Front.Comput.Sci.

RESEARCH ARTICLE

Informative and diverse emotional conversation generation with mode-enriched variational recurrent pointer-generator

Weichao WANG[1], Shi FENG[1], Kaisong SONG[2], Daling WANG(✉)[1], Shifeng LI[1]

1 School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China
2 Alibaba Group, Hangzhou 310000, China

Abstract Generating emotional responses is critical to the open-domain conversation systems, because the emotion can increase user engagement and improve user empathy. Existing studies ignore the emotional trigger expressions embedded in the multi-turn conversation context, and consequently fail to generate informative and diverse emotional responses. In this paper, we propose a novel Mode-Enriched Variational Recurrent Pointer-Generator network, which conducts a pilot study of generating emotionally rich and informative responses with diverse emotional trigger expressions. Specifically, first, we incorporate the emotion factor into the Pointer-Generator network to copy informative emotional trigger entities from context, and retain the ability to generate such words from the predefined vocabulary. Second, we refine the pointing calculation of copy process with the sequential latent variables, which is used to reproduce diverse and valid emotional trigger expressions. Finally, we incorporate an emotional mode into the mode controller, so that the generated responses can be further enriched with explicit emotional words. Experiments conducted on a real-world multi-turn conversation dataset show that our proposed model outperforms the strong baseline algorithms with large margins.

Keywords emotional trigger expressions, informative and diverse response generation, Pointer-Generator network, variational recurrent neural network, mode controller

1 Introduction

Recent development in neural conversation modeling has generated significant excitement in the dialogue system community [1, 2], and has many practical applications in a wide range of fields, such as customer service, chat-bot and e-commerce. Existing neural network based studies on dialogue systems mainly focus on increasing the diversity of responses [3,4], modeling the personality of the speakers [5,6], maintaining the semantic coherence [7, 8] and developing retrieval-based response selection techniques [9, 10]. The previous dialogue systems have achieved promising results, which, however, ignore incorporation of the desired emotion to make responses more natural and emotionally acceptable.

Emotional response generation, in essence, is a natural extension of typical dialogue systems, which aims to understand the speaker's feelings and intentions, and then generate emotionally rich and empathic responses for smooth human-machine communication [11–14]. Some traditional studies on this task either controlled the response with emotion categories, such as incorporating memory module [13], conditional variational autoencoders [14] and discriminative emotion classifier [15], or explored the natural emotion incorporation, including utilizing affect-based encoding and decoding techniques [16] and affective attention mechanism [17]. Although these methods can enrich the response with desired emotion, they always ignore the relevant entities triggering the emotion, resulting in generating safe and dull emotional

Received month dd, yyyy; accepted month dd, yyyy

E-mail: wangdaling@cse.neu.edu.cn

responses.

In this paper, we study the controllable multi-turn emotional response generation task, where a response is generated according to the conversation context and the given emotion category. For the multi-turn conversation, the context refers to the conversation history of the response's previous turns between two speakers. Inspired by the psychology theory on the **emotional trigger expressions** [18, 19], in this task, we conjecture that taking the expressions into consideration can refine conversation generation and enhance user satisfaction. Specifically, the emotional trigger expressions can be defined as: the entity-related words and phrases which can provoke and cause the desired emotion of any response [20, 21]. Table 1 depicts a motivating exemplar, which presents the the suitable ground-truth responses with different emotions, and the response generated from a well-known model, called ECM [13]. We can observe that the ground-truth responses are informative and semantic coherent to the contexts, which consist of emotional words "*blame*" and "*pray*", ordinary words "*should*" and "*for*", and emotional trigger expression "*Boeing 737 Max*", "*Ethiopian Airlines*" and "*passengers and crew members*". Intuitively, these emotional trigger expressions can increase the informativeness of the generated response, and further locate the reason and origin of the desired emotion based on the complex and noisy context. By contrast, although the ECM model can generate the explicit emotion word "*sad*", the meaningless responses without emotional trigger expressions will decrease user experience greatly. Furthermore, according to the ground-truth responses of "anger" emotion, there exists various appropriate emotional trigger expressions, which should be exploited to improve the diversity of generated responses. Our motivation is to express the valid and various emotional trigger entities, simultaneously maintaining the informativeness and diversity for emotional response generation.

Compared with the single-turn conversation, the context of multi-turn conversation contains complex content with helpful emotional trigger expressions[1)]. Although the traditional methods can generate appropriate responses with given emotions in single-turn conversation, they underperform in the multi-turn conversation for the following reasons. First, the emotional trigger expressions mainly exist in conversation context and they can not be easily reproduced by the traditional vocabulary-sampling-based strategies, and the performance will worsen for those rare words and out-of-vocabulary words (OOVs). Second, the conversational utterances are all casual and non-goal-oriented, and there is no exact matching between context and response. Therefore, the diversity should be taken into account when predicting the emotional trigger entities.

Table 1 A motivating example of emotional conversation generation.

● **emotional trigger expressions** ● **emotion words** ● **ordinary words**

Context
A: Frequent air disasters have occurred in recent years. B: I heard the Ethiopian Airlines Flight 302 crashed last year. A: Really, what's the specific situation? B: Boeing 737 Max killed passengers and crew members.
Response (anger)
A_{truth}: Boeing 737 Max should be to blame for this accident. ✓ A_{truth}: Ethiopian Airlines should be to blame. ✓
Response (sad)
A_{truth}: Let's pray for all the passengers and crew members in this air crash. ✓ A_{ECM}: It's so sad. (omit emotional trigger expressions.) X

Faced with the above severe challenges that are not addressed well by existing studies of the emotional response generation task, in this paper, we propose a novel Mode-Enriched Variational Recurrent Pointer-Generator model (dubbed as **ME-VRPG**), which can generate informative and diverse emotional responses through the following three efforts: (1) We refer to the well-known Pointer-Generator model [22] with emotion factor to copy informative emotional trigger entities from the multi-turn conversation context. (2) Then we incorporate the sequential latent variables to assist with diverse reproduction of valid emotional trigger expressions. (3) We further model the explicit emotional expression with a mode controller, which can generate response words by three generative modes (i.e. copy mode, generic mode and emotional mode). Our main contributions are summarized in three folds:

- We are the first to explore the controllable emotional conversation generation task with the consideration of emotional trigger expressions, which are used to generate more informative responses. The Pointer-Generator network incorporated with emotional factor is applied to the multi-turn emotional conversation scenario.
- We further propose to incorporate the sequential latent variables into the pointing calculation of copy process, which can capture the high variability of emotional trigger expressions to generate diverse responses. Besides, an emotion mode is incorporated into the mode controller to enrich generated response with explicit emotional words.

[1)] According to our statistics on 500 randomly selected multi-turn conversations from reddit.com, 98% responses contain emotional trigger expressions, and 60% of such expressions are contained in conversation context.

- We conduct comprehensive experiments on a real-world multi-turn emotional conversation dataset. The experimental results demonstrate that our models consistently outperform strong baseline methods with large margins in promoting informativeness and diversity.

2 Related Work

2.1 Emotional Response Generation

In recent years, the emotional response generation is an emerging research trend in dialogue systems. On one hand, to control the response generation towards given emotion, the Emotional Chatting Machine (ECM) [13] utilized an internal memory module to model implicit emotional changes and an external memory module to help generate explicit emotional words. Zhou et al. [14] proposed a conditional variational autoencoders-based emotional response generation model, which exploited large scale emoji labels. Song et al. [15] incorporated emotional words and a discriminative emotion classifier to improve emotional expression. Peng et al. [23] combined topic and emotion for conversation generation. On the other hand, some other works explored the natural emotional response generation without utilizing additional emotion categories. Asghar et al. [16] introduced a sequence-to-sequence model to generate emotional response with three extensions, including affective word embedding, affect-based objective function, and affective beam search. Zhong et al. [17] introduced additional affective notations to embed words, and considered the effect of negators and intensifiers via an affective attention mechanism.

Despite the success of these pioneer works in single-turn conversation, these methods ignored the expressions of emotional trigger entities in the complex multi-turn conversation. In multi-turn emotion conversation generation, Rashkin et al. [11] proposed the retrieval-based and generative-based models for empathetic conversation generation, but this study also failed to maintain the informativeness of generated responses.

The emotional trigger expressions can intrigue the empathy between the user and the conversational system, because they can help to understand user's emotional concerns, and make the response not only informative but also emotionally acceptable.

2.2 The Pointer-Generator Network

The Pointer-Generator model [22] is the hybrid of pointer network [24] and sequence-to-sequence network, which has been widely applied in natural language generation tasks, such as text summarization [25–27], image captioning [28, 29] and question generation [30–32]. In these works, the Pointer-Generator network was used to extract entity words from source input, and generate accurate natural language sentences.

In conversation response generation task, Tam et al. [33] applied the Pointer-Generator network to generate response grounded by external knowledge. Yavuz et al. [34] proposed a fused Pointer-Generator network to copy entity words from fact sentences and conversation context. However, both methods did not explore the helpful emotional factor and the diversity of entity words.

