Neural Computing and Applications (2021) 33:5581–5589
https://doi.org/10.1007/s00521-020-05338-z

ORIGINAL ARTICLE

Non-deterministic and emotional chatting machine: learning emotional conversation generation using conditional variational autoencoders

Kaichun Yao[1] · Libo Zhang[2] · Tiejian Luo[1] · Dawei Du[1] · Yanjun Wu[2]

Received: 21 January 2020 / Accepted: 2 September 2020 / Published online: 21 September 2020
© Springer-Verlag London Ltd., part of Springer Nature 2020

Abstract
Conversational responses are non-trivial for artificial conversational agents. Artificial responses should not only be meaningful and plausible, but should also (1) have an emotional context and (2) should be non-deterministic (i.e., vary given the same input). The two factors enumerated, respectively, above are involved and this is demonstrated such that previous studies have tackled them individually. This paper is the first to tackle them together. Specifically, we present two models both based upon conditional variational autoencoders. The first model learns disentangled latent representations to generate conversational responses given a specific emotion. The other model explicitly learns different emotions using a mixture of multivariate Gaussian distributions. Experiments show that our proposed models can generate more plausible and diverse conversation responses in accordance with designated emotions compared to baseline approaches.

Keywords Chatting machine · Conditional variational autoencoders · Non-deterministic · Neural dialog

1 Introduction

Understanding the emotional content of conversations and empathizing accordingly is a challenge for artificial conversational agents. Having emotional intelligence, i.e., to enable machine to understand affect and emotion [21], has been a long-term goal for artificial intelligence. Moreover, to express the diverse emotional contents of conversation is another important factor to generate successful artificial conversational agents. To build an interactive human like chatbot, it is absolutely essential to equip the machine with the ability of expressing and understanding emotions and learning diversity of natural languages.

The success of deep neural networks in natural language processing tasks [2, 3] promotes the exploration of the paradigm of neural dialogue generation greatly. In existing conversation-generating systems based on the neural network techniques, an encoder-decoder framework [27] has shown great potential in modeling open-domain conversations [4, 5], However, a vanilla encoder-decoder model is prone to generate dull and generic responses. To improve the quality of responses in the conversation generation, latest efforts include diversity promoting objective functions [13], diverse decoding [15], topic-introducing approaches [32] and latent variable modeling for diversity [6].

These approaches make the responses suitable in diverse contexts, while being more informative and interesting. Although a variety of models have been proposed for conversation generation from large-scale social data, it is still challenging to generate non-deterministic responses in a diverse emotional contexts. We are still far from our goal of building autonomous neural agents that can consistently carry out interesting human-like conversations. Human is able to express emotions not only naturally and coherently, but also can give non-deterministic responses at discourse level. For instance, if a user says "My dog got lost yesterday," the most appropriate response would be "It's so

Kaichun Yao and Tiejian Luo were contributed equally and should be considered as co-first authors.

✉ Libo Zhang
libo@iscas.ac.cn

[1] School of Computer Science and Technology, University of Chinese Academy of Sciences, Beijing, China

[2] State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences, No. 4 North Road, Beijing, China

sad. I am so sorry to hear about that." or "I am sorry. You must be heartbroken." to express sadness. We can also say "Bad things always happen. I hope you will be happy soon." or "Don't be sad. Your dog may come back one day." to express comfort. However, it is still difficult to incorporate the emotion factor in existing neural network models for conversation generation. This is because it is difficult to produce grammatically correct sentences with appropriate emotions. To account for emotion expression, the Emotional Chatting Machine (ECM) model [35] is proposed to generate appropriate responses not only in content (relevant and grammatical) but also in emotion (emotionally consistent). However, it usually fails to capture the diversity of emotional expression.

Our goal in this work is to generate responses for the same input sentence that carry different emotional context. We follow the NLPCC Emotion Classification Challenge where the emotions are categorized to *Anger*, *Disgust*, *Happiness*, *Like*, *Sadness* or *Other*. Adopted from the CVAE formalism [26], we propose two models. The first approach augments the unstructured variables z with a set of structured variables e, each of which corresponds to a salient and independent semantic feature (e.g., emotion categories) of responses. The second model, inspired by Gaussian Mixture model (GMM), structures the latent z space by using a set of K Gaussian priors with different means and standard deviations, which correspond to different types of emotion categories.

To this end, we design an artificial conversational chatting machine termed as non-deterministic emotional chatting machine (NECM), which generates non-deterministic responses given the same input with different emotional context that is empathetically consistent. The main contributions of this work are summarized as follows:

- We develop two novel neural dialogue models, based upon CVAE, that are able to respond colloquially under diverse emotional contexts given the same input. To the best of our knowledge, our work is the first to apply CVAE to emotional dialogue generation.
- Emotive characteristics are captured and learnt by an emotion-guided mechanism, which models the representation facilitated by a latent space.
- We elaborate an emotion gate mechanism to alleviate the problem of generating grammatically inaccurate responses, while rendering the responses emotionally correct.

