2Seq-att*, *MMI-bidi*, *MARM,* and *Seq2Seq+IDF* baselines and our model. We can see that: (1) Given an utterance, there are multiple emotion categories that are suitable for its corresponding response in conversation. However, *Seq2Seq-att*, *MMI-bidi*, and *Seq2Seq+IDF* generate a response with a random emotion. Moreover, the generated responses can not well describe the responder state. (2) Although the *MARM* baseline can generate multiple responses from different latent mechanisms, it is difficult to distinguish the responses with respect to the emotion categories. Here, we show the top responses with highest probability score from generated response candidates for each utterance, and they are all quite general and short. (3) Our *DFGen*$_d$ model can generate emotional responses conditioned on each emotion category. Take the case 1 "长沙有这么有意境的地方? ! Is there such an interesting place in Changsha?" for example; the emotion of the ground-truth response is "other," while the generated responses are related to different emotions. The response "I still haven't gone there and I am sad" with $s = 2$ focuses on the "sadness" emotion, while another response, "Oh my god. I am so tired" with $s = 4$ focuses on the "anger" emotion. In case 3 with $s$ as 1 and 4, we can find words such as "浪漫(romantic)" and "歧视(discrimination)," which explicitly express *like* and *anger* emotions, respectively, by applying the discrete-based intervention mechanism for deciding the words. All these responses are appropriate not only in content but also in emotion to the utterance, indicating the effectiveness of our model for considering the cause of the responder state and the input utterance for the response simultaneously.

*4.5.4 Analysis on Usage Representations.* We also conduct some analysis to understand the usage representations of words denoting its usage preference in different emotion expressions. We randomly sampled four words from our *DFGen*$_d$; we show the top-5 similar words based on

Table 8. Target Words and Their Top-5 Similar Words under Usage and Semantic Representations, Respectively, in Our **DFGen**$_d$ with Respect to Sentiment

| 爱情(love) | | 蛋糕(cake) | |
|---|---|---|---|
| Usage | Semantic | Usage | Semantic |
| 花(flower) | 人生(life) | 团圆(reunion) | 烧烤(barbecue) |
| 等待(wait) | 感情(feeling) | 生日(birthday) | 寿司(sushi) |
| 坚强(strong) | 香水(perfume) | 圣诞(Christmas) | 巧克力(chocolate) |
| 天涯(skyline) | 亲情(family) | 礼物(gift) | 饺子(dumplings) |
| 放弃(abandon) | 女人(woman) | 幸福(gift) | 饼干(cookie) |
| 傻子(fool) | | 七夕(Chinese Valentine's Day) | |
| Usage | Semantic | Usage | Semantic |
| 恶意(malice) | 傻瓜(fool) | 礼物(gift) | 国庆(National Day) |
| 欺压(oppress) | 疯子(madman) | 情侣(lovers) | 中秋(Mid-Autumn Festival) |
| 猪(pig) | 笨蛋(idiot) | 幸福(happiness) | 端午(Dragon Boat Festival) |
| 无赖(rogue) | 坏人(bad person) | 怀抱(embrace) | 春节(Spring Festival) |
| 胡说八道(nonsense) | 混蛋(wretch) | 倍感(feel) | 节日(festival) |

cosine similarity under both representations in Table 8. We can see that the nearest neighbors of a same word are quite different under two representations. Take the word "七夕(Chinese Valentine's Day)" as an example; the neighbors based on usage representations are "礼物(gift)," "情侣(lovers)," and "怀抱(embrace)," which often happen at the expression with the emotion category "happiness." Meanwhile, the neighbors based on semantic representations are "国庆(National Day)," "中秋(Mid-Autumn Festival)," and "端午(Dragon Boat Festival)," which are semantically related about the specific festivals.

### 4.6 Evaluation Results for the Continuous-based Generation

We first compare our proposed DFGen model against existing generative models with respect to topic specificity.

*4.6.1 Model Analysis.* We first analyze our models trained with different distant supervision information. For each model, given a test utterance, we vary the continuous state variable $s$ by setting it to five different values (i.e., 0, 0.2, 0.5, 0.8, 1) to check whether the learned model can actually achieve different topic specificity levels. As shown in Table 10, we find that: (1) The $DFGen_c$ model trained with NIRF cannot work well. The test performances are almost the same with different $s$ value. This is surprising, since the NIRF definition seems to be directly corresponding to the topic preference of the responder. By conducting further analysis, we find that even though the conversation dataset is large, it is still limited, and a general response could appear very few times in this corpus. In other words, the inverse frequency of a response is very weakly correlated with the response topic specificity. (2) The $DFGen_c$ model trained with NIWF can achieve our purpose. By varying the continuous state variable $s$ from 0 to 1, the generated responses turn from general to specific as measured by the distinct metrics. The results indicate that the max inverse word frequency in a response is a good distant label for the response topic specificity. (3) When we compare the generated responses against ground-truth data, we find the $DFGen_c^{\text{NIWF}}$ model with the continuous state variable $s$ set to 0.5 can achieve the best performances. The results indicate that there are diverse responses in real data in terms of topic specificity, and it is necessary to take a balanced setting if we want to fit the ground-truth data.