2.3 Variational Autoencoder Model

The variational autoencoder model has been widely applied in image generation [35, 36] and text generation [37, 38], which can produce highly realistic data. Furthermore, some more advanced models, the conditional variational autoencoder [39], and the variational recurrent neural network [40] were proposed to deal with more complex tasks. The conditional variational autoencoder introduces a latent variable, which can capture the latent distribution over response for conversation generation [41, 42]. The variational recurrent neural network introduces sequential latent variables to model high variability of structured data, including machine translation [43] and response generation [44]. In conversation generation, although incorporating latent variables could help to generate diverse responses, they failed to well reproduce diverse entity words from the context.

To the best of our knowledge, we are the first to incorporate the sequential latent variables into the Pointer-Generator network, which combines the advantages of the two models, and exploits emotional trigger expressions for informative and diverse emotional conversation generation.

3 Proposed Models

In this section, we first formulate the task (Section 3.1) and then introduce our proposed framework with three innovations. First, the Pointer-Generator network is utilized to generate informative responses in multi-turn emotional conversation scenario, which is incorporated with emotion factor to

reproduce emotional trigger expressions (Section 3.2). Furthermore, we incorporate the sequential latent variables into the pointing of copy process, that is, the posterior distributions over target words participates in the attention calculation over source context in each decoding step, which can facilitate the response generation with diverse emotional trigger expressions (Section 3.3). Finally, based on above improvements, we propose a mode controller to estimate the generative probability of each word, where the distribution over an emotion vocabulary is further constructed to promote the generation of explicit emotional words (Section 3.4).

3.1 Problem Formalization

Our key notations used in this paper is described in Table 2. Suppose that there are N training conversations, and for each conversation, let $U = (w_1, ...w_i, ..., w_M)$ denotes multi-turn conversation context, where w_i is i-th word in U. Let $Y = (y_1, ..., y_T)$ denotes the emotional response sentence with specific emotion type l. Note that we assume the emotion type l of the response to be generated is given, as there exists multiple appropriate emotion categories for the same conversation context. The task paradigm of incorporating emotion is same as [13, 14]. During training, we aim to estimate the generation probability $p(Y|U, l)$ from the training dataset. During inference, we aim to find $Y^* = \underset{Y}{argmax}\, p(Y|U, l)$ for generating informative and diverse responses consistent with the specified emotion.

Table 2 Description of the key notations in our model.

Notation	Description
U	multi-turn conversation context
$\boldsymbol{e}(w_i)$	word embedding of i-th word in the context
M	number of words in the context
Y	emotional response
$\boldsymbol{e}(y_t)$	word embedding of t-th word in the response
T	number of words in response
$\boldsymbol{e}(l)$	emotion representation of emotion category l
$\boldsymbol{h}_i^U$	encoding representation of word w_i
$\boldsymbol{h}_t^Y$	decoder state in t-th step
$\boldsymbol{\alpha}_{t_i}$	attention value of context word w_i in t-th decoding step
$\boldsymbol{c}_t$	context representation in t-th decoding step
$\boldsymbol{z}_t$	sampled latent variable in t-th step
$\boldsymbol{p}_t^s$	soft switch probability
$\boldsymbol{m}_t$	generative mode in t-th step
Ω	set of out-of-vocabulary words
ϕ	parameters of recognition networks
θ	parameters of prior networks
δ	all parameters of our ME-VRPG network
V^E	emotion vocabulary
V^G	generic vocabulary
$\mathcal{L}(\delta, U, l, Y)$	objective function

3.2 Emotion-based Pointer-Generator Network

The emotional trigger expressions (including rare words and OOVs) widely exist in context of multi-turn conversations, which are difficult to be reproduced by traditional methods that are highly dependent on the quality of word embeddings. To tackle this challenge, we naturally resort to the copy capability of the Pointer-Generator network [22]. The purpose is to reproduce emotional trigger words and phrases compatible with given emotion, and maintain the informativeness of generated emotional responses.

The framework of our Emotion-based Pointer-Generator network (dubbed as **E-Pointer-Generator**) is shown in Fig. 1, where the emotion representation is incorporated into the decoding input part and the pointing calculation of copy process. Specifically, during the encoding process, the Bi-LSTM [45] is formulated as a context encoder, which contains a forward $\overrightarrow{LSTM}$ and a backward $\overleftarrow{LSTM}$. Given the word embedding matrix $E_U = [\boldsymbol{e}(w_1), ...\boldsymbol{e}(w_i), ..., \boldsymbol{e}(w_M)]$ of the context U, the output $\overrightarrow{\boldsymbol{h}}_i^U$ of word w_i encoded by the forward $\overrightarrow{LSTM}$ can be computed as:

$$\overrightarrow{\boldsymbol{\xi}}_i^U = \sigma(\overrightarrow{W}_\xi^U[\boldsymbol{e}(w_i); \overrightarrow{\boldsymbol{h}}_{i-1}^U]). \tag{1}$$

$$\overrightarrow{\boldsymbol{f}}_i^U = \sigma(\overrightarrow{W}_f^U[\boldsymbol{e}(w_i); \overrightarrow{\boldsymbol{h}}_{i-1}^U]). \tag{2}$$

$$\overrightarrow{\boldsymbol{o}}_i^U = \sigma(\overrightarrow{W}_o^U[\boldsymbol{e}(w_i); \overrightarrow{\boldsymbol{h}}_{i-1}^U]). \tag{3}$$

$$\overrightarrow{\tilde{\boldsymbol{c}}}_i^U = tanh(\overrightarrow{W}_c^U[\boldsymbol{e}(w_i); \overrightarrow{\boldsymbol{h}}_{i-1}^U]). \tag{4}$$

$$\overrightarrow{\boldsymbol{c}}_i^U = \overrightarrow{\boldsymbol{f}}_i^U \odot \overrightarrow{\boldsymbol{c}}_{i-1}^U + \overrightarrow{\boldsymbol{\xi}}_i^U \odot \overrightarrow{\tilde{\boldsymbol{c}}}_i^U. \tag{5}$$

$$\overrightarrow{\boldsymbol{h}}_i^U = \overrightarrow{\boldsymbol{o}}_i^U \odot tanh(\overrightarrow{\boldsymbol{c}}_i^U). \tag{6}$$

where $\overrightarrow{W}_\xi^U$,$\overrightarrow{W}_f^U$,$\overrightarrow{W}_o^U$ and $\overrightarrow{W}_c^U$ are learnable parameters, $\sigma(\cdot)$ is the sigmoid activation function, $\overrightarrow{\boldsymbol{\xi}}$, $\overrightarrow{\boldsymbol{f}}$ and $\overrightarrow{\boldsymbol{o}}$ are the input gate, forget gate and output gate, respectively. Similarly, we can obtain the output $\overleftarrow{\boldsymbol{h}}_i^u$ of word w_i encoded by backward $\overleftarrow{LSTM}$. Finally, the semantic representation of word w_i encoded by Bi-LSTM is the concatenation of representation obtained by $\overrightarrow{LSTM}$ and $\overleftarrow{LSTM}$, which is defined as:

$$\boldsymbol{h}_i^U = [\overrightarrow{\boldsymbol{h}}_i^U; \overleftarrow{\boldsymbol{h}}_i^U]. \tag{7}$$

where $i \in [1, M]$.

During decoding process, the decoder is formulated as a unidirectional LSTM. The initial state $\boldsymbol{h}_0^Y$ of the decoder LSTM is defined as:

$$\boldsymbol{h}_0^Y = [\overrightarrow{\boldsymbol{h}}_M^U; \overleftarrow{\boldsymbol{h}}_1^U]. \tag{8}$$

In order to capture the characters of different emotion category, we take as input the emotion category l of a response to

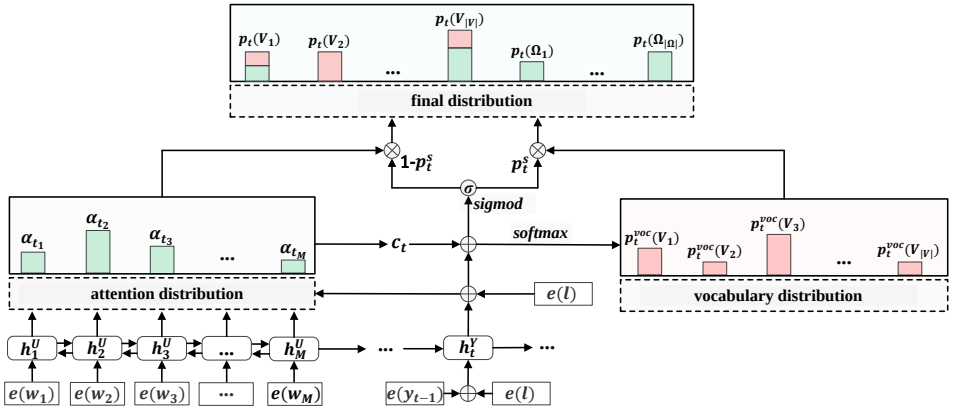

Fig. 1 The Emotion-based Pointer-Generator network (E-Pointer-Generator). Notation ⊕ denotes the vector concatenation operation.