2 Related work

2.1 Neural dialog models

Deep neural networks have achieved huge success in natural language processing tasks such as machine translation [27] and text summarization [23]. Recently, much attention has been drawn to the problem of generating diverse responses. To generate coherent and diverse responses, several works have focused on enhancing the input of encoder-decoder models by introducing richer context information. The speaker's characteristics are captured with an encoder-decoder model to generate more specific responses by encoding background information and speaking style into the distributed embeddings [14]. Topic encoding model-based LDA [3] is proposed to augument the model to produce more topic coherent responses [31]. Other works aim to improve the architecture of the encoder-decoder models. In order to improve the diversity in the responses, a search-based loss is introduced to optimize directly the decoder networks for beam search decoding [30]. The mutual information between input and output is proposed and maximized to optimize the standard encoder-decoder model [13]. Reinforcement learning is used for optimizing the MLE objective of an encoder-decoder model in order to encourage responses that have long-term payoff [16].

More recently, the Variational Autoencoder (VAE) [12] becomes a popular framework for dialogue generation. Viewing the dialog contexts as the conditional attributes, a novel dialogue model based on conditional variational autoencoders is proposed in [33] to generate diverse responses. To directly capture the variability in possible responses to a given input, the work of [6] introduces the latent variable in a dialogue generation model to reduce the lack of variations in the output dialogues. In question generation, CVAE have been used for image question generation in order to output multiple questions given an image [10]. The latent variable and an additional observed variable are introduced in question generation model based on text to generate different types of questions [19, 32]. Additionally, a variant of CVAE for image captioning has been used for producing diverse and accurate image description [29].

2.2 Affective response generation

In human-machine interactions, the ability to perceive human emotions and to generate appropriate response can enrich communication. However, it is difficult to incorporate the emotional contexts in generating conversations. Past work has incorporated emotions in retrieval-based or

slot-based spoken dialogue systems [21, 22] by constructing hand-crafted speech and text-based features. Recently, several studies have focused on generating text from controllable variables. Conditioned on certain attributes of the language such as sentiment, speaker's age, cultural background and gender, a neural generative model was proposed to generate sentences by combining variational autoencoders and attribute discriminators [9]. Affect Language Model (Affect-LM) was used for generate text based on the context and affect categories [8]. In large-scale conversation generation, an emotional chatting machine was first proposed to produce responses with different type of emotions by leveraging an internal memory and external memory [35]. Differed from Affect-LM, an affective dialogue system has been proposed to produce emotionally rich responses by modeling affective word embedding [1]. Similar to the work of Affect-LM, AR-S2S, an end-to-end affect-rich open-domain neural conversational model incorporating external affect knowledge, extends the Seq2Seq model and adopts VAD (Valence, Arousal and Dominance) affective notations to embed each word with affects [34]. More recently, pipeline-style methods are proposed. E-SCBA, a syntactically constrained bidirectional-asynchronous approach for emotional conversation generation, introduces pre-generated emotion keywords and topic keywords into the process of decoding [17]. A reinforcement learning (RL)-based conversation content generation model combines RL strategy with emotional editing constraints to generate more meaningful and customizable emotional responses [18]. The above end-to-end or pipeline methods used novel mechanisms to capture emotion attributes in responses, but they failed to take diversity of languages into account. Our proposed method is different, aiming at not only expressing and understanding emotions, but also learning diversity of natural languages in an end-to-end manner.

3 NECM

In this section, we introduce Non-deterministic and Emotional Chatting Machine (**NECM**), which aims to generate colloquial responses that vary with the same input and emotional context (i.e., Angry, Disgust, Happy, Like, Sad or Other). We build our framework based on conditional variational autoencoders (CVAE) that captured and learnt the emotive characteristics by latent variables z. The framework is also be able to produce meaningful and plausible responses in diverse contexts.

Standard CVAE with a fixed Gaussian prior employs an unstructured vector z in which the dimensions are entangled (i.e., emotive semantic characteristic of language mix each other in our case). To model and incorporate emotion in an interpretable way in dialogue generation, we capture different emotive semantics by using a hybrid latent space or multiple latent spaces, respectively. Therefore, we implement our approach in two different ways. In our first model, we learn a latent space to capture different emotive semantics by using an unstructured variable z which is augmented with a set of structured variables e, each of which corresponds to a salient and independent semantic feature (e.g., emotion categories) of responses. The first model is named as **NECM-Z&E**. The second model learns multiple latent spaces and explicitly structures the latent space around K components corresponding to different types of emotions and chooses one of components to create priors. We named it **NECM-GMM**. Figure 1 demonstrates the overall structure of **NECM**. The details for our two approaches are described in the following sections.