Table 9. Examples of Response Generation from the EC Test Data

| |
|---|
| **Utterance**: 长沙有这么有意境的地方? ! Is there such an interesting place in Changsha?! |
| **Ground-truth Response**: 石燕湖。小林同学。 Shiyan Lake. Lin. |
| **Seq2Seq-att**: 我很久没有去啦。 It looks good, very good! |
| **MMI-bidi**: 真的吗。? Really? |
| **MARM**: 真好。Good. |
| **Seq2Seq+IDF**: 不错。Good. |
| **s=1**: 好美，好美，我喜欢。 So beautiful. I like it. |
| **s=2**: 还是没去，伤心啊。 I still haven't gone there and I am sad. |
| **s=3**: 无期望就不会有失望啦。 There will be no disappointment without expectation. |
| **s=4**: 我勒个去，好累。 Oh my god. I am so tired. |
| **s=5**: 感觉很好，我很开心。 I feel good and I am happy. |
| **Utterance**: 离出嫁还早着呢，离一个贤惠的老婆还有一定的距离同志。 It is still early to leave the marriage, and there is still a certain distance from a virtuous wife. |
| **Ground-truth Response**: 好的，看来我还得努力啊，不过你的要求也很高啊。 OK. It seems I have to work hard, but your requirements are also high. |
| **Seq2Seq-att**: 好的。Yeah. |
| **MMI-bidi**: 就是嘛就是嘛，我会努力的。Yes, Yes. I will try my best. |
| **MARM**: 我会努力的。I will try my best. |
| **Seq2Seq+IDF**: 真不错。Really good. |
| **s=1**: 谢谢支持，祝你好运！ Thanks for your support. Good luck! |
| **s=2**: 孤单寂寞感到冷。 I feel lonely and cold. |
| **s=3**: 那可说不准哦。 That can't be said. |
| **s=4**: 我心里难受啊。I feel bad in my heart. |
| **s=5**: 会实现的，哈哈。 It will be realized. Haha. |
| **Utterance**: 爱情的力量是不能用地域来衡量的。 The power of love cannot be measured by the geography. |
| **Ground-truth Response**: 晚上搞基吧。反正不分地域。Let's do it at night, regardless of the region, anyway. |
| **Seq2Seq-att**: 相信爱情。Believe love. |
| **MMI-bidi**: 爱情的力量是真的吗? Is the power of love true? |
| **MARM**: 这是真的吗? Is this true? |
| **Seq2Seq+IDF**: 爱情啊。Love. |
| **s=1**: 真浪漫，那就是向往中的爱情。It is really romantic. That is the desirable love. |
| **s=2**: 我再也不相信爱情了。I don't believe in love anymore. |
| **s=3**: 这是真的还是假的。Is this true? |
| **s=4**: 赤裸裸的爱情歧视。It is naked love discrimination |
| **s=5**: 哈哈，我和我的小伙伴都惊呆了。Haha, I and my friends are stunned. |
| **Utterance**: 十分同情你的遭遇，我们也有类似的困境。We have a similar dilemma with you, and I am very sympathetic to your experience. |
| **Ground-truth Response**: 有人胆敢打劫? 反了天了! Anyone dare to rob? Oh my God! |
| **Seq2Seq-att**: 谢谢啦。Thanks. |
| **MMI-bidi**: 谢谢你的关心。Thanks for your care. |
| **MARM**: 哎。Well. |
| **Seq2Seq+IDF**: 一起加油吧。 Let's do it together. |

(Continued)

Table 9. Continued

**s=1**: 困境可以锻炼自己。The dilemma can exercise yourself.
**s=2**: 可怜已经无法形容我的困境了。Poorness has been unable to describe my dilemma.
**s=3**: 不公平啊。It is unfair.
**s=4**: 真的很难受啊。I feel so bad.
**s=5**: 身处困境需要乐观的心态。You need an optimistic attitude when you are in a difficult dilemma.