be generated, which is represented by a low dimensional embedding $\boldsymbol{e}(l)$ and randomly initialized before training. In the t-th decoding step, we feed the concatenation of emotion category embedding $\boldsymbol{e}(l)$, word embedding of the previous word $\boldsymbol{e}(y_{t-1})$ and the context vector $\boldsymbol{c}_{t-1}$ (details are described later) into the decoder to calculate the decoder state $\boldsymbol{h}_t^Y$, which is defined as:

$$\boldsymbol{h}_t^Y = LSTM(\boldsymbol{h}_{t-1}^Y, [\boldsymbol{e}(y_{t-1}); \boldsymbol{e}(l); \boldsymbol{c}_{t-1}]). \tag{9}$$

and the specific decoding process of the unidirectional LSTM is defined as:

$$\boldsymbol{\xi}_t^Y = \sigma(W_\xi^Y[\boldsymbol{e}(y_{t-1}); \boldsymbol{h}_{t-1}^Y; \boldsymbol{c}_{t-1}; \boldsymbol{e}(l)]). \tag{10}$$
$$\boldsymbol{f}_t^Y = \sigma(W_f^Y[\boldsymbol{e}(y_{t-1}); \boldsymbol{h}_{t-1}^Y; \boldsymbol{c}_{t-1}; \boldsymbol{e}(l)]). \tag{11}$$
$$\boldsymbol{o}_t^Y = \sigma(W_o^Y[\boldsymbol{e}(y_{t-1}); \boldsymbol{h}_{t-1}^Y; \boldsymbol{c}_{t-1}; \boldsymbol{e}(l)]). \tag{12}$$
$$\hat{\boldsymbol{c}}_t^Y = tanh(W_c^Y[\boldsymbol{e}(y_{t-1}); \boldsymbol{h}_{t-1}^Y; \boldsymbol{c}_{t-1}; \boldsymbol{e}(l)]). \tag{13}$$
$$\boldsymbol{c}_t^Y = \boldsymbol{f}_t^Y \odot \boldsymbol{c}_{t-1}^Y + \boldsymbol{\xi}_t^Y \odot \hat{\boldsymbol{c}}_t^Y. \tag{14}$$
$$\boldsymbol{h}_t^Y = \boldsymbol{o}_t^Y \odot tanh(\boldsymbol{c}_t^Y). \tag{15}$$

where W_ξ^Y,W_f^Y,W_o^Y and W_c^Y are learnable parameters.

Different from the traditional Pointer-Generator network-based methods [22, 34] which only formulate the decoder state $\boldsymbol{h}_t^Y$ as the attention query, we further incorporate $\boldsymbol{e}(l)$ into the attentive calculation process. The purpose is to establish the compatibility between emotion category and emotional trigger expressions. Therefore, the concatenation of $\boldsymbol{h}_t^Y$ and $\boldsymbol{e}(l)$ is used to attend over encoding representations $\{\boldsymbol{h}_i^U | i \in [1, M]\}$ of each word in the context. Specifically, we calculate the attention distribution $\boldsymbol{\alpha}_{t_i}$ and the context vector

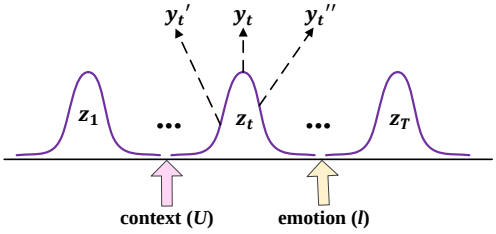

Fig. 2 Given the context and desired emotion, there exists diverse valid response words in each decoding step. z_t follows multivariate Gaussian distribution.

$\boldsymbol{c}_t$ as:

$$\boldsymbol{d}_{t_i} = v^T tanh(W_{attn}[\boldsymbol{h}_t^Y; \boldsymbol{e}(l); \boldsymbol{h}_i^U] + b_{attn}). \tag{16}$$
$$\boldsymbol{\alpha}_{t_i} = softmax(\boldsymbol{d}_{t_i}). \tag{17}$$
$$\boldsymbol{c}_t = \sum_{i=1}^{M} \boldsymbol{\alpha}_{t_i} \boldsymbol{h}_i^U. \tag{18}$$

where v, W_{attn} and b_{attn} are learnable parameters. The context vector is concatenated with the decoder state $\boldsymbol{h}_t^Y$ and emotion representation $\boldsymbol{e}(l)$, and then fed into a linear layer for producing the vocabulary distribution by softmax function. The output of the linear layer is defined as:

$$\boldsymbol{f}^V = W_V[\boldsymbol{h}_t^Y; \boldsymbol{c}_t; \boldsymbol{e}(l)] + b_V. \tag{19}$$

where W_V and b_V are learnable model parameters. For the target word y_t in the predefined vocabulary V, the vocabulary distribution is defined as:

$$\boldsymbol{p}_t^{voc}(y_t | y < t, U, l) = \frac{e^{\boldsymbol{f}_{y_t}^V}}{\sum_{j=1}^{|V|} e^{\boldsymbol{f}_j^V}}. \tag{20}$$

where $\boldsymbol{f}_{y_t}^V$ denotes the value of y_t in $\boldsymbol{f}^V$, and $\boldsymbol{f}_j^V$ denotes the value corresponding to j-th word in $\boldsymbol{f}^V$.

In t-th step of decoding, the word y_t can be generated either by vocabulary-sampling according to the probability $\boldsymbol{p}_t^{voc}(y_t)$,

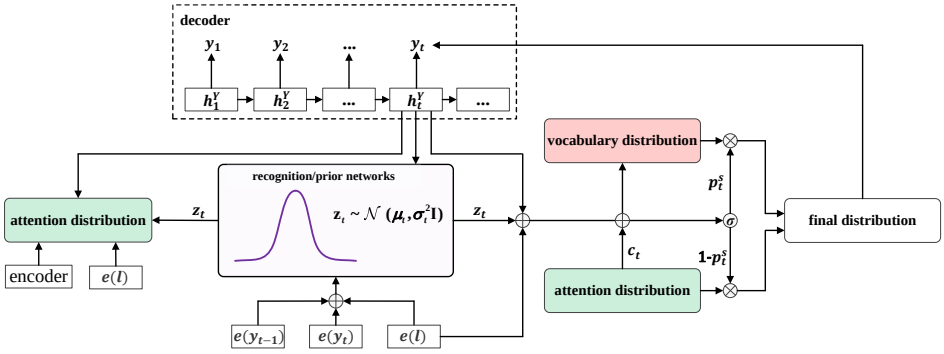

Fig. 3 The Emotion-based Variational Recurrent Pointer-Generator network (E-VRPG). The encoder part is similar to that described in Fig. 1.

or by sampling from the attention distribution. Therefore, a soft switch probability $\boldsymbol{p}_t^s \in [0, 1]$ is utilized to combine the two generation methods, which is defined as:

$$\boldsymbol{p}_t^s = \sigma(W_s[\boldsymbol{c}_t; \boldsymbol{h}_t^Y; \boldsymbol{e}(\boldsymbol{y}_{t-1}); \boldsymbol{e}(l)] + b_s). \tag{21}$$

where W_s and b_s are learnable model parameters. For each conversation, the response is generated based on the union of the predefined vocabulary and the words of conversation context. Finally, the pointing probability distribution is defined as:

$$\boldsymbol{p}_t^{copy}(y_t|y < t, U, l) = \sum_{i:w_i=y_t} \boldsymbol{\alpha}_{t_i}. \tag{22}$$

and the final generation probability distribution for predicting y_t is defined as:

$$\boldsymbol{p}_t(y_t|y < t, U, l) = \boldsymbol{p}_t^s \boldsymbol{p}_t^{voc}(y_t|y < t, U, l) + (1 - \boldsymbol{p}_t^s)\boldsymbol{p}_t^{copy}(y_t|y < t, U, l). \tag{23}$$

where $y_t \in \{V \bigcup \Omega\}$, and Ω is the set of out-of-vocabulary words contained in the context U.

3.3 Emotion-based Variational Recurrent Pointer-Generator Network

In the E-Pointer-Generator network, the decoder state is utilized to predict the time of copy via Eq. 21, and address the location of copy via Eq. 17, which is used to accurately reproduce information from context. However, in the casual and non-goal-oriented conversation utterances, multiple valid emotional trigger entities exist for emotional response generation, and only focusing on the accuracy of information reproduction could ignore the diversity character of conversation utterances. To simultaneously maintain the informativeness and diversity for emotional conversation generation task, we intend to capture the high variability of emotional trigger entities in response to refine the copy mechanism. As shown in Fig. 2, various appropriate words can be sampled from the latent distribution in each decoding step. Therefore we conjecture that incorporating the sequential latent variables into the pointing calculation of copy process can reproduce diverse emotional trigger expressions in the response.