3.1 NECM-Z&E

NECM-Z&E model aims at learning disentangled representations from the unstructured part z of the representation by using the structured variable e (emotion category). The unstructured part z of the representation is modeled as continuous variables with standard Gaussian prior $p(z)$, while the structured variable e can contain discrete variables to encode different attributes of affection with appropriate prior $p(e)$. Given an input post X and an expected emotion category e of the response, we want our generative model to condition on the combined vector (z, e) and generate responses Y that fulfill the corresponding emotional attributes as specified in the structured variable e. z and e are independent in our case. n and m are the length of the input posts and the responses. We define the conditional distribution $P(Y, (z, e)|X) = P(Y|(z, e), X)P(z|X)P(e|X)$, where $P(e|X) = P(e)$ is a prior of the variable e corresponding to attributes of affection, and we explicitly specify it given observation X in our case. Therefore, our goal is to use deep neural networks (parametrized by θ) to approximate $P(z|X)$ and $P(Y|(z, e), X)$. We refer to $P_\theta(z|X)$ as the encoder and $P_\theta(Y|(z, e), X)$ as the response decoder. Then the generative process of Y can be depicted as: (1) Sample the latent variable z from the encoder $P_\theta(z|X)$. (2) Generate Y through the response decoder $P_\theta(Y|(z, e), X)$.

At training time, we follow the variational autoencoder framework [12] and introduce an approximation network $Q_\phi(z|X, Y)$ to approximate the true posterior distribution $P(z|X, Y)$. Q is a diagonal Gaussian whose parameters depend on X and Y in our case. We thus have the following evidence lower bound (ELBO) [26]:

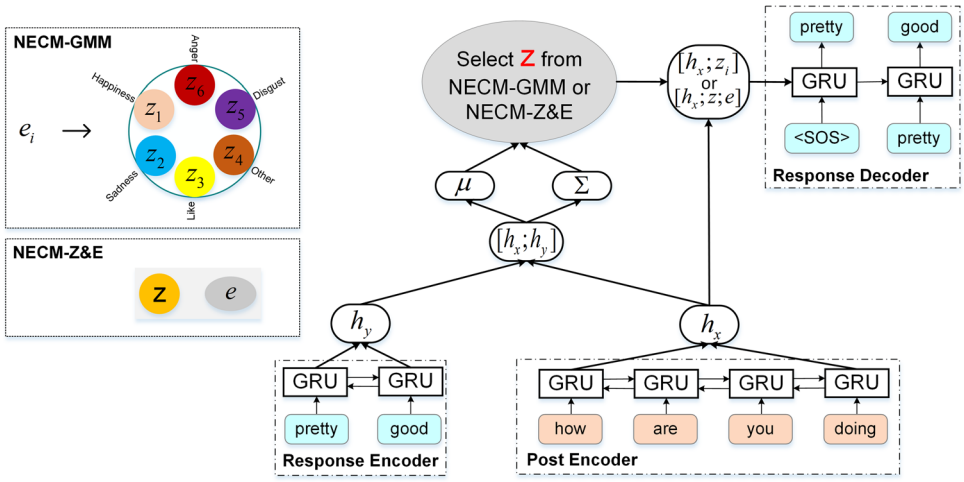

Fig. 1 Overview of our NECM model. The latent variable z is selected from either NECM-GMM or NECM-Z&E. For NECM-GMM, there are six clusters corresponding to six emotion categories and only one cluster is selected every time according to the emotion category

$$\begin{aligned}\mathcal{L}(\theta,\phi;X,Y,e) &= -KL(Q_\phi(z|X,Y)||P_\theta(z|X)) \\ &+ E_{Q_\phi(z|X,Y)}[\log P_\theta(Y|(z,e),X)] \\ &\leq \log P(Y|X)\end{aligned} \tag{1}$$

and the KL divergence between the approximate posterior and the prior will be optimized:

$$\begin{aligned}D_{\mathrm{KL}}[Q_\phi(z|X,Y),P_\theta(z|X)] &= \log\left(\frac{\sigma}{\sigma_\phi}\right) \\ &+ \frac{\sigma_\phi^2+||\mu_\phi-\mu||_2^2}{2\sigma^2}-\frac{1}{2}\end{aligned} \tag{2}$$

Given an input sentence X and a response Y, we run two separate encoders (i.e., Post Encoder and Response Encoder in Fig. 1), consisting of a bidirectional recurrent neural network (BRNN) [24] with a gated recurrent unit (GRU) [5], over their word embeddings x_i and y_i. We concatenate the final states of each and obtain our representations h_x and h_y of X and Y. Finally, we estimate the mean and variance of the approximation network Q as:

$$\begin{cases}\mu = W_\mu[h_x;h_y]+b_\mu \\ \log(\Sigma) = \mathrm{diag}(W_\Sigma[h_x;h_y]+b_\Sigma)\end{cases} \tag{3}$$

where $[h_x;h_y]$ denotes the concatenation of h_x and h_y, and *diag* denotes inserting along the diagonal of a matrix. W and b are parameters.