$s$ = 1, 2, 3, 4, 5 are the outputs of our $\mathbf{DFGen}_d$ with the like, sadness, disgust, anger, and happiness emotions, respectively.

Table 10. Model Analysis of Our $\mathbf{DFGen}_c$ with Respect to Topic Specificity under the Automatic Evaluation

| | Models | Distinct-1 | Distinct-2 | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|---|---|
| $\mathbf{DFGen}_c^{\text{NIRF}}$ | $s$ = 1 | 5,258/0.064 | 16,195/0.269 | 15.109 | 7.023 | 0.578 | 0.380 |
| | $s$ = 0.8 | 5,337/0.065 | 16,105/0.271 | 15.112 | 7.003 | 0.578 | 0.381 |
| | $s$ = 0.5 | 5,318/0.065 | 16,183/0.269 | 15.054 | 7.001 | 0.578 | 0.380 |
| | $s$ = 0.2 | 5,323/0.065 | 16,087/0.270 | 15.168 | 7.032 | 0.580 | 0.380 |
| | $s$ = 0 | 5,397/0.066 | 16,319/0.271 | 15.093 | 7.011 | 0.577 | 0.380 |
| $\mathbf{DFGen}_c^{\text{NIWF}}$ | $s$ = 1 | **11,588/0.116** | **27,144/0.347** | 12.392 | 5.869 | 0.554 | 0.353 |
| | $s$ = 0.8 | 6,006/0.051 | 17,843/0.257 | 11.492 | 5.703 | 0.553 | 0.350 |
| | $s$ = 0.5 | 2,835/0.050 | 9,537/0.235 | **16.122** | **7.674** | **0.609** | **0.399** |
| | $s$ = 0.2 | 1,534/0.048 | 5,117/0.218 | 8.313 | 4.058 | 0.542 | 0.335 |
| | $s$ = 0 | 1,038/0.046 | 3,154/0.211 | 4.417 | 3.283 | 0.549 | 0.334 |

Table 11. Comparisons between Our $\mathbf{DFGen}_c$ with Respect to Topic Specificity and the Baselines under the Automatic Evaluation

| Models | Distinct-1 | Distinct-2 | BLEU-1 | BLEU-2 | Average | Extrema |
|---|---|---|---|---|---|---|
| Seq2Seq-att | 5,048/0.060 | 15,976/0.168 | 15.062 | 6.964 | 0.575 | 0.376 |
| MMI-bidi | 5,074/0.082 | 12,162/0.287 | 15.772 | 7.215 | 0.586 | 0.381 |
| MARM | 2,566/0.096 | 3,294/0.312 | 7.321 | 3.774 | 0.512 | 0.336 |
| Seq2Seq+IDF | 4,722/0.052 | 15,384/0.229 | 14.423 | 6.743 | 0.572 | 0.369 |
| $\mathbf{DFGen}_c^{\text{NIWF},s=1}$ | **11,588/0.116** | **27,144/0.347** | 12.392 | 5.869 | 0.554 | 0.353 |
| $\mathbf{DFGen}_c^{\text{NIWF},s=0.5}$ | 2,835/0.050 | 9,537/0.235 | **16.122** | **7.674** | **0.609** | **0.399** |

*4.6.2 Baseline Comparison.* The performance comparisons between our *DFGen*$_c$ model and the baselines are shown in Table 11. We have the following observations: (1) By using MMI as the objective, *MMI-bidi* can improve the topic specificity (in terms of distinct ratios) over the traditional *Seq2Seq-att* model. (2) By modeling different responding mechanisms as latent embeddings, *MARM* can achieve the best distinct ratios among the baseline methods, but the worst in terms of the distinct numbers. The results indicate that *MARM* tends to generate specific but very short responses. Meanwhile, its low BLEU scores also show that the responses generated by *MARM* deviate from the ground truth significantly (p-value $< 0.01$). The results again demonstrate that depending only on the input utterance is not suitable for the response generation. (3) By using the IDF information of the response as the reward to train the Seq2Seq model, the *Seq2Seq+IDF* does not show much advantages, but only achieves comparable results as MMI-bidi. (4) By setting the continuous state