The framework of incorporating latent variables is shown in Fig. 3. We extend the E-Pointer-Generator network into Emotion-based Variational Recurrent Pointer-Generator (dubbed as **E-VRPG**) by adapting the variational recurrent neural network (VRNN) [40], where the latent distributions over target words participate in the copy process. We assume that a latent variable sequence $\boldsymbol{Z} = z_1, z_2, ..., z_T$ exists in the latent space, where each latent variable is iteratively constructed to assist with expressing diverse emotional trigger entities. The generation probability of our E-VRPG model can be defined as:

$$p(Y, Z|U, l) = \prod_{t=1}^{T} p(\boldsymbol{y}_t|y < t, z \leq t, U, l) \cdot p(\boldsymbol{z}_t|y < t, z < t, U, l). \tag{24}$$

In t-th step decoding process, by assuming that the $\boldsymbol{z}_t$ follows the multivariate Gaussian distribution, the recognition networks parameterized by ϕ are introduced to approximate:

$$\begin{aligned} q_\phi(\boldsymbol{z}_t|y \leq t, z < t, U, l) \sim &\mathcal{N}(\boldsymbol{z}_t; \mu_t(y \leq t, z < t, U, l), \\ &\sigma_t(y \leq t, z < t, U, l)^2 \mathbf{I}). \end{aligned} \tag{25}$$

and the prior networks parameterized by θ are introduced to approximate:

$$\begin{aligned} p_\theta(\boldsymbol{z}'_t|y < t, z' < t, U, l) \sim &\mathcal{N}(\boldsymbol{z}'_t; \mu'_t(y < t, z' < t, U, l), \\ &\sigma'_t(y < t, z' < t, U, l)^2 \mathbf{I}). \end{aligned} \tag{26}$$

In the recognition networks, we project $\boldsymbol{e}(y_{t-1})$, $\boldsymbol{h}_t^Y$, $\boldsymbol{e}(l)$ and target word representation $\boldsymbol{e}(y_t)$ into the latent space to obtain $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$:

$$\boldsymbol{h}_t^z = W_\phi^z[\boldsymbol{e}(y_{t-1}); \boldsymbol{h}_t^Y; \boldsymbol{e}(y_t); \boldsymbol{e}(l)] + b_\phi^z. \tag{27}$$

$$\boldsymbol{\mu}_t = W_\phi^\mu \boldsymbol{h}_t^z + b_\phi^\mu. \tag{28}$$

$$log\boldsymbol{\sigma}_t^2 = W_\phi^\sigma \boldsymbol{h}_t^z + b_\phi^\sigma. \tag{29}$$

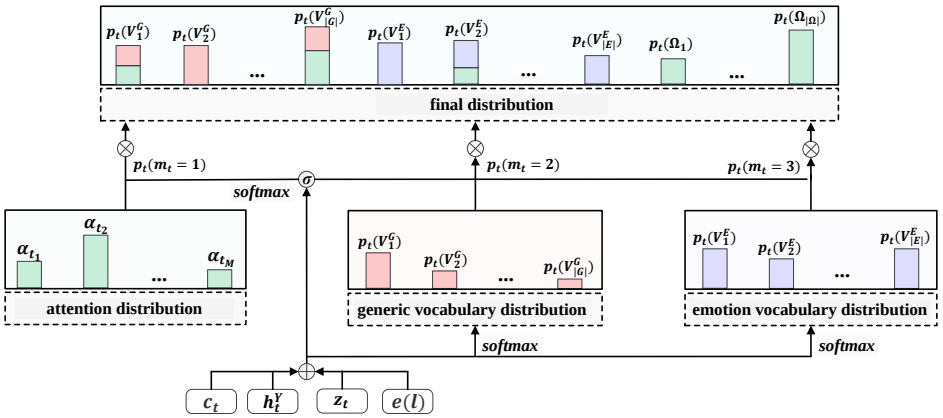

Fig. 4 The Mode-Enriched Variational Recurrent Pointer-Generator network (ME-VRPG), where the soft switch mechanism of the E-VRPG model is replaced with a novel mode controller. The attention distribution and latent variables are described in Fig. 3.

where W_ϕ^z, b_ϕ^z, W_ϕ^μ, b_ϕ^μ, W_ϕ^σ and b_ϕ^σ are trainable parameters.

During inference, the target word $\boldsymbol{e}(y_t)$ is unavailable, and the prior networks are used to obtain $\boldsymbol{\mu}_t'$ and $\boldsymbol{\sigma}_t'$, which are defined as:

$$\boldsymbol{h}_t^{z'} = W_\theta^{z'}[\boldsymbol{e}(y_{t-1}); \boldsymbol{h}_t^Y; \boldsymbol{e}(l)] + b_\theta^{z'}. \tag{30}$$

$$\boldsymbol{\mu}_t' = W_\theta^{\mu'}\boldsymbol{h}_t^{z'} + b_\theta^{\mu'}. \tag{31}$$

$$log\boldsymbol{\sigma'}_t^2 = W_\theta^{\sigma'}\boldsymbol{h}_t^{z'} + b_\theta^{\sigma'}. \tag{32}$$

where $W_\theta^{z'}$, $b_\theta^{z'}$, $W_\theta^{\mu'}$, $b_\theta^{\mu'}$, $W_\theta^{\sigma'}$ and $b_\theta^{\sigma'}$ are trainable parameters. All the latent variables are sampled by reparameterization trick [46]. The posterior distribution parameterized by $\boldsymbol{\mu}_t$ and $\boldsymbol{\sigma}_t$ is driven to be close to the prior distribution parameterized by $\boldsymbol{\mu}_t'$ and $\boldsymbol{\sigma}_t'$ using the Kullback-Leibler Divergence [47].

In the decoding process, the calculation of the t-th step decoder state in E-VRPG is defined as:

$$\boldsymbol{h}_t^Y = LSTM(\boldsymbol{h}_{t-1}^Y, [\boldsymbol{e}(y_{t-1}); \boldsymbol{e}(l); \boldsymbol{c}_{t-1}; \boldsymbol{z}_{t-1}]). \tag{33}$$

In the attention calculation, different from the attention distribution of the E-Pointer-Generator network defined in Eq. 16, we further incorporate the sampled latent variable $\boldsymbol{z}_t$ into the attention query part, which is defined as:

$$\boldsymbol{d}_{t_i} = v^T tanh(W_{attn}[\boldsymbol{h}_t^Y; \boldsymbol{e}(l); \boldsymbol{h}_i^U; \boldsymbol{z}_t] + b_{attn}). \tag{34}$$

and the pointing calculation of copy process is same as Eq. 17 and Eq. 22.

Similarly, different from the soft switch of the E-Pointer-Generator network defined in Eq. 21, the latent variable $\boldsymbol{z}_t$ is incorporated to refine the switch process between vocabulary-sampling and attention-sampling, which is defined as:

$$\boldsymbol{p}_t^s = \sigma(W_s[\boldsymbol{c}_t; \boldsymbol{e}(l); \boldsymbol{z}_t; \boldsymbol{h}_t^Y; \boldsymbol{e}(\boldsymbol{y}_{t-1})] + b_s). \tag{35}$$

3.4 Mode-Enriched Variational Recurrent Pointer-Generator

In the emotional conversation, an appropriate emotional response is usually composed of three parts, i.e. emotional trigger expressions, ordinary words and emotion words. In the E-VRPG model, the emotion representation $e(l)$ is incorporated for emotional expression in an implicit way, which is not sufficient to generate explicit emotion words.

Inspired by [48, 49] on handling the compatibility for generating different types of words, we propose a Mode-Enriched Variational Recurrent Pointer-Generator model (**ME-VRPG**), which further extends the E-VRPG model with a mode controller to construct different generative modes during each decoding step. Specifically, as shown in the Fig. 4, the mode controller is used to estimate a distribution over three generative modes, which is defined as:

$$p_t(\boldsymbol{m}_t|y < t, z \le t, U, l) = softmax(W_m[\boldsymbol{e}(\boldsymbol{y}_{t-1}); \boldsymbol{h}_t^Y; \boldsymbol{c}_t; \boldsymbol{z}_t; \boldsymbol{e}(l)] + b_m). \tag{36}$$

where W_m and b_m are trainable parameters, $\boldsymbol{m}_t = 1, 2, 3$ denotes the word y_t is generated from copy mode, generic mode and emotional mode, respectively. The overall generation distribution is defined as:

$$p_t(y_t|y < t, z \le t, U, l) = \sum_{s=1}^{3} p_t(\boldsymbol{m}_t = s|y < t, z \le t, U, l) \cdot p_t(y_t|y < t, z \le t, U, l, \boldsymbol{m}_t = s). \tag{37}$$

Copy mode: The emotional trigger expressions can be reproduced from the context by the copy mode, and the probability of generating y_t in the copy mode is defined as:

$$p_t(y_t|y < t, z \le t, U, l, \boldsymbol{m}_t = 1) = p_t^{copy}(y_t). \tag{38}$$

where $p_t^{copy}(y_t)$ is defined in Eq. 22, and the attention probability distribution is over the predefined vocabulary V and the set of out-of-vocabulary words Ω.