We then use the reparametrization trick [12] to obtain samples of z from approximation network Q and initialize the hidden state of the decoder GRU with the nonlinear transformation of these concatenated representation $s_0 = \tanh(W_0[h_x;z;e]+b_0)$, where W_0 and b_0 are learning parameters.

3.2 NECM-GMM

Standard CVAE with a fixed Gaussian prior can capture the latent distribution over all the valid responses with different attributes (e.g., affection). The dimensions in the latent z space are entangled each other and these attributes are hard separated from the unstructured vector z. Starting from the idea of multiple Gaussian priors, we encourage the latent z space to have a multi-modal structure composed of K clusters, each corresponding to different types of emotional attributes. That is, the distribution of z vectors are represented by using a Gaussian Mixture model (GMM). Then, we can model the prior $P(z|X)$ as follows:

$$P(z|X) = \sum_{i=1}^{K}\pi_i\mathcal{N}(z|\mu_i,\sigma_i^2\mathcal{I}) \tag{4}$$

where π_i is defined as the weights and μ_i represents the mean vector of the i-th component. Following the work of [29], for all components, we use the same standard deviation σ. In our current work, each component corresponds to one emotion category in {*Anger*, *Disgust*, *Happiness*, *Like*, *Sadness* and *Other*}.

NECM-GMM has the same formalism of variational lower bound in Eq. (1) as **NECM-Z&E**. The difference is that the former uses a set of K Gaussian priors (GMM prior) and the latter uses a fixed Gaussian prior, as shown in Fig. 1. In each step of training phase, according to the expected emotion category e_i of each response, we sample z from the corresponding Gaussian component. The KL divergence term in Eq. (2) need to be approximated as follows:

$$D_{\mathrm{KL}}[Q_\phi(z|X,Y,e_i),P_\theta(z|X,e_i)] = \log\left(\frac{\sigma_i}{\sigma_\phi}\right) + \frac{\sigma_\phi^2 + ||\mu_\phi - \mu_i||_2^2}{2\sigma_i^2} - \frac{1}{2} \tag{5}$$

At testing phase, we first specify an emotion category and then sample z from the corresponding component distribution. We initialize the hidden state of the decoder GRU with the nonlinear transformation of these concatenated representation $s_0 = \tanh(W_1[h_x; z] + b_1)$, where W_1 and b_1 are learning parameters.

As shown in Fig. 1, we can note that **NECM-GMM** employs the same encoder, the approximation network and response decoder as **NECM-Z&E**. However, different with **NECM-Z&E** learning disentangled latent representations, **NECM-GMM** expects each part of the latent representation to focus on one aspect of the samples. That is, each of the latent variables z_i (i.e., z_1,z_2,z_3,z_4,z_5 and z_6 in Fig. 1) only concentrates on the generated responses with certain emotion in our case.

3.3 Response decoder and emotion category embedding

We employ an attention-based GRU response decoder to predict the words in response Y sequentially. In affective response generation task, an emotion category can provide a high-level abstraction of an emotion expression. Therefore, we embed the emotion category into low dimensional and real-valued vector and take it as additional input of the response decoder. That is, at decoding time step t, the GRU decoder reads the previous word embedding y_{t-1}, the vector of an emotion category v_e, and context vector c_{t-1} to compute the new hidden state s_t.

$$s_t = \mathrm{GRU}(s_{t-1}, [c_t; y_{t-1}; v_e]) \tag{6}$$

The context vector c_t for current time step t is computed through the attention mechanism [20] as follows:

$$\begin{cases} e_{t,i} = v^T \tanh(W_e s_{t-1} + U_e h_i^e) \\ \alpha_{t,i} = \dfrac{\exp(e_{t,i})}{\sum_{i=1}^{n} \exp(e_{t,i})} \\ c_t = \sum_{i=1}^{n} \alpha_{t,i} h_i^e \end{cases} \tag{7}$$

where W_e, U_e and v^T are learning parameters. h_i^e is the hidden state of representation of the i-th word in the input X.

In practice, it is beneficial to capture emotional attributes of sentences in response generation by introducing emotion category embedding in the decoder. However, this scheme often hurts grammaticality of generated responses. Therefore, we introduce an emotion gate mechanism to balance the weights between grammatical coherence and emotional attributes of a generated response. Then, it can automatically decide how much emotion will be considered at the next decoding step.

$$\mathrm{egate} = \mathrm{sigmoid}(W_g[s_{t-1}; c_t] + b_g) \tag{8}$$

where W_g and b_g are parameters. Then the emotion category embedding vector v_e is updated as $v_e = \mathrm{egate} * v_e$. Finally, the probability of each target word y_t is predicted based on all the previously generated words (i.e., $y_{<t}$) and the input sentence X.

$$P(y_t|X, y_{<t}) = \mathrm{softmax}(V[s_t; c_t] + b) \tag{9}$$

where V and b are learnable parameters.