Table 12. Results on the Human Evaluation of Our **DFGen**$_c$ with Respect to Topic Specificity

| | +2 | +1 | +0 | kappa |
|---|---|---|---|---|
| Seq2Seq-att | 29.32% | 25.27% | 45.41% | 0.448 |
| MMI-bidi | 30.40% | 24.85% | 44.75% | 0.471 |
| MARM | 20.11% | 27.96% | 51.93% | 0.404 |
| Seq2Seq+IDF | 28.81% | 23.87% | 47.33% | 0.418 |
| **DFGen**$_c^{\text{NIWF}, s=1}$ | 42.47% | 14.29% | 43.24% | 0.507 |
| **DFGen**$_c^{\text{NIWF}, s=0.5}$ | 20.62% | 40.16% | 39.22% | 0.451 |
| **DFGen**$_c^{\text{NIWF}, s=0}$ | 14.34% | 46.38% | 39.28% | 0.526 |

variable $s$ to 1, our $DFGen_c^{\text{NIWF}}$ model can achieve the best specificity performance as evaluated by the distinct metrics. By setting the state variable $s$ to 0.5, our $DFGen_c^{\text{NIWF}}$ model can best fit the ground-truth data as evaluated by the BLEU scores, Average and Extrema. All the improvements over the baseline models are statistically significant (p-value $< 0.01$). These results demonstrate the effectiveness as well as the flexibility of our dual-factor generation model.

Table 12 shows the human evaluation results of our **DFGen**$_c$ model and baselines. We can observe that: (1) $DFGen_c^{\text{NIWF}, s=1}$ generates the most informative responses and interesting (labeled as "+2") and the least general responses than all the baseline models. Meanwhile, $DFGen_c^{\text{NIWF}, s=0}$ generates the most general responses (labeled as "+1"). (2) *MARM* generates the most bad responses (labeled as "+0"), which indicates the drawbacks of the unknown latent responding mechanisms. (3) By increasing the state variable $s$ from 0 to 1 in our **DFGen**$_c$ model, the percentage of the irrelevant and ungrammatical responses (labeled as "+0") is increasing. When we look at those cases (label "0" given $s = 1$), we find that some input utterances may not need specific responses, while the model is forced to generate the most specific response so it is likely to generate <UNK> words in this case. This means that it is not always the best to assume that the topic preference of the responder is very specific denoted as $s = 1$, and that is why $s = 0.5$ achieves the best in terms of BLEU, average, and extrema metrics in automatic evaluation. (4) The kappa values of our models are all larger than 0.4, considered as "moderate agreement" regarding quality of responses. The largest kappa value is achieved by $DFGen_c^{\text{NIWF}, s=0}$, which seems reasonable, since it is easy to reach an agreement on general responses. Sign tests demonstrate the improvements of $DFGen_c^{\text{NIWF}, s=1}$ to the baseline models are statistically significant (p-value $< 0.01$). All the human judgment results again demonstrate the effectiveness of our intervention mechanism.

*4.6.3 Case Study.* To better understand how different models perform, we conduct some case studies. We randomly sample four utterances from the test dataset and show the responses generated by different models. As shown in Table 13, we can find that: (1) The responses generated by the four baselines are often quite general and short, which may quickly lead to an end of the conversation. (2) $DFGen_c^{\text{NIWF}}$ with large continuous state variable values (i.e., $s > 0.5$) can generate very long and specific responses. In these responses, we can find many informative words. For example, in case 2 with $s$ as 1 and 0.8, we can find words such as "眼妆 (eye make-up)," "气质 (temperament)," and "雪亮 (bright)," which are quite specific and strongly related to the conversation topic of "beauty." Specifically, the ground-truth response is relatively general compared with the responses generated by our model. In case 1 with $s$ as 1, we can find the word "牌子 (brand)" is quite related to the topic "帐篷 (tent)." (3) When we decrease the continuous state variable value, the generated responses become more and more general and shorter from our $DFGen_c^{\text{NIWF}}$ model.

Table 13. Examples of Response Generation from the STC Test Data

**Utterance**: 三室一厅的帐篷，你值得拥有！ A tent with three rooms and one hall, you're worth it!
**Ground-truth Response**: 我想看看里面的样子。I want to see what it looks like inside.
**Seq2Seq-att**: 这个不错，很好! It looks good, very good!
**MMI-bidi**: 这个不错，我喜欢。It looks good, I like it.
**MARM**: 不错。Good.
**Seq2Seq+IDF**: 帐篷不错。The tent is good.
**s=1**: 我想知道这帐篷是什么牌子的？ I want to know what the brand of this tent is?
**s=0.8**: 我想知道这是在哪买的？ I want to know where to buy it?
**s=0.5**: 我也想买一个,可惜买不起！ I want one too, but it's a pity that I can't afford it.
**s=0.2**: 不错不错！Good, good.
**s=0**: 真的吗？ Really?