Generic mode: The generic mode is used to generate ordinary words, and it also retains the ability to express emotional trigger entities from the generic vocabulary. The probability of generating y_t in the generic mode can be estimated as:

$$p_t(y_t|y < t, z \leq t, U, l, \boldsymbol{m}_t = 2) = softmax(W_G[\boldsymbol{h}_t^Y; \boldsymbol{c}_t; \boldsymbol{z}_t]+b_G). \quad (39)$$

where W_G and b_G are trainable parameters, and the probability distribution $(p_t(V_1^G), p_t(V_2^G), ..., p_t(V_{|G|}^G))$ is over the generic vocabulary V^G. The size of V^G is denoted as $|G|$, and $V^G \in V$.

Emotional mode: The emotional expressions are highly related to the emotion words which carry stronger emotions than entity words and ordinary words. Therefore, we incorporate an emotional mode into the mode controller to generate explicit emotion words. In the emotional mode, we predefine an emotion vocabulary set V^E, and the probability of generating y_t in this mode is defined as:

$$p_t(y_t|y < t, z \leq t, U, l, \boldsymbol{m}_t = 3) = softmax(W_E[\boldsymbol{h}_t^Y; \boldsymbol{c}_t; \boldsymbol{e}(l); \boldsymbol{z}_t] + b_E). \quad (40)$$

where W_E and b_E are trainable parameters, and the probability distribution $(p_t(V_1^E), p_t(V_2^E), ..., p_t(V_{|E|}^E))$ is over the emotion vocabulary V^E. The size of V^E is denoted as $|E|$. Note that $V = V^E \bigcup V^G$.

After incorporating the emotional mode with the mode controller, our ME-VRPG model can not only promote the informativeness and diversity, but also express explicit emotional words in the generated responses.

3.5 Objective Function

Our ME-VRPG model is trained by maximizing the sum of variational lower bound losses at each decoding step. Each variational lower bound loss contains a log-likelihood loss and a KL divergence loss, which is defined as:

$$\begin{aligned}\mathcal{L}(\delta, U, l, Y) = \sum_{t=1}^{T}\{&-KL(q_\phi(\boldsymbol{z}_t|y \leq t, z < t, U, l) \\ &\|p_\theta(\boldsymbol{z}_t'|y < t, z < t, U, l)) \\ &+\mathbb{E}_{q_\phi(\boldsymbol{z}_t|y\leq t, z<t, U, l)}[log p_\theta(y_t|y < t, \boldsymbol{z} \leq t, U, l)]\}.\end{aligned} \quad (41)$$

where δ denotes parameters of our ME-VRPG network, and $\theta, \phi \in \delta$. We use the KL annealing strategy [50] to avoid the vanishing latent variable problem.

3.6 Algorithm Procedure

The details of training process of our ME-VRPG model is described in Algorithm 1, which contains five main steps.

Step 1 contains Line 1-2. Initialize the model parameters δ and word embeddings.

Algorithm 1 Training process of the ME-VRPG model

Input:
Training dataset: $\Lambda = \{(U_k, l_k, Y_k)\}_{k=1}^N$; Model parameters: δ; Batch size: ϵ; Number of training iterations: Φ
Output:
Learned model parameters δ of the ME-VRPG model

1: Initialize parameters δ with the truncated normal distribution
2: Initialize the word representation with pretrained Glove word embedding
3: **repeat**
4: Randomly select ϵ conversation instances Λ' from Λ
5: **for all** $(U, l, Y) \in \Lambda'$ **do**
6: For the context, compute each encoding step representation $\boldsymbol{h}_i^U$ according to Eqs. 1-7
7: Compute the overall context representation as the initial state of decoder $\boldsymbol{h}_0^Y$ according to Eq. 8
8: **for all** decoding step $t \in [1, 2, ..., T]$ **do**
9: Input emotion representation $\boldsymbol{e}(l)$ and word representation $\boldsymbol{e}(y_{t-1})$
10: Obtain the decoder state $\boldsymbol{h}_t^Y$ according to Eq. 33
11: Input the target word representation $\boldsymbol{e}(y_t)$, and compute the latent variable $\boldsymbol{z}_t$ according to Eqs. 27-29
12: Input the sampled latent variable $\boldsymbol{z}_t$, and obtain the attention (pointing) distribution $\boldsymbol{\alpha}_t$ according to Eq. 34 and Eq. 17
13: Compute the context vector $\boldsymbol{c}_t$ based on the attention distribution $\boldsymbol{\alpha}_t$ and encoder states $\boldsymbol{h}_1^U, ..., \boldsymbol{h}_M^U$, which is shown in Equation 18
14: Compute the mode distribution according to Equation 36
15: Compute the generation probability of y_t over three modes according to Equation 37-40
16: **end for**
17: Compute the loss $\mathcal{L}(\delta, U, l, Y)$ by Equation 41
18: **end for**
19: Obtain the loss over Λ': $\frac{1}{\epsilon}\sum_{(U,l,Y)\in\Lambda'}\mathcal{L}(\delta, U, l, Y)$
20: Update the model parameters δ by using the Adam optimizer
21: **until** Reach the training iterations: Φ
22: return δ

Step 2 contains Line 6-7. For each conversation instance, encode the context utterances, and obtain the word-level representation and the context-level representation. The word representations are attentively read during decoding, and the context representation is utilized as decoding initial state.

Step 3 contains Line 8-17, which corresponds to the decoding part of our model. Specifically, first, the emotion factor is utilized to reproduce emotional trigger expressions for maintaining the informativeness of generated respons-

es. Then, the sequential latent variables are incorporated into the decoding process to reproduce diverse entity words. Finally, a model controller is proposed to enrich the generated response with explicit emotional words.

Step 4 contains Line 19-20. The objective function is calculated, and the Adam optimizer is utilized to update the model parameters.

Step 5 repeats step 2, step 3 and step 4 until reaching the number of training iterations Φ, and obtain the learned model parameters δ.

4 Experiment

4.1 Dataset and Experiment Settings

A public available conversation dataset named **MED** [51] is used for our conversation generation experiment, which is the first corpus constructed for the multi-turn emotional response generation task. This MED dataset is extracted from the Reddit[2)] comments, and annotated by the well trained emotion classifier with five different emotion categories, including *Disgust*, *Happy*, *Like*, *Anger* and *Other*. The conversations are kept with appropriate length, say 4-7 turns, and each utterance has 5-30 words. The average number of words is 96.59 in context, and 24.40 in response, respectively.

The details of MED dataset statistics are displayed in Table 3, where the training set contains 803,178 conversations, validation set contains 15,000 conversations, and test set contains 15,000 conversations, respectively. We learn the response generation model from the training set, and the model performing best on validation set is used for inference. For the test data, the contexts are used for response generation, and their original responses are used as ground truth to evaluate the quality of different generative models.

We initialize the word representation with the 300d Glove embeddings [52], and all weights are initialized by the truncated normal distribution with standard deviation of 0.0001. We set the size of Bi-LSTM encoder state to 600, and set the size of unidirectional LSTM decoder state to 300. The size of latent variable is 100, and the size of mini-batch is 30. The dimension of hidden layer in the recognition network and prior networks is 300. We use KL-annealing strategy [50] varying from 0 to 1 after 100k iterations of batch training. We select 40,000 most common words as the vocabulary, including 4,289 emotion words retrieved from emotion lexicon [53]. We adopt beam search strategy [54] and the size is set to 5.

[2)] http://www.reddit.com

Table 3 The statistics of the MED dataset.

Training set	Contexts		803,178
	Emotional responses	Disgust	93,458
		Happy	166,719
		Like	78,947
		Anger	198,203
		Other	265,851
Validation set	Contexts		15,000
Test set	Contexts		15,000

All datasets are tokenized with the NLTK tokenizer [55]. We optimize our model end-to-end using Adam [56] with learning rate of 0.0002.

Pretraining is vital to the success of our ME-VRPG model, which can drive the log-likelihood losses to keep pace with KL divergences during training process, and alleviate the vanishing latent variable problems. Therefore, in this paper, we first pretrain the E-Pointer-Generator network with the mode controller, which is then used to initialize the parameters of ME-VRPG model.

4.2 Baseline methods

We compare our proposed model with several popular baseline models.

Seq2Seq: A vanilla sequence-to-sequence model with attention mechanism [57] is used to generate responses.

E-Seq2Seq: The emotion category is incorporated into the decoder part of sequence-to-sequence model, where the emotion is represented as low dimensional embedding.