3.4 Emotion-guided mechanism

We assume that the emotional features are beneficial for our model to learn meaningful latent z variable in our case. We concatenate the semantic representation of input post X and z and pass it through an multilayer perceptron (MLP) to predict emotion category $e = \mathrm{MLP}(z, X)$. To encode emotion-related information into the latent z space more efficiently, we rewrite our loss function in Eq. 1 as follows:

$$\begin{aligned} \mathcal{L}(\theta,\phi;X,Y,e) = &-\mathrm{KL}(Q_\phi(z|X,Y)||P_\theta(z|X)) \\ &+ E_{Q_\phi(z|X,Y)}[\log P_\theta(Y|(z,e),X)] \\ &+ E_{Q_\phi(z|X,Y)}[\log P_\theta(e|z,X)] \end{aligned} \tag{10}$$

4 Experiments

4.1 Dataset

We evaluate the proposed method on the dataset provided by the Emotional Conversation Generation Challenge.[1] The dataset is constructed from postings and follow-up comments from the Chinese social media platform Weibo (www.weibo.com) . It contains more than 1 million utterance-response pairs.

The dataset includes the original posts, the corresponding responses, and labels of each post and response. These labels are obtained by an emotion classifier that is based on a bidirectional LSTM model. The classifier was trained on the data from the NLPCC Emotion Classification Challenge.[2] In this dataset, emotions are divided into the six

[1] http://aihuang.org:8000/p/challenge.html.

[2] http://tcci.ccf.org.cn/conference/2014/dldoc/evatask1.pdf.

basic emotion categories (i.e., *Anger*, *Disgust*, *Happiness*, *Like*, *Sadness* and *Other*).

4.2 Implementation details

We choose the top 30,000 most frequent tokens for the source and target vocabulary. All other tokens outside the vocabulary list are replaced by the *UNK*(unknown) symbol. The dimension of word embedding is set to 200. All our model variants use a single-layer bidirectional GRU encoder and a single-layer GRU decoder. The size of the GRU hidden unit is set to 256. We update the model parameters using the Adam [11] algorithm with a mini-batch size of 128. The dimension of the latent variable z is set to 200, and the size of emotion category embedding is set to 30. For decoding, we use beam search with a width of 10.

4.3 Approaches

We conducted experiments using the following dialogue generation approaches.

seq2seq_emb: We adapt a general seq2seq model with attention mechanism [2].

CVAE_emb: We use "vanilla" version of the CVAE model with a fixed Gaussian prior following [33].

SentiGAN: we adapt SentiGan [28] for emotional dialogue generation, in which GAN and the reinforcement learning strategy are used to support the generation of emotional text.

E-SCBA: We implemented a syntactically constrained bidirectional-asynchronous approach for emotional conversation generation [17].

ECM: We implemented the Emotional Chatting Machine proposed in [35] for generating appropriate responses not only in content but also in emotion.

NECM-Z&E: We implemented the first approach presented in Sect. 3.1. The model augments the unstructured variables z with a set of structured variables e, each of which corresponds to a salient and independent semantic feature.

NECM-GMM: We proposed the second approach presented in Sect. 3.2. The model clusters the latent z space around six components corresponding to six emotion categories and combines components to create priors for dialogue generation.

4.4 Evaluation metrics

Existing metrics (e.g., BLEU, ROUGE and METEOR) are not suitable for evaluating dialogue generation since these metrics have not been correlated with human judgments [19], and more importantly, the accuracy of the sentiment cannot be evaluated by these metrics. Therefore, we adopt the perplexity metric from the work in [35] to evaluate the performance of the proposed models, reflecting that whether the generated responses are grammatically correct grammatical and relevant at the content level. We also adopted the emotive accuracy to evaluate the model at the emotional context. The emotion accuracy is used to evaluate the agreement between the expected emotion category (one of the inputs to the model) and the predicted emotion category of a generated response. The emotive accuracy is computed as the classification scores of the generated input classified by a bidirectional LSTM-based emotion classifier that is trained on the NLPCC emotion classification dataset.

In addition, we perform also human evaluation to evaluate the quality of the generated responses at content level and at emotive level, as in [35].

- *Content Relevancy* is evaluated as whether the response is meaningful and appropriate to a post and could exhibit ability that equivalent or distinguished from a human. The metric proposed in [25] has been widely accepted in conversation generating tasks.
- *Emotion Consistency* is evaluated as whether the emotion expression of a generated response corresponds to the given emotion category. The responses generated by our models are randomized and presented to five human raters, who are asked to score a response in terms of *Content* (rating scale is 0, 1, 2)[3] and *Emotion* (rating scale is 0, 1).[4]

In our experiments, we also evaluate the generative abilities of our different approaches, using a method similar to the work of [1]. Specifically, *syntactic diversity* and *affective diversity* was evaluated by five human raters in this experiment. The former evaluates the discourse-level diversity and the latter judges the responses under different emotional contexts. The rating scale is 0, 1, 2 and 3 with labels *bad*, *satisfactory*, *good* and *very good*, respectively.