**Utterance**: 哎哟! 被发现了! 汤唯有必要那么美吗? Oh! Got caught! Why is Tang Wei so beautiful?
**Ground-truth Response**: 我超喜欢她！ 有人可以帮我告诉她吗？ I really like her! Can someone help me tell her?
**Seq2Seq-att**: 我超喜欢她！ I really like her!
**MMI-bidi**: 好美。我超喜欢她！ So beautiful. I really like her!
**MARM**: 我不喜欢汤唯。I don't like Tang Wei.
**Seq2Seq+IDF**: 我超喜欢她的，我喜欢。I really like her, I like.
**s=1**: 我超喜欢她的眼妆，她的气质很美！ I really like her eye make-up, and her temperament is so beautiful!
**s=0.8**: 我超喜欢她的，她的眼睛很雪亮！ I really like her, and her eyes are so bright!
**s=0.5**: 我觉得她很美，很喜欢她！ I think she is very beautiful, and I really like her!
**s=0.2**: 我觉得她很美！ I think she is very beautiful.
**s=0**: 是的。Yes.

**Utterance**: 将来也要养一只这么乖的的萨摩，让它陪着孩子一起长大。 I will keep such a sweet samoyed, and let it accompany my children to grow up.
**Ground-truth Response**: 好乖的萨摩，也想养狗狗了。What a good samoyed. I want to have a dog too.
**Seq2Seq-att**: 这不是萨摩吗？ Is it a samoyed?
**MMI-bidi**: 好可爱的萨摩。What a lovely samoyed.
**MARM**: 好可爱啊。 So lovely.
**Seq2Seq+IDF**: 好可爱的狗狗。What a lovely dog.
**s=1**: 萨摩耶好可爱，真想养一只! The samoyed is so lovely, and I really want to keep one!
**s=0.8**: 萨摩好可爱，好想掐掐。The samoyed is so lovely, and I really want to pinch it.
**s=0.5**: 好可爱的狗狗，好可爱的狗狗。What a lovely dog, what a lovely dog.
**s=0.2**: 好可爱！ 好可爱！ So lovely, so lovely!
**s=0**: 好可爱！ So lovely!

**Utterance**: 其实。这张照片整幅都是画出来的。很有欺骗性啊！ Actually, this picture is a drawing. It is very deceptive!
**Ground-truth Response**: 画出来的! 太牛了吧。It is a drawing! Too great.
**Seq2Seq-att**: 真的吗？ Really?
**MMI-bidi**: 这是画出来的吗？ Is it a drawing?
**MARM**: 这不是画的吗？ Is it not a drawing?
**Seq2Seq+IDF**: 这是画的吗？ Is it a drawing?

(Continued)

Table 13. Continued

| |
|---|
| **s=1**: 这是个体力活啊！ It is a physical labor! |
| **s=0.8**: 这是画出来的吗? 好厉害！ Is it a drawing? So amazing! |
| **s=0.5**: 我也想知道你的照片是什么？ I also want to know what your picture is? |
| **s=0.2**: 这是画的吗？ Is it a drawing? |
| **s=0**: 不错。Good. |

$s$ = 1, 0.8, 0.5, 0.2, 0 are the outputs of our **DFGen**$_c^{\text{NIWF}}$ with different $s$ values.


(a) usage (b) semantic

Fig. 5. t-SNE embeddings of usage and semantic vectors in our **DFGen**$_c^{\text{NIWF}}$ with respect to topic preference.

*4.6.4 Analysis on Usage Representations.* We also conduct some analysis to understand the usage representations of words, denoting its usage preference under different topic specificity levels. We randomly sample 500 words from our *DFGen*$_c^{\text{NIWF}}$ and apply t-SNE [26] to visualize both usage and semantic embeddings. As shown in Figure 5, we can see that the two distributions are quite different. Two-tailed t-tests demonstrate the difference between these two distributions are statistically significant (p-value $< 0.01$). In the usage space, words such as "脂肪肝 (fatty liver)" and "久坐 (outsit)" lie closely, which are both specific words, and both are far from the general words like "胖 (fat)." On the contrary, in the semantic space, "脂肪肝 (fatty liver)" is close to "胖 (fat)," since they are semantically related, and both are far from the word "久坐 (outsit)."

Furthermore, given some sampled target words, we also show the top-5 similar words based on cosine similarity under both representations in Table 14. Again, we can see that the nearest neighbors of a same word are quite different under two representations. Neighbors based on semantic representations are semantically related, while neighbors based on usage representations are not so related but with similar topic specificity levels. Namely, the usage representation can well capture the the usage preference of the responder with respect to his/her knowledge state.

