TERG: Topic-Aware Emotional Response Generation for Chatbot

Pei Huo†, Yan Yang* †, Jie Zhou†, Chengcai Chen@, Liang He†

†*Department of Computer Science and Technology, East China Normal University*

@*Xiaoi Research, Xiaoi Robot Technology Co., Ltd*

Shanghai, China

*yanyang@cs.ecnu.edu.cn

***Abstract*—A more intelligent chatbot should be able to express emotion, in addition to providing informative responses. Despite much works in designing neural dialogue generation systems in recent years, few studies consider both emotion to be expressed and topic relevance in the generation process. To address this problem, we present a Topic-aware Emotional Response Generation (TERG) model, which can not only exactly generate desired emotional response but perform well in topic relevance. Specifically, TERG equips an encoder-decoder structure with an emotion aware module to control the emotional sentence generation and a topic aware module to enhance topic relevance. We evaluate our model on a large real-world dataset of conversations from social media. Experimental results show that our model obtains a significant improvement against several strong baseline methods on both automatic and human evaluation.**

***Index Terms*—dialogue generation, emotion, topic aware commonsense, latent variable, Seq2Seq, CVAE**

I. Introduction

With the availability of large-scale dialogue corpus, there is a boom in research on open-domain chit-chat dialogue systems. Emotion expression is an important inherent attribute in the dialogue system. In recent years, some research is about how to supply chatbots with an emotion expression ability. The studies [1], [2] have proved that the emotional chatbot can significantly improve the user satisfaction and enrich the human-computer interactions.

Early related studies [3]–[5] are either rule-based, retrieval-based, or limited to small-scale data that can hardly express complex, various emotions and difficult to scale well to large datasets. Most recently, sequence to sequence (Seq2Seq) with attention [6] represents a good neural network framework for dialogue generation. Zhou et al. [7] proposed an emotional chatting machine (ECM) based Seq2Seq that is able to generate a specific emotional response. Immediately after this work, Asghar et al. [8] constructed an emotional dialogue system by adding affective word embeddings, the affective object function and diverse beam search algorithms.

Although current emotional generative conversation models have achieved promising results, they still suffer from the following issues. First, these models tend to generate trivial or universally relevant responses with little meaning like " Haha", " I love you ", "I hate you" due to the addition of emotional factor. Second, they tend to ignore topic relevance in generating emotional responses. As widely acknowledge, the conversations between humans are usually limited to a particular topic during a period of time. What's more, we use the efficient unsupervised topic model BTM [9] to analyse the topic relevance of real-word conversations much higher than the generated ones, more details in Section *V.D*. As shown in Figure 1, we can intuitively found the importance of topic relevance in an emotional dialogue system. Given a message about basketball, the natural responses should also be basketball related, but the responses from existing generative models are rarely related to basketball.

Emotion Label	Message: My favorite sport is playing basketball.	
	Generated responses	**Real-life responses**
Happy	What a happy day!	I will be very excited after playing basketball.
Like	I like it.	I also like playing basketball.
Disgust	I hate playing piano.	The basketball game is too difficult for me.
Sad	I am so sad.	The team lost yesterday, and the players were frustrated.
Angry	I am so angry.	I am very annoyed because I cannot throw the ball into the basket.
Null	We will not play basketball tomorrow.	No one doubts Kobe's talent on basketball.

Fig. 1. The comparison of generated responses and real responses. Emotion-related words are in red, keywords in blue and others are ordinary words.

To comprehensively consider emotion and topic factors in response generation, we present a Topic-aware Emotional Response Generation (TERG) model. In the following, we refer to the architecture with the abbreviation TERG. Our model equips an encoder-decoder structure with an Emotion-Aware module (EA) to control the emotional sentence generation and a Topic Commonsense-Aware module (TCA) to enhance topic relevance. In EA module, we use a learnable latent variable to learn the semantic information of the specific emotion response and three kinds of word distribution: emotion-related words, keywords and ordinary words. In decoding, the latent variable and emotion label embedding are fed into each decoder unit and the word type distribution obtained by the latent variable will be used to explicit modulate the generation distribution of the entire vocabulary. In addition, we introduce the TCA module to enrich the dialogue topic relevance, which can obtain external topic commonsense and then integrated into generation process in the form of attention fusion.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

Automatic and human evaluations demonstrate that our model improves both the topic relevance and emotion expression precision, compared to strong baselines.