4.5 Results and analysis

4.5.1 Perplexity and emotive accuracy

The results are presented in Table 1. Note that we expect the perplexity to be as lower as possible and accuracy to be as high as possible. As can be seen, NECM-GMM obtains the best performance in both perplexity and emotive accuracy. In our baselines, E-SCBA performs best. Its

[3] 0, 1 and 2 are content scores. 0 denotes content irrelevancy, 1 denotes moderately relevant content and 2 denotes content relevancy.

[4] 0 and 1 are emotion scores. 0 denotes that the emotion in response generated by our models is inconsistent with the given emotion category, and 1 denotes that the emotion in response is consistent with the given emotion category.

Table 1 Objective automatic evaluation with perplexity and accuracy

Model	Perplexity	Accuracy
seq2seq	69.2	0.168
CVAE	60.3	0.196
SentiGan	60.1	0.761
E-SCBA	59.2	0.786
ECM	60.5	0.767
NECM-Z&E	56.2	0.772
w/o embed	58.8	0.763
w/o emo-gate	59.6	0.757
w/o emo-guided	58.1	0.755
NECM-GMM	**55.4**	**0.792**
w/o embed	56.5	0.778
w/o emo-gate	57.6	0.764

The best performing method for each column is highlighted in bold

underlying reason may be to explicitly introduce emotion keywords and topic keywords into the process of decoding. Moreover, NECM-Z&E performs comparably with ECM in emotion accuracy. It indicates our method can disentangle the representation to separate emotion information from the unstructured z space using the combined variables (z, e). Notably, the emotive accuracies of seq2seq and CVAE is extremely low. This is because the former generates the same response for different emotion categories and the latter is difficult to produce responses with different emotional attributes.

We also conduct ablation test to investigate the influence of different modules in our method, with the results shown in Table 1. Specifically, we remove one of the three modules (i.e., emotion category embedding, emotion gate mechanism and emotion-guided mechanism) from NECM in each set of experiment. These results show that when the emotion gate mechanism has been removed, both NECM-Z&E and NECM-GMM models produce lower accuracy and higher perplexity. This indicates that the emotion gate mechanism helps generating responses not only emotionally correct, but also grammatically correct. This helps to alleviate the problem that the model may generate grammatically inaccurate responses when using the emotion category embedding model alone. After removing the emotion-guided mechanism, the emotive accuracy decreases the most. This is because it can encode emotion-related information into our latent variable for robust performance.

4.5.2 Content and emotion

Two hundred posts are randomly sampled from the test set as input to all approaches. Each generated response will be asked to be scored by five human raters. The numbers in Table 2 are the average scores calculated by scores for all responses. In generating phase, the emotion category label is needed as the extra input. As shown in Table 2, NECM-GMM outperforms the other approaches slightly in terms of *Emotion* and *Content*. We also evaluate inter-annotator consistency using Fleiss' Kappa [7] and obtained the score of 0.445 and 0.761 for *Content* and *Emotion*, interpreted as "Moderate agreement" and "Substantial agreement" among the judges, respectively.

4.5.3 Diversity

The results of diversity evaluation are presented in Table 3. As can be seen, the performance of our NECM methods in diversity is significantly better than other baselines. CVAE obtained better performance than other baselines in *syntactic diversity*. This is because CVAE also uses the latent variables to capture the diversity by a fixed Gaussian prior. Different from CVAE, our NECM-GMM method uses GMM priors to achieve better diversified responses. The Fleiss' Kappa score for diversity is 0.463 for NECM-GMM, indicating a "Moderate agreement" among the judges.

Table 2 Manual evaluation of the generated responses in terms of *Content* and *Emotion*

Model	Content	Emotion
seq2seq	1.262	0.146
CVAE	1.278	0.152
SentiGan	1.301	0.452
E-SCBA	1.31	0.461
ECM	1.309	0.447
NECM-Z&E	1.312	0.451
NECM-GMM	**1.316**	**0.465**

The best performing method for each column is highlighted in bold

Table 3 Diversity evaluation of the generated responses in terms of *Syntax diversity* and *Affective diversity*

Model	Syntactic diversity	Affective diversity
seq2seq	0.66	0.25
CVAE	1.92	0.31
SentiGan	1.72	1.78
E-SCBA	1.59	1.76
ECM	1.61	1.77
DEMC-Z&E	2.24	2.21
DEMC-GMM	**2.46**	**2.35**