### 4.7 Comparison between Discrete-based and Continuous-based Generation

To further understand how our model works, we compare the responses generated by DFGen$_c^{\text{NIWF}}$ and DFGen$_d$, respectively, on the EC dataset. Here, we choose the distinct and accuracy evaluation metrics to analyze the differences between continuous-based and discrete-based generation. As shown in Table 15, we can find that: (1) For the distinct metric, the relative order of *DFGen*$_c^{\text{NIWF}}$ model with different topic specificity levels on the EC dataset is quite consistent with that on the STC dataset as shown in Table 10. The results again demonstrate the effectiveness of our continuous-based generation mechanism. The *DFGen*$_d$ model with different sentiments on the EC dataset perform almost the same, which indicates that the emotion category is very weakly correlated with the topic specificity. Meanwhile, the overall performance of *DFGen*$_d$ model with

Table 14. Target Words and Their Top-5 Similar Words under Usage and Semantic Representations, Respectively, in Our $\mathbf{DFGen}_c^{\text{NIWF}}$ with Respect to Topic Specificity

| 爸爸(dad) | | 水果(fruits) | |
|---|---|---|---|
| Usage | Semantic | Usage | Semantic |
| 更好(better) | 妈妈(mother) | 尝试(attempt) | 蔬菜(vegetables) |
| 睡觉(sleep) | 哥哥(brother) | 诱惑(tempt) | 牛奶(milk) |
| 快乐(happy) | 老公(husband) | 表现(express) | 西瓜(watermelon) |
| 无聊(boring) | 爷爷(grandfather) | 拥有(own) | 米饭(rice) |
| 电影(movie) | 姑娘(girl) | 梦想(dream) | 巧克力(chocolate) |
| 脂肪肝(fatty liver) | | 单反相机(DSLR) | |
| Usage | Semantic | Usage | Semantic |
| 坐久(outsit) | 胖(fat) | 亚洲杯(Asian Cup) | 照相机(camera) |
| 素食主义(vegetarian) | 减肥(diet) | 读取(read) | 摄影(photography) |
| 散步(walk) | 高血压(hypertension) | 半球(hemispherical) | 镜头(shot) |
| 因果关系(causality) | 亚健康(sub-health) | 防辐射(anti-radiation) | 影楼(studio) |
| 哑铃(dumbbell) | 呕吐(emesis) | 无人机(UAV) | 写真(image) |
| 番茄酱(ketchup) | | 男朋友(boyfriend) | |
| Usage | Semantic | Usage | Semantic |
| 放错 (misplaced) | 柿子 (persimmon) | 相信 (believe) | 女朋友 (girlfriend) |
| 苦涩 (bitter) | 双皮奶 (milk) | 棒 (good) | 女友 (girlfriend) |
| 汉堡包 (hamburger) | 小龙虾 (crawfish) | 回忆(memory) | 前女友 (ex-girlfriend) |
| 面食(pasta) | 菠萝蜜 (jackfruit) | 工作 (work) | 老公 (husband) |
| 薯条(chips) | 白萝卜 (mooli) | 强大 (powerful) | 父亲 (father) |

Table 15. Comparisons between $\mathbf{DFGen}_c$ and $\mathbf{DFGen}_d$ on the EC Dataset under the Automatic Evaluation

| | Models | Distinct-1 | Distinct-2 | Accuracy (%) |
|---|---|---|---|---|
| $\mathbf{DFGen}_c^{\text{NIWF}}$ | $s = 1$ | 0.098 | 0.323 | 34.3 |
| | $s = 0.8$ | 0.071 | 0.265 | 31.4 |
| | $s = 0.5$ | 0.052 | 0.199 | 37.4 |
| | $s = 0.2$ | 0.048 | 0.184 | 36.6 |
| | $s = 0$ | 0.041 | 0.181 | 38.0 |
| $\mathbf{DFGen}_d$ | $s = 1$ | 0.064 | 0.231 | - |
| | $s = 2$ | 0.061 | 0.225 | - |
| | $s = 3$ | 0.066 | 0.236 | - |
| | $s = 4$ | 0.068 | 0.237 | - |
| | $s = 5$ | 0.069 | 0.237 | - |
| | $s$ = ground-truth | 0.081 | 0.284 | 76.86 |

Specifically, $s$ denotes the topic specificity level and sentiment in $\mathbf{DFGen}_c$ and $\mathbf{DFGen}_d$, respectively.