II. Related Work & Background

Neural response generation models are built upon the encoder-decoder framework [6]. The research of generation emotional response is an important step for building a more intelligent chatbot. Zhou et al. [7] proposed an emotional chat machine (ECM) utilizing emotion category embeddings, internal emotion states, and external emotion vocabulary. ECM only performs better in several specific emotion categories in which there are sufficient training data. Immediately after this work, Asghar et al. [8] used an affective dictionary to add three dimensions for each word embedding for constructing the affective word embeddings. And they also proposed an affective object function and an affective diverse beam search algorithm to generate proper emotional response. Some people use multi-task learning for building emotional conversation [10], but the model is so rude that the experiment is not good. Song et al. [11] proposed an emotional dialogue system (EmoDS) that is able to put a specific emotion into responses explicitly or implicitly. Ekman et al. [12] proposed the emotion classification method whose author is one of the earliest emotion theorists. According to their theory, there are six basic emotions: anger, disgust, fear, joy, sadness and surprise. In our work, we made a slight adjustment to this classification metrics according to the existing corpus.

In our model, we harness a latent variable in the Conditional Variational Autoencoder (CVAE) [13] framework to project different emotional responses into a latent space. CVAE based model is developed from VAE by introduce additional condition. More specifically, CVAE characterizes the conditional generation problem using three random variables: message X, target response Y and latent variable z, which is used for modelling the latent distribution of semantic over responses given a message. The generative objective function can be expressed as $P(Y,z|X) = P(z|X)P(Y|X,z)$. Assuming the given message X as the condition, the prior distribution of latent variable z can be determined as $p_\theta(z|X)$. Each response Y can sample a latent variable from this prior distribution $p_\theta(z|X)$, then Y can be generated by the decoder $p_\theta(Y|X,z)$. In inference stage, the training data X and Y are used to get posterior distribution $q_\phi(z|X,Y)$, shown as Figure 2. Then the model adjusts parameters of $p_\theta(z|X)$ by minimize the KL divergence between $q_\phi(z|X,Y)$ and $p_\theta(z|X)$.

Meanwhile, we employed an unsupervised topic model to construct topic-aware commonsense, namely BTM [9], which is an efficient topic model specifically for short documents. Prior studies on responses generation only focused on one inherent attribute while our work generates specific emotion response but also can perform well in topic relevance. The introduce of external topic-aware commonsense can lead our model to associate external topic related words for the message. For example, there is a message: ”I like playing basketball”. Our model can associate many topic-related words such as, ’NBA’, ’Kobe’, ’referee’, ’dribble’, ’coach’ etc. Besides, we fine-tune the BERT [14] model for training the classifiers to evaluate model performance. The BERT is a new method of pre-training language representations which obtains state-of-the-art results on a wide array of Natural Language Processing (NLP) tasks. Thus, the experimental evaluation results are credible.

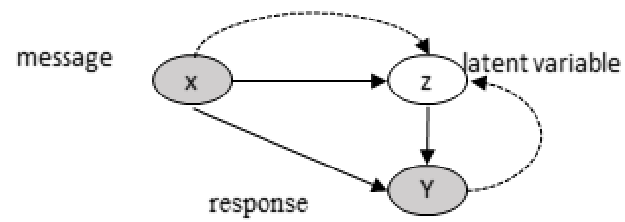

Fig. 2. A simple directed graph is used to illustrate the inference and generation process of CVAE. Dashed lines represent the inference of z. Solid lines represent the generation process.

III. Our TERG Model

A. Task Definition and Model Overview

Our problem is formulated as follows: given a message $X = x_1, x_2, ..., x_n$ and a specified emotion label el, the goal is to generate a response $Y = y_1, y_2, ..., y_m$ that not only match the emotion category el but also is topic-related with the message. Essentially, the generation model aim to maximize the generation probability of Y conditioned on X and el.

$$P(Y|z,X,el,v) = \prod_{t=1}^{m} p(y_t|y_{<t},z,X,el,v) \tag{1}$$

where the variable v is denoted as topic commonsense related to the message X.

The model architecture of our Topic-aware Emotional Response Generation model (TERG) is presented in Figure 3. In the encoding stage, encoder transforms the message and response(which solely used in the training process) into hidden representations. The Q and P nets are two networks to draw latent variable samples during training and test respectively [15]. The implementation principle of Q(P) networks adopted from the CVAE framework [13]. The latent variable z also captures specific emotional information by an Emotion Supervisor. To improve the topic relevance of generated responses, we propose to use the TCA module, which obtained external topic commonsense and then integrated into generation through the form of attention fusion. In decoding, the latent variable, the fused attention and emotion label are used as input features to update the decoder hidden state. Additionally, we introduce a word type selector to explicitly affect word distribution by obtaining word type distribution in each decoding position.