The best performing method for each column is highlighted in bold

Post	Model		Response(Original)	Response(Translated)
早上的天气不错啊！哈哈哈哈哈哈 The weather is nice in the morning! Hahahahahaha	Like	ECM	今天的天气不错啊！	Today's weather is great!
		NECM-Z&E	早上好舒服好舒服	It's very comfortable in the morning, so comfortable!
			亲爱的，天气真好！	Darling, it's a nice day!
		NECM-GMM	那是相当不错	That's pretty good
			是啊，我也觉得很不错，很不错！	Yes, I think it's nice too. It's good!
	Happy	ECM	今天的天气真好啊！	Today's weather is so good!
		NECM-Z&E	呵呵，早上好！	Well, good morning!
			是啊是啊	Hear hear
		NECM-GMM	哈哈，好吧，早上好	Ha-ha, alright, Good morning
			哈哈，那是那是！哈哈！	Ha-ha, hear hear! Ha-ha
	Sad	ECM	天气好，心情不好。	Weather is good, but I am not in a good mood.
		NECM-Z&E	是啊，好累啊！	Yeah, I am so tired.
			可怜的孩纸伤不起啊	Poor me. I can't take this anymore
		NECM-GMM	唉，我是可怜的娃啊！	Uggh, poor me!
			那是相当的悲催啊！	What a pity!
	Disgust	ECM	这是哪里啊?求解释	Where am I? Please explain
		NECM-Z&E	天气太热了吧?	Isn't it too hot?
			你是不是很无聊???	Aren't you boring???
		NECM-GMM	是啊，这天气太热了，搞不动了	Yeah, the weather is too hot to do anything
			还真有点冷	It's a bit cold
	Angry	ECM	怎么了？什么情况?	What's up? What's going on?
		NECM-Z&E	啊。你是故意的！！！	Arg. You did it on purpose!!!
			气死我了气死我了	I am so mad
		NECM-GMM	我还以为你们是故意的！！！	I thought you did it on purpose!!!
			坑爹啊坑爹	Bloody hell. It's non-sense.
知道真相的我眼泪笑下来 Knowing the truth, I laughed myself to tears	Like	ECM	太可爱了，太可爱了！	Too lovely, it's too lovely.
		NECM-Z&E	真相帝，我爱你！	I love the truth!
			真相的力量太强大了	The power of the truth is too strong
		NECM-GMM	其实我也很喜欢这句话	I really like it, too.
			这是什么意思！！！笑死我了	What do you mean by this!!! I am dying in laughter.
	Happy	ECM	笑得我眼泪都出来了。	I can't help laugh myself to tears.
		NECM-Z&E	我也是这么想的，哈哈。	I think so, ha-ha.
			真相帝笑死我了	The truth made me laugh so hard
		NECM-GMM	开心开心地笑	Laugh happily and happily
			哈哈哈。。。笑死我了	Ha-ha-ha… I am dying in laughter
	Sad	ECM	知道真相的我眼泪掉下来。	After knowing the truth, I burst into tears.
		NECM-Z&E	真相的结局.....	The end of the truth…
			你看到的时候很开心吗?	Were you happy when you've seen it?
		NECM-GMM	我哭的稀里哗啦的.....	I cried so hard…
			知道真相的我眼泪掉下来	After knowing the truth, I burst into tears.
	Disgust	ECM	真是真的还是假的?	Is this true?
		NECM-Z&E	真相，你懂的。	The truth, you know…
			嗯嗯，真相帝说的是事实啊，你懂的	Hmmm, the truth is the truth, you know.
		NECM-GMM	这是赤裸裸的讽刺啊！！！鄙视你	This is truthly sarcastic!!! I despise you
			我还以为是真相了，原来是这样的人啊。	I thought it was the truth, I didn't know you were such a person.
	Angry	ECM	什么意思？没看懂。	What do you mean? I don't understand.
		NECM-Z&E	我还以为你是真相帝啊.....	I thought you knew the truth…
			至于你信不信，反正我信了！	Whatever you believe or not, I believe in it!
		NECM-GMM	真相！！！吓死我了	The truth freaked me out
			是真相了吗?你说的是事实吗??	Is it the truth? Are you talking about the truth??

Fig. 2 Sample responses generated by the Baseline (ECM) and our NECM (original Chinese and English translation)

We can observe that NECM-GMM slightly outperforms DEMC-Z&E on all evaluation metrics. The potential reason may be that the former can better capture the emotive semantics than DEMC-Z&E by learning multiple latent spaces, instead of learning a hybrid latent space. In addition, to further illustrate the diversity of the generated responses given a post, some examples generated from our NECM are shown in Fig. 2. We note that NECM is capable of producing non-deterministic and emotional responses under the corresponding emotion category.

5 Conclusion and future work

In this work, we propose a non-deterministic and emotional chatting machine(NECM) to learn understanding and expressing emotion and generating diversified responses in dialogue generation. Two proposed models are based on conditional variational autoencoders. We introduce emotion gate mechanism and emotion-guided mechanism to help the models generate multiple emotionally and grammatically correct responses based on the same input. Experiments show that our approach can generate responses appropriate that is capable of not only capturing emotions but also learning diversity of natural languages. The potential future direction would be to explore the area that allows the system to gauge the emotional content automatically. Meanwhile, we would utilize adversarial learning to generate more human-like responses in terms of emotion attributes and diversity expression of languages.