different sentiments is close to $DFGen_c^{\text{NIWF}}$ model with $s = 0.8$, while $DFGen_c^{\text{NIWF}}$ model with ground-truth sentiments performs better. The reason of good specificity performance given by $DFGen_d$ model might be that emotional expressions usually contain specific and diverse words or phrases. (2) For the accuracy metric, we compare the $DFGen_c^{\text{NIWF}}$ model with different topic

specificity levels and $DFGen_d$ model with the ground-truth emotion. Specifically, the emotion accuracy is measured between the predicted emotion category of generated responses by the emotion classifier and the ground-truth emotion category. Note, we do not show the accuracy results of $DFGen_d$ model with different given sentiments (i.e., $s = 1, 2, 3, 4, 5$), since it is not reasonable to compare the predicted emotion category with the ground truth. $DFGen_c^{\text{NIWF}}$ performs worse than $DFGen_d$ model, due to the fact that it mainly focuses on the topic specificity instead of emotion. Specifically, $DFGen_c^{\text{NIWF}}$ with $s = 0$ performs better than that with other $s$ values. The reason might be that the general responses usually contain indicative words (e.g., "Good," "Haha," and "Thanks") of the "like" and "happiness" emotion, which contributes to the improvement of the overall emotion classification accuracy. (3) By testing $DFGen_c$ and $DFGen_d$ on the same dataset, we can conclude that it is necessary to employ the specific intervention mechanism to control the generation of the response with respect to a given responder state.

## 5 CONCLUSION

In this article, we formulated the conversation task as a dual-factor generation problem, where the objective factor of the input utterance and the subjective factor of the responder affect the response simultaneously. We proposed a novel Dual-Factor Generation model, which can well solve the blandness and intervention issue in existing generation-based models. Specifically, we introduced an explicit state variable into the Seq2Seq model, which interacts with the usage representation of words to generate responses affected by the responder's emotion state or topic preference state. Empirical results showed that our model is capable of controlling the generation of the response as restricted by a human and significantly outperform state-of-the-art generation methods under both automatic and human evaluations.

This article focused on the problem of the single-turn conversation generation. In the future work, we plan to control the response generation in the multi-turn conversation task, which is critical in many natural language processing applications, e.g., customer services, intelligent assistant, and chatbot. Furthermore, we would like to investigate the proposed DFGen model. For example, we can attempt to learn to adjust the most appropriate state variable in a data-driven way instead of specifying a state, which can better unveil the responder's state. In addition, we can consider other important factors that may affect the response generation in practice.

## ACKNOWLEDGMENTS

We thank our anonymous reviewers for their helpful comments and valuable suggestions.

## REFERENCES

[1] Rami Al-Rfou, Marc Pickett, Javier Snaider, Yun-hsuan Sung, Brian Strope, and Ray Kurzweil. 2016. Conversational contextual cues: The case of personalization and history for response ranking. *arXiv preprint arXiv:1606.00372* (2016).

[2] Nabiha Asghar, Pascal Poupart, Jesse Hoey, Xin Jiang, and Lili Mou. 2018. Affective neural response generation. In *Proceedings of the European Conference on Information Retrieval*. Springer, 154–166.

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR'15)*.

[4] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683* (2016).

[5] Zhongxia Chen, Ruihua Song, Xing Xie, Jian-Yun Nie, Xiting Wang, Fuzheng Zhang, and Enhong Chen. 2019. Neural response generation with relevant emotions for short text conversation. In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 117–129.