B. Encoder

We adopt the bidirectional gated recurrent unit (GRU) [16] as the encoder to transform a message and a response $X = x_1, x_2, ..., x_n$, $Y = y_1, y_2, ..., y_m$ into their respectively vector

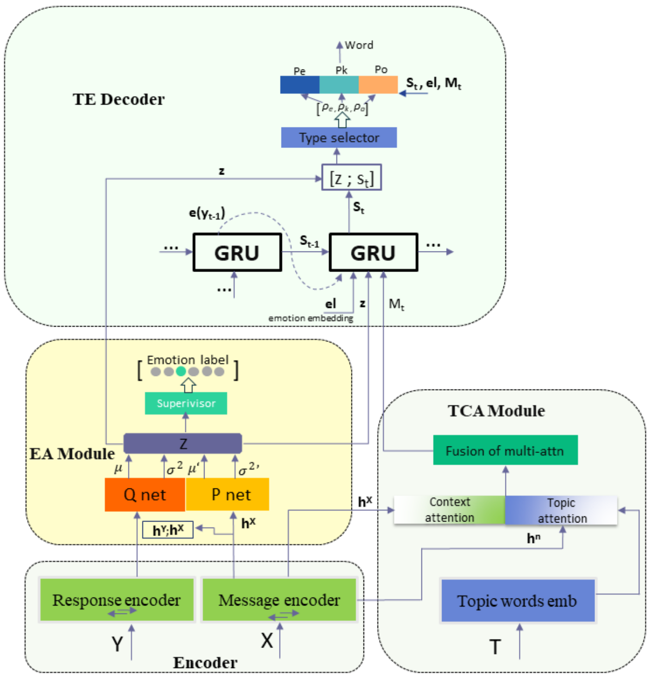

Fig. 3. The architecture of TERG which consists of four parts: Encoder, Emotion-aware (EA) module, Topic commonsense-aware (TCA) module and Topic-aware Emotion (TE) Decoder. The model graph is the t-th step state in decoding stage, where h^X is all hidden states in encoder, h^n is final step hidden state in Bi-GRU encoder. Besides, the response encoder only exists during the training process.

representation. Formally, the hidden states of the encoder can be computed as follows:

$$\overrightarrow{h_t} = GRU_{\mathrm{f}}\left(\tilde{x_t}, \overrightarrow{h_{t-1}}\right); \overleftarrow{h_t} = GRU_{\mathrm{b}}\left(\tilde{x_t}, \overleftarrow{h_{t+1}}\right) \tag{2}$$

where $\tilde{x_t}$ is the embedding of word x_t. In this paper, we represent lower case letters with wavy lines as words embedding. $\overrightarrow{h_t}$ and $\overleftarrow{h_t}$ are the j-th hidden states of forward and backward GRU respectively. The hidden h_t is the concatenation of the two hidden states, denoted as $h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}]$. The response encoder is similar to the message encoder. Significantly, the response encoder is only used in training.

C. Topic Commonsense-Aware Module

The human's conversations are usually under a particular topic during a period of time. After receiving a message, the topic commonsense is essential for continuing the conversation. In this paper, the topic commonsense is a set of topic words. We dynamically construct a specific topic-related lexicon for each message. Specifically, we employ the bi-term topic model (BTM) [9] (an efficient topic model specifically for short texts) to obtain topic-related words. BTM models the generation procedure of bi-terms in a short text collection, evolving from the LDA model. Here we omit the exhaustive background description of BTM because the topic model is not the main point of this paper. The procedure of topic words selection is as follows. Firstly, the BTM will assign the most related topic RT for the current message. Then we will pick the top N topic words with the highest probability under topic RT. Besides, in consideration of the noise of the topic model, we also apply keyword extraction algorithms like TextRank and named entity recognition (NER) tools to obtain keywords of the message as a part of topic words set. Finally, we obtain the topic words set: $T = t_1, t_2, ..., t_l$

The TCA module includes multi-attention and the fusion of multi-attention. Following [17], multi-attention is the concatenation of context attention and topic attention. The calculation of context and topic attention can refer to Equation 3-5.