Acknowledgements This work was supported by the National Natural Science Foundation of China, Grant No. 61807033, the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBS-LY-JSC038. Libo Zhang was supported by Youth Innovation Promotion Association, CAS (2020111) and Outstanding Youth Scientist Project of ISCAS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Asghar N, Poupart P, Hoey J, Jiang X, Mou L (2018) Affective neural response generation. In: ECIR, pp 154–166
2. Bahdanau D, Cho K, Bengio Y (2014)Neural machine translation by jointly learning to align and translate. CoRR arXiv:abs/1409.0473
3. Blei DM, Ng AY, Jordan MI (2001) Latent Dirichlet allocation. In: NIPS, pp 601–608
4. Callejas Z, Griol D, López-Cózar R (2011) Predicting user mental states in spoken dialogue systems. EURASIP J Adv Signal Process 2011:6
5. Chung J, Gülçehre Ç, Cho K, Bengio Y (2014) Empirical evaluation of gated recurrent neural networks on sequence modeling. CoRR arXiv:abs/1412.3555
6. Clark S, Cao K (2017) Latent variable dialogue models and their diversity. In: EACL, pp 182–187
7. Fleiss JL (1971) Measuring nominal scale agreement among many raters. Psychol Bull 76(5):378–382
8. Ghosh S, Chollet M, Laksana E, Morency L, Scherer S (2017) Affect-lm: a neural language model for customizable affective text generation. In: ACL, pp 634–642
9. Hu Z, Yang Z, Liang X, Salakhutdinov R, Xing EP (2017) Toward controlled generation of text. In: ICML, pp 1587–1596
10. Jain U, Zhang Z, Schwing AG (2017) Creativity: generating diverse questions using variational autoencoders. In: CVPR, pp 5415–5424
11. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization. CoRR arXiv:abs/1412.6980
12. Kingma DP, Welling M (2013) Auto-encoding variational bayes. CoRR arXiv:abs/1312.6114
13. Li J, Galley M, Brockett C, Gao J, Dolan B (2016) A diversity-promoting objective function for neural conversation models. In: NAACL, pp 110–119
14. Li J, Galley M, Brockett C, Spithourakis GP, Gao J, Dolan WB (2016) A persona-based neural conversation model. In: ACL
15. Li J, Monroe W, Jurafsky D (2016) A simple, fast diverse decoding algorithm for neural generation. CoRR arXiv:abs/1611.08562
16. Li J, Monroe W, Ritter A, Jurafsky D, Galley M, Gao J (2016) Emnlp, pp 1192–1202
17. Li J, Sun X (2018) A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. In: EMNLP. Association for Computational Linguistics, pp 678–683
18. Li J, Sun X, Wei X, Li C, Tao J (2019) Reinforcement learning based emotional editing constraint conversation generation. CoRR arXiv:abs/1904.08061
19. Liu C, Lowe R, Serban I, Noseworthy M, Charlin L, Pineau J (2016) How NOT to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation. In: EMNLP, pp 2122–2132
20. Luong T, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: EMNLP, pp 1412–1421
21. Picard RW (2002) Affective computing. Technical Report vol 1(1), pp 71–73
22. Pittermann J, Pittermann A, Minker W (2010) Emotion recognition and adaptation in spoken dialogue systems. Int J Speech Technol 13(1):49–60
23. Rush AM, Chopra S, Weston J (2015) A neural attention model for abstractive sentence summarization. In: EMNLP, pp 379–389
24. Schuster M, Paliwal KK (1997) Bidirectional recurrent neural networks. IEEE Trans Signal Process 45(11):2673–2681
25. Shang L, Lu Z, Li H (2015) Neural responding machine for short-text conversation. In: ACL, pp 1577–1586
26. Sohn K, Lee H, Yan X (2015) Learning structured output representation using deep conditional generative models. In: NIPS, pp 3483–3491
27. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. In: NIPS, pp 3104–3112
28. Wang K, Wan X (2018) Sentigan: generating sentimental texts via mixture adversarial networks. In: IJCAI, pp 4446–4452. ijcai.org
29. Wang L, Schwing AG, Lazebnik S (2017) Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space. In: NIPS, pp 5758–5768
30. Wiseman S, Rush AM (2016) Sequence-to-sequence learning as beam-search optimization. In: EMNLP, pp 1296–1306
31. Xing C, Wu W, Wu Y, Liu J, Huang Y, Zhou M, Ma W (2016) Topic augmented neural response generation with a joint attention mechanism. CoRR arXiv:abs/1606.08340
32. Xing C, Wu W, Wu Y, Liu J, Huang Y, Zhou M, Ma W (2017) Topic aware neural response generation. In: AAAI, pp 3351–3357
33. Zhao T, Zhao R, Eskénazi M (2017)Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In: ACL, pp 654–664
34. Zhong P, Wang D, Miao C (2019) An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In: AAAI, pp 7492–7500. AAAI Press
35. Zhou H, Huang M, Zhang T, Zhu X, Liu B (2018) Emotional chatting machine: Emotional conversation generation with internal and external memory. In: AAAI

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.