[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. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[7] Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2016. Towards end-to-end reinforcement learning of dialogue agents for information access. *arXiv preprint arXiv:1609.00777* (2016).
[8] Joseph L. Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ. Psychol. Meas.* 33, 3 (1973), 613–619.
[9] Jianfeng Gao, Michel Galley, Lihong Li, et al. 2019. Neural approaches to conversational AI. *Found. Trends® Inf. Ret.* 13, 2–3 (2019), 127–298.
[10] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.
[11] Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL'14)*. 263–272.
[12] Chenyang Huang, Osmar Zaiane, Amine Trabelsi, and Nouha Dziri. 2018. Automatic dialogue generation with expressed emotions. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, Vol. 2. 49–54.
[13] Anil Jain, Karthik Nandakumar, and Arun Ross. 2005. Score normalization in multimodal biometric systems. *Pattern Recog.* 38, 12 (2005), 2270–2285.
[14] Zongcheng Ji, Zhengdong Lu, and Hang Li. 2014. An information retrieval approach to short text conversation. *arXiv preprint arXiv:1408.6988* (2014).
[15] Michael Kearns. 2000. Cobot in LambdaMOO: A social statistics agent. In *AAAI/IAAI*.
[16] Diederik Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.
[17] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL'16)*.
[18] Jiwei Li, Michel Galley, Chris Brockett, Georgios P. Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Meeting of the Association for Computational Linguistics*.
[19] Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data distillation for controlling specificity in dialogue generation. *arXiv preprint arXiv:1702.06703* (2017).
[20] Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541* (2016).
[21] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547* (2017).
[22] Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*. 150–157.
[23] Feng Liu, Qirong Mao, Liangjun Wang, Nelson Ruwa, Jianping Gou, and Yongzhao Zhan. 2018. An emotion-based responding model for natural language conversation. *World Wide Web-internet Web Inf. Syst.* 9 (2018), 1–19.
[24] Yi Luan, Chris Brockett, Bill Dolan, Jianfeng Gao, and Michel Galley. 2017. Multi-task learning for speaker-role adaptation in neural conversation models. *arXiv preprint arXiv:1710.07388* (2017).
[25] Liangchen Luo, Wenhao Huang, Qi Zeng, Zaiqing Nie, and Xu Sun. 2019. Learning personalized end-to-end goal-oriented dialog. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 6794–6801.
[26] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *J. Mach. Learn. Res.* 9, Nov. (2008), 2579–2605.
[27] Yang Min, Zhao Zhou, Zhao Wei, Xiaojun Chen, and Zigang Cao. 2017. Personalized response generation via domain adaptation. In *Proceedings of the 40th International ACM SIGIR Conference*.
[28] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 2016. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of the International Conference on Computational Linguistics (COLING'16)*.
[29] 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 Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.
[30] Diana Perez-Marin. 2011. *Conversational Agents and Natural Language Interaction: Techniques and Effective Practices: Techniques and Effective Practices*. IGI Global.
[31] Qian Qiao, Minlie Huang, and Xiaoyan Zhu. 2018. Assigning personality/identity to a chatting machine for coherent conversation generation. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI'18)*.
[32] Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei Chu. 2017. ALIME chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 498–503.

[33] 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, 583–593.

[34] Stephen E. Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *Proceedings of the SIGIR'94*. Springer, 232–241.

[35] 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 *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'16)*. 3776–3784.

[36] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[37] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*.

[38] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A conditional variational framework for dialog generation. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics*.

[39] 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* (2015).

[40] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS'14)*. 3104–3112.

[41] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? An empirical study on context-aware neural conversational models. In *Proceedings of the 55th Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 231–236.

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

[43] Marilyn A. Walker, Rebecca Passonneau, and Julie E. Boland. 2001. Quantitative and qualitative evaluation of DARPA Communicator spoken dialogue systems. In *Proceedings of the 39th Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 515–522.

[44] Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[45] Jason Williams, Antoine Raux, Deepak Ramachandran, and Alan Black. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL Conference*. 404–413.

[46] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627* (2016).

[47] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI*. 3351–3357.

[48] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to respond with deep neural networks for retrieval-based human-computer conversation system. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 55–64.

[49] Rui Yan, Yiping Song, Xiangyang Zhou, and Hua Wu. 2016. Shall I be your chat companion?: Towards an online human-computer conversation system. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 649–658.

[50] Rui Yan, Dongyan Zhao, et al. 2017. Joint learning of response ranking and next utterance suggestion in human-computer conversation system. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 685–694.

[51] Liu Yang, Junjie Hu, Minghui Qiu, Chen Qu, Jianfeng Gao, W. Bruce Croft, Xiaodong Liu, Yelong Shen, and Jingjing Liu. 2019. A hybrid retrieval-generation neural conversation model. *arXiv preprint arXiv:1904.09068* (2019).

[52] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response ranking with deep matching networks and external knowledge in information-seeking conversation systems. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 245–254.

[53] Kaisheng Yao, Baolin Peng, Geoffrey Zweig, and Kam-Fai Wong. 2016. An attentional neural conversation model with improved specificity. *arXiv preprint arXiv:1606.01292* (2016).

[54] Lantao Yu, Weinan Zhang, Jun Wang, and Yong Yu. 2017. SeqGAN: Sequence generative adversarial nets with policy gradient. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*.

[55] Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. In *Proceedings of the 56th Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1108–1117.

[56] Wei Nan Zhang, Qingfu Zhu, Yifa Wang, Yanyan Zhao, and Ting Liu. 2017. Neural personalized response generation as domain adaptation. *World Wide Web-internet Web Inf. Syst.* 4 (2017), 1–20.

[57] Ganbin Zhou, Ping Luo, Rongyu Cao, Fen Lin, Bo Chen, and Qing He. 2017. Mechanism-aware neural machine for dialogue response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'17)*. 3400–3407.

[58] Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI'18)*.

[59] 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. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 372–381.

[60] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating emotional responses at scale. In *Proceedings of the Meeting of the Association for Computational Linguistics*.

Received October 2019; revised February 2020; accepted April 2020