Pointer-Generator: The Pointer-Generator network [22] is used to accurately copy entity words and rare words for response generation.

TA-Seq2seq: The topic-aware dialogue generation model [58] can generate informative responses with utilizing topic information.

VRNN: We use a variant of variational recurrent neural network for dialogue generation, which is similar to VRNMT [43].

ECM: The emotional chatting machine model utilizes an internal emotion memory and an external memory to generate emotional responses [13].

EmoDS: Both the explicit emotional words and implicit neutral words are exploited for dialogue generation conditioned on the specific emotion [15].

Mojitalk: The emotion category embedding is incorporated into a conditional variational autoencoders model [14].

4.3 Automatic Evaluation

The automatic evaluation metrics used in this paper are defined as:

Perplexity: The perplexity measures the quality of the trained language model, and the lower perplexity (*PPL*) value reflects better fluency [59]. The calculation method is defined as:

$$PPL = exp\{-\frac{1}{N}\sum_{k=1}^{N} log(p(Y_k|U_k, l_k))\}. \quad (42)$$

where N is the number of training conversations.

Embedding Average: The embedding average method calculates the sentence level semantic similarity between generated responses and ground-truth responses [60], which is denoted as *Eavg*. For the target response Y, the sentence representation is defined as the mean of the word embeddings of tokens:

$$\bar{e}(Y) = \frac{\sum_{t=1}^{T} \boldsymbol{e}(y_t)}{|\sum_{t=1}^{T} \boldsymbol{e}(y_t)|}. \quad (43)$$

Similarly, we can obtain the generated response representation $\bar{\boldsymbol{e}}(\hat{Y})$. The *Emb-avg* is the cosine similarity between $\bar{\boldsymbol{e}}(Y)$ and $\bar{\boldsymbol{e}}(\hat{Y})$, and the higher value reflects better semantic consistence.

Distinct1/Distinct2: The numbers of distinct unigrams/bigrams in the generated responses are denoted as *Dist1/Dist2* [3]. The higher values reflect better diversity.

Entropy: We use the entropy [61] (denoted as *Entro*) to measure the informativeness of generated responses, which is based on the information theory that low-frequent words carry more information. The higher values reflect better informativeness.

Entity: We use the name entity recognition method of spaCy[3)] to count the number of generated responses which contain entity words. The ratio (denoted as *Enti*) can be used to evaluate the informativeness of the response.

OOVs: We count the number of different out-of-vocabulary words in generated responses. The larger the number is, the stronger the ability to reproduce rare entity words and phrases is.

Accuracy: The emotion accuracy (*Acc*) is the agreement between the inputted emotion category and the predicted emotion category of a generated response. A BERT-based emotion classifier [62] is trained on 72,165 Reddit comments naturally tagged with emojis, which are divided into five emotion categories. Finally, the classifier for evaluating *Acc* obtains the accuracy of 0.674.

[3)] https://spacy.io/

The results displayed in the Table 4 are organized in two groups: 1) the models without utilizing emotion information; 2) the models for emotional dialogue generation task.

In the first group, we remove the emotion embedding from E-VRPG model for comparison, denoted as VRPG. Compared with the sequence-to-sequence model, the models incorporating copy mechanism (i.e. Pointer-Generator network and our VRPG model) can reproduce emotional trigger expressions from context to maintain the the informativeness of generated responses, and therefore obtain higher values in terms of *Entro* and *Enti*. Although the TA-Seq2Seq model incorporates the topic information, it fails to reproduce diverse emotional trigger expressions. The models incorporating latent variables (i.e. VRNN and VRPG) perform better in terms of *PPL*, *Eavg* and *Dist2*, as the latent variables can capture the high variability of target words. Furthermore, our VRPG model performs best in the first group in terms of all metrics except *Dist1*, and this is because the latent variables are incorporated to refine the copy process, and maintain the informativeness and diversity simultaneously. The reason that our VRPG model does not prevail in terms of *Dist1* is mainly that the VRPG model usually generates longer responses than the Pointer-Generator network.

In the second group, compared with the Pointer-Generator network, our E-Pointer-Generator model obtains obvious improvement in terms of *Acc*, indicating that the emotion embedding can guide the emotional response generation. Furthermore, after incorporating the emotion factor and model controller into VRPG, our ME-VRPG model performs best in terms of all evaluation matrix. This is because the the mode controller can project each word into appropriate generative mode, and the emotion factor can help copy compatible emotional trigger entities, which in turn refine accurate emotion expression.

In the Table 5, in order to investigate the influence of different factors on ME-VRPG, i.e. emotion embedding, latent variables and mode controller, we conduct ablation tests where one of the three factors is removed from ME-VRPG each time. After removing the the emotion embedding, the emotion accuracy decreases by 13.5%, indicating the emotion embedding can guide the emotional response generation, and leads to a high emotion accuracy. After removing the latent variables, the *Eavg* decreases by 3.6%, the *Dist2* decreases by 4.8% and the emotion accuracy decreases by 2.1%, because the latent variables can model the latent distribution of target words, and the ability to reproduce various emotional trigger expressions can improve semantic coherence, diversity and emotion accuracy. After removing the mode controller, the

Table 4 The automatic evaluation results. The best results in each group are highlighted for reading convenience.

Models	*PPL*	*Eavg*	*Dist1*	*Dist2*	*Entro*	*Acc*	*Enti*	*OOVs*
Seq2Seq	75.21	0.672	0.011	0.045	7.86	0.287	0.084	-
Pointer-Generator	63.78	0.689	**0.025**	0.113	8.21	0.324	0.129	546
TA-Seq2seq	59.24	0.693	0.018	0.119	8.23	0.321	0.131	-
VRNN	45.21	0.710	0.016	0.126	8.29	0.318	0.137	-
VRPG	**30.93**	**0.721**	0.021	**0.150**	**8.64**	**0.337**	**0.158**	**689**
E-Seq2Seq	73.14	0.685	0.012	0.054	7.88	0.583	0.085	-
E-Pointer-Generator	62.36	0.693	0.026	0.114	8.24	0.654	0.131	571
ECM	73.16	0.692	0.014	0.061	7.94	0.661	0.098	-
EmoDS	68.32	0.698	0.016	0.065	7.96	0.669	0.102	-
Mojitalk	47.61	0.707	0.023	0.110	8.14	0.656	0.126	-
ME-VRPG	**28.82**	**0.731**	**0.032**	**0.167**	**8.82**	**0.698**	**0.187**	**731**

Table 5 The ablation evaluation results.

Models	*PPL*	*Eavg*	*Dist1*	*Dist2*	*Entro*	*Acc*	*Enti*	*OOVs*
No emb	29.14	0.728	0.029	0.157	8.79	0.563	0.185	718
No latent variables	60.52	0.695	0.027	0.119	8.26	0.677	0.134	605
No mode controller (E-VRPG)	29.17	0.723	0.028	0.151	8.76	0.664	0.182	706
ME-VRPG	**28.82**	**0.731**	**0.032**	**0.167**	**8.82**	**0.698**	**0.187**	**731**

model is equivalent to the E-VRPG, and the emotion accuracy decreases by 3.4%, because incorporating the emotional mode could help enrich the generated responses with explicit emotion words.

4.4 Human Evaluation

We manually compare our ME-VRPG with three strong baselines (i.e. ECM, EmoDS and Mojitalk) which can generate emotional responses. We randomly sample 200 cases from the test set, and generate 800 responses corresponding to four emotion categories (i.e. *disgust*, *happy*, *like*, *anger*) for each model. Three graduate students major in sentiment analysis are recruited as human annotators. Given the conversation context and an emotion category, responses generated from the three models are randomized and presented to each annotator. The annotators are required to score a response in terms of content, emotion and informativeness, respectively. The rating scale of each criteria is 0 or 1. **Content** is defined as whether the response is coherent to the context and reasonable in logic and grammar. **Emotion** is defined as whether the emotion expression of a response agrees with the given emotion category. **Emotional trigger** is defined as whether the generated response contains appropriate emotional trigger expressions. The results are calculated by averaging the annotations from the three judges.

As shown in Table 6, the overall performances of ME-VRPG are consistently in line with the human perspective in all three criteria. It indicates that our model can express desired emotional trigger entities, which can promote content reasonability and emotion accuracy.

Besides, the pairwise comparison is applied to evaluate the overall quality of generated responses, which is shown in Table 7. The evaluated responses are generated based on randomly sampled 200 cases and the emotion category of target responses. The ME-VRPG model performs better than E-Seq2Seq (win-loss), ECM, EmoDS and Mojitalk, and improves by 47.0%, 24.2%, 23.1% and 26.7%, respectively, which further demonstrates that our novel ME-VRPG model could generate the highest quality emotional responses.

We calculate the Fleiss' Kappa [63] to measure the inter-

Table 6 Results of human evaluation.

Models	*happy*			*digust*		
	C	E	T	C	E	T
ECM	0.378	0.442	0.137	0.418	0.380	0.115
EmoDS	0.432	0.465	0.173	0.426	0.432	0.127
Mojitalk	0.488	0.396	0.245	0.423	0.348	0.148
ME-VRPG	**0.671**	**0.537**	**0.412**	**0.680**	**0.490**	**0.337**
Models	*anger*			*like*		
	C	E	T	C	E	T
ECM	0.395	0.415	0.173	0.408	0.445	0.265
EmoDS	0.402	0.443	0.175	0.413	0.482	0.270
Mojitalk	0.410	0.367	0.177	0.433	0.410	0.273
Ours	**0.623**	**0.502**	**0.365**	**0.528**	**0.562**	**0.388**

Models	*sum*		
	C	E	T
ECM	1.599	1.682	0.69
EmoDS	1.673	1.822	0.745
Mojitalk	1.724	1.521	0.843
Ours	**2.502**	**2.091**	**1.502**

Note: The criteria include content (C), emotion (E) and emotional trigger (T).

Table 7 Pairwise human evaluation.

Ours vs.	*win*	*loss*	*tie*
E-Seq2Seq	0.605	0.135	0.260
ECM	0.472	0.230	0.298
EmoDS	0.467	0.236	0.297
Mojitalk	0.492	0.225	0.283

rater consistency. Fleiss' kappa is 0.464 for results in Table 6, and 0.492 for results in Table 7, indicating "moderate agreement" for the human evaluations.

4.5 Case Study

Table 8 The case study for comparison.

Context	u_1: we may not be the best team, but we're going to look handsome.	
	u_2: for sure, whats your top 5 handsome teams in the nba?	
	u_3: easy. 1.washington 2.cleveland 3.boston 4.76ers 5.portland	
	u_4: i feel that washington is pretty top heavy with wall and kelly. i put cleveland on top with kyrie, love, osman, thompson and korver	
EmoDS	*hap*	i like them very much.
	lik	i'm sure you'll enjoy it.
	disg	they performed badly this year.
	ang	i'm worried about the wall.
Mojitalk	*hap*	i love nba teams very much.
	lik	the boston is not a good team in the nba.
	disg	they performed poorly in this season.
	ang	the games are always companied with controversy.
Ours	hap^1	i believe thompson is one of the best defenders.
	lik	i like osman and i'm sure he's pretty good.
	disg	cleveland has a rough time, and i've seen a lot of problems with kyrie.
	ang	damn, i'm not a fan of the cavs.
Ours	hap^2	i think washington is the best team in the nba.
	hap^3	well, i am also a fan of the korver.
	hap^4	cleveland is an excellent team with love.

Note: The topic of the conversation is about different teams and players of the NBA. The emotion categories include *happy* (*hap*), *like* (*lik*), *disgust* (*disg*) and *anger* (*ang*). The emotion/emotional trigger expressions are denoted by red/green colour.

We list one case from test corpus to compare different methods in Table 8. Note that four responses based on *happy* emotion are presented for our ME-VRPG model, which is used to verify the ability to reproduce diverse emotional trigger expressions. The EmoDS model performs well in emotional expressions of four emotion categories, but it tends to generate dull responses which omit emotional trigger expressions, e.g. responses in "*happy*", "*like*" and "*disgust*". Although the Mojitalk model can generate semantic coherence responses, it performs poorly in emotional expression (e.g. response in "*like*"), and in maintaining informativeness (e.g. responses in "*disgust*" and "*anger*"). Our ME-VRPG model can reproduce appropriate emotional trigger expressions from context, e.g. responses in "*happy*", "*like*" and "*disgust*", and retain the ability to express emotional trigger entities through the predefined vocabulary e.g. response in "*anger*", and therefore generate informative responses. Furthermore, our ME-VRPG model can reproduce various emotional trigger entities based on "*happy*" emotion (e.g. "washington", "korver", "cleveland" and "love"), which is used to maintain the diversity of generated responses. Besides, our model can generate the out-of-vocabulary entity words, e.g. "*osman*".

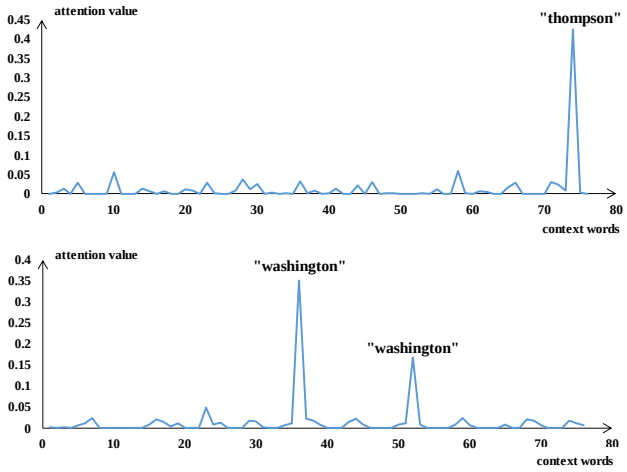

Fig. 5 Visualization of the attention (pointing) values over each word of the context in Table 8. (a) Visualization when generating word "thompson"; (b) Visualization when generating word "washington".

Furthermore, we visualize the attention distribution during decoding in Fig. 5. The Fig. 5(a) corresponds to the prediction of word "thompson" in generated response "i believe thompson is one of the best defenders.", and the Fig. 5(b) corresponds to the prediction of word "washington" in response "i think washington is the best team in the nba.". For the *happy* emotion, our model can assign higher attention weights to various emotional trigger entities in the context (i.e. "thompson" and "washington"), and therefore reproduce these words in the response with copy mechanism to generate informative and diverse emotional responses.

5 Conclusion

In this paper, we explore the multi-turn emotional conversation generation with informativeness and diversity, and propose a novel ME-VRPG model. The emotion embedding is incorporated into the Pointer-Generator network to copy informative entity words compatible to the given emotion. By incorporating the latent variables, the modelled latent distributions over target words can refine the pointing process to copy diverse emotional trigger entities. The mode controller is utilized to enrich the generated responses with explicit emotion words. The experimental results demonstrate the effectiveness of our methods. In our future work, we will explore the structured knowledge graph to extract more abundant emotional trigger expressions, which can help to interact

with users efficiently based on the shared knowledge information.

Acknowledgment

The work was supported by the National Key R&D Program of China under grant 2018YFB1004700, and National Natural Science Foundation of China (61872074, 61772122).

References

1. Chen H, Liu X, Yin D, Tang J. A survey on dialogue systems: Recent advances and new frontiers. SIGKDD Explorations, 2017, 19(2): 25–35
2. Huang M, Zhu X, Gao J. Challenges in building intelligent open-domain dialog systems. ACM Trans. Inf. Syst., 2020, 38(3): 21:1–21:32
3. Li J, Galley M, Brockett C, Gao J, Dolan B. A diversity-promoting objective function for neural conversation models. In: Knight K, Nenkova A, Rambow O, eds, NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016. 2016, 110–119
4. Zhang R, Guo J, Fan Y, Lan Y, Xu J, Cheng X. Learning to control the specificity in neural response generation. In: Gurevych I, Miyao Y, eds, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. 2018, 1108–1117
5. Song H, Zhang W, Hu J, Liu T. Generating persona consistent dialogues by exploiting natural language inference. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. 2020, 8878–8885
6. Zheng Y, Zhang R, Huang M, Mao X. A pre-training based personalized dialogue generation model with persona-sparse data. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. 2020, 9693–9700
7. Zhang H, Lan Y, Guo J, Xu J, Cheng X. Reinforcing coherence for sequence to sequence model in dialogue generation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden. 2018, 4567–4573
8. Wang W, Feng S, Wang D, Zhang Y. Answer-guided and semantic coherent question generation in open-domain conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. 2019, 5065–5075
9. Yan Z, Duan N, Bao J, Chen P, Zhou M, Li Z. Response selection from unstructured documents for human-computer conversation systems. Knowl. Based Syst., 2018, 142: 149–159
10. Zhang Z, Li J, Zhu P, Zhao H, Liu G. Modeling multi-turn conversation with deep utterance aggregation. In: Bender E M, Derczynski L, Isabelle P, eds, Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018. 2018, 3740–3752
11. Rashkin H, Smith E M, Li M, Boureau Y. Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Korhonen A, Traum D R, Màrquez L, eds, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. 2019, 5370–5381
12. S Z Q, Z X Q, L L, X M, H X J. Generating responses with a specific emotion in dialog. In: Korhonen A, Traum D R, Màrquez L, eds, Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers. 2019, 3685–3695
13. H Z, L H M, Y Z T, Y Z X, B L. Emotional chatting machine: Emotional conversation generation with internal and external memory. In: Proceedings of the AAAI. 2018, 730–739
14. D Z X, Y W W. Mojitalk: Generating emotional responses at scale. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, 1128–1137
15. Q S Z, Q Z X, L L, M X, J H X. Generating responses with a specific emotion in dialog. In: Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019, 3685–3695
16. Asghar N, Poupart P, Hoey J, Jiang X, Mou L. Affective neural response generation. In: Pasi G, Piwowarski B, Azzopardi L, Hanbury A, eds, Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings. 2018, 154–166
17. Zhong P, Wang D, Miao C. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019. 2019, 7492–7500
18. C S A. The handbook of emotion and memory: Research and theory. Psychology Press, 2014
19. Bynum IV W E, Artino Jr A R, Uijtdehaage S, Webb A M, Varpio L. Sentinel emotional events: The nature, triggers, and effects of shame experiences in medical residents. Academic Medicine, 2019, 94(1): 85–93
20. Lubis N, Sakti S, Neubig G, Yoshino K, Toda T, Nakamura S. A study of social-affective communication: Automatic prediction of emotion triggers and responses in television talk shows. In: 2015 IEEE Work-

shop on Automatic Speech Recognition and Understanding (ASRU). 2015, 777–783

21. L N, S S, N G, T T, P A, N S. Emotion and its triggers in human spoken dialogue: Recognition and analysis. In: Situated Dialog in Speech-Based Human-Computer Interaction, 103–110. Springer, 2016
22. S A, L P J, M C D. Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017, 1073–1083
23. Peng Y, Fang Y, Xie Z, Zhou G. Topic-enhanced emotional conversation generation with attention mechanism. Knowl. Based Syst., 2019, 163: 429–437
24. Vinyals O, Fortunato M, Jaitly N. Pointer networks. In: Cortes C, Lawrence N D, Lee D D, Sugiyama M, Garnett R, eds, Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. 2015, 2692–2700
25. Chung T L, Xu B, Liu Y, Ouyang C. Main point generator: Summarizing with a focus. In: Pei J, Manolopoulos Y, Sadiq S W, Li J, eds, Database Systems for Advanced Applications - 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part I. 2018, 924–932
26. Shen X, Zhao Y, Su H, Klakow D. Improving latent alignment in text summarization by generalizing the pointer generator. In: Inui K, Jiang J, Ng V, Wan X, eds, Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019. 2019, 3760–3771
27. Zhu J, Li H, Liu T, Zhou Y, Zhang J, Zong C. MSMO: multimodal summarization with multimodal output. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. 2018, 4154–4164
28. Li Y, Yao T, Pan Y, Chao H, Mei T. Pointing novel objects in image captioning. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. 2019, 12497–12506
29. Yao T, Pan Y, Li Y, Mei T. Incorporating copying mechanism in image captioning for learning novel objects. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017. 2017, 5263–5271
30. Gao S, Ren Z, Zhao Y E, Zhao D, Yin D, Yan R. Product-aware answer generation in e-commerce question-answering. In: Culpepper J S, Moffat A, Bennett P N, Lerman K, eds, Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, WSDM 2019, Melbourne, VIC, Australia, February 11-15, 2019. 2019, 429–437
31. Ma X, Zhu Q, Zhou Y, Li X. Improving question generation with sentence-level semantic matching and answer position inferring. In: The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020. 2020, 8464–8471
32. Sun X, Liu J, Lyu Y, He W, Ma Y, Wang S. Answer-focused and position-aware neural question generation. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. 2018, 3930–3939
33. Tam Y C, Ding J, Niu C, Zhou J. Cluster-based beam search for pointer-generator chatbot grounded by knowledge. In: The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, Dialog System Technology Challenges Workshop. 2019
34. Yavuz S, Rastogi A, Chao G, Hakkani-Tür D. Deepcopy: Grounded response generation with hierarchical pointer networks. In: Nakamura S, Gasic M, Zuckerman I, Skantze G, Nakano M, Papangelis A, Ultes S, Yoshino K, eds, Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue, SIGdial 2019, Stockholm, Sweden, September 11-13, 2019. 2019, 122–132
35. Hou X, Shen L, Sun K, Qiu G. Deep feature consistent variational autoencoder. In: 2017 IEEE Winter Conference on Applications of Computer Vision, WACV 2017, Santa Rosa, CA, USA, March 24-31, 2017. 2017, 1133–1141
36. Razavi A, Oord v. d A, Vinyals O. Generating diverse high-fidelity images with VQ-VAE-2. In: Wallach H M, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E B, Garnett R, eds, Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada. 2019, 14837–14847
37. Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing E P. Toward controlled generation of text. In: Precup D, Teh Y W, eds, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. 2017, 1587–1596
38. Yang Z, Hu Z, Salakhutdinov R, Berg-Kirkpatrick T. Improved variational autoencoders for text modeling using dilated convolutions. In: Precup D, Teh Y W, eds, Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017. 2017, 3881–3890
39. Yan X, Yang J, Sohn K, Lee H. Attribute2image: Conditional image generation from visual attributes. In: Leibe B, Matas J, Sebe N, Welling M, eds, Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part IV. 2016, 776–791
40. J C, K K, L D, K G, C. C A, Y B. A recurrent latent variable model for sequential data. In: Advances in Neural Information Processing Systems. 2015, 2980–2988
41. Serban I V, Sordoni A, Lowe R, Charlin L, Pineau J, Courville A C, Bengio Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In: Singh S P, Markovitch S, eds, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. 2017, 3295–3301
42. Zhao T, Zhao R, Eskénazi M. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: Barzilay R, Kan M, eds, Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver,

Canada, July 30 - August 4, Volume 1: Long Papers. 2017, 654–664

43. Su J, Wu S, Xiong D, Lu Y, Han X, Zhang B. Variational recurrent neural machine translation. In: McIlraith S A, Weinberger K Q, eds, Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018. 2018, 5488–5495
44. Du J, Li W, He Y, Xu R, Bing L, Wang X. Variational autoregressive decoder for neural response generation. In: Riloff E, Chiang D, Hockenmaier J, Tsujii J, eds, Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018. 2018, 3154–3163
45. Graves A, Fernández S, Schmidhuber J. Bidirectional LSTM networks for improved phoneme classification and recognition. In: Duch W, Kacprzyk J, Oja E, Zadrozny S, eds, Artificial Neural Networks: Formal Models and Their Applications - ICANN 2005, 15th International Conference, Warsaw, Poland, September 11-15, 2005, Proceedings, Part II. 2005, 799–804
46. Kingma D P, Welling M. Auto-encoding variational bayes. In: Bengio Y, LeCun Y, eds, 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings. 2014
47. Kullback S, Leibler R A. On information and sufficiency. The annals of mathematical statistics, 1951, 22(1): 79–86
48. Ke P, Guan J, Huang M, Zhu X. Generating informative responses with controlled sentence function. In: Gurevych I, Miyao Y, eds, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. 2018, 1499–1508
49. Wang Y, Liu C, Huang M, Nie L. Learning to ask questions in open-domain conversational systems with typed decoders. In: Gurevych I, Miyao Y, eds, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers. 2018, 2193–2203
50. Bowman S R, Vilnis L, Vinyals O, Dai A M, Józefowicz R, Bengio S. Generating sentences from a continuous space. In: Goldberg Y, Riezler S, eds, Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning, CoNLL 2016, Berlin, Germany, August 11-12, 2016. 2016, 10–21
51. C W W, S F, W G, L W D, F Z Y. A cue adaptive decoder for controllable neural response generation. In: The Web Conference 2020. 2020, 2570–2576
52. Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation. In: Moschitti A, Pang B, Daelemans W, eds, Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL. 2014, 1532–1543
53. Hu M, Liu B. Mining and summarizing customer reviews. In: Kim W, Kohavi R, Gehrke J, DuMouchel W, eds, Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Seattle, Washington, USA, August 22-25, 2004. 2004, 168–177
54. Wiseman S, Rush A M. Sequence-to-sequence learning as beam-search optimization. In: Su J, Carreras X, Duh K, eds, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. 2016, 1296–1306
55. Bird S. NLTK: the natural language toolkit. In: Calzolari N, Cardie C, Isabelle P, eds, ACL 2006, 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, Sydney, Australia, 17-21 July 2006. 2006
56. Kingma D P, Ba J. Adam: A method for stochastic optimization. In: Bengio Y, LeCun Y, eds, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015
57. D B, K C, Y B. Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. 2015
58. Xing C, Wu W, Wu Y, Liu J, Huang Y, Zhou M, Ma W. Topic aware neural response generation. In: Singh S P, Markovitch S, eds, Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA. 2017, 3351–3357
59. Vinyals O, Le Q V. A neural conversational model. CoRR, 2015, abs/1506.05869
60. Liu C, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In: Su J, Carreras X, Duh K, eds, Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016. 2016, 2122–2132
61. Mou L, Song Y, Yan R, Li G, Zhang L, Jin Z. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In: Calzolari N, Matsumoto Y, Prasad R, eds, COLING 2016, 26th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, December 11-16, 2016, Osaka, Japan. 2016, 3349–3358
62. Devlin J, Chang M, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Burstein J, Doran C, Solorio T, eds, Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers). 4171–4186
63. F J L, C J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. Educational and psychological measurement, 1973, 33(3): 613–619