$$C_t = \sum_{j=1}^{n} \alpha_{tj} h_j; TC_t = \sum_{j=1}^{l} \beta_{tj} \tilde{t_j} \tag{3}$$

where α_{tj} measures the semantic relevance between state s_{t-1} and hidden state h_j, β_{tj} denotes the weight between hidden state s_{t-1} and the j-th topic word in T, which are given by:

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{k=1}^{n} exp(e_{tk})}; e_{tj} = \eta(s_{t-1}, h_j) \tag{4}$$

$$\beta_{tj} = \frac{exp(w_{tj})}{\sum_{k=1}^{N} exp(w_{tk})}; w_{tj} = g(s_{t-1}, h^n, \tilde{t_j}); \tag{5}$$

where η and g are deep neural networks such as multiple layer perceptions (MLPs). In order to extract the most relevant feature of the external topic words lexicon, we use the previous hidden state s_{t-1} of the decoder, the last message encoder hidden state h^n, and the embedding of j-th topic word as the input of g to get the weight score.

The fusion of multi-attention makes the topic commonsense to be more naturally integrated into the generation process. The concatenation ($[C_t; TC_t]$) between context attention and topic attention is input into another deep neural network g^* to get the final attention M_t.

$$M_t = g^*(C_t; TC_t) \tag{6}$$

where $\mathrm{C_t} \in \mathbb{R}^{1 \times d}$, $\mathrm{TC_t} \in \mathbb{R}^{1 \times d}$, and the final attention $\mathrm{M_t} \in \mathbb{R}^{1 \times d}$. Through such a fusion mechanism, the module can weaken the noise effect of topic words that are irrelevant to the message in generation and can seamlessly plug proper topic words into the generated texts at the right time steps.

D. Emotion-Aware Module

The emotion-aware module consists of Q(P) networks and an emotion supervisor. Following [15], we use two networks to draw latent variable samples during training and test respectively [15]. The implementation principle adopted from CVAE [13] framework.

a) Q(P) Networks: According to the theory of CVAE, we should have sampled the latent variable from the true posterior distribution $P(z|Y, X)$, but the posterior distribution is intractable. Therefore, we input the messages $[X...]$ and responses $[Y...]$ into the Q network to get an approximate posterior distribution $q_\phi(z|Y, X)$ during training process. Besides, we input the z into an emotional supervisor to predict the emotion label of response. After the training of large samples, the latent variable z can map different kinds of emotional responses into different regions in a latent space.

In specific practice, we assume that latent variable z follows the Gaussian distribution whose covariance matrix is diagonal. During training, we construct the Q network to output the pivotal parameter μ and σ^2 of the approximate posterior distribution $q_\phi(z|Y,X)$ and then sample latent variable z. The Q network is a multiple layer perception (MLP):

$$MLP_{q_\phi}([Y;X]) \Longrightarrow [\mu;\sigma^2]; q_\phi(z|Y,X) \sim \mathcal{N}(\mathbf{z};\boldsymbol{\mu},\boldsymbol{\sigma}^2 I) \tag{7}$$

However, during prediction process, there is no encoding feature of the response Y. Therefore, we adopt another MLP, namely P network, to approximate the true prior distribution, which is implemented in the same way:

$$MLP_{p_\theta}(X) \Longrightarrow [\mu^{'};\sigma^{'2}]; p_\theta(z|X) \sim \mathcal{N}(\mu^{'},\sigma^{'2} I) \tag{8}$$

Our model is trained to minimize the KL divergence between the prior and posterior distribution so that our model can approximate the posterior distribution accurately using the prior distribution. The lower KL loss, the closer distance between the two distributions, which is defined as :

$$\begin{aligned} D_{kl} &= KL[q_\phi(z|Y,X)||p_\theta(z|X)] \\ &= \sum_{i=1}^{N_z} q_\phi(z=z_i|X,Y)\log\frac{q_\phi(z=z_i|X,Y)}{p_\theta(z=z_i|X)} \end{aligned}$$

where N_z is the dimension of latent variable z, θ and ϕ denote the model parameters. Then, during the inference process, the model samples a latent variable z merely based on the prior distribution.

b) Emotion Supervisor: Furthermore, there is an emotion supervisor that guides the latent variable to encode emotional information in the response with emotion label. Following [15], the supervisor takes z as input and then predicts the emotion label:

$$P(el|z) = softmax(W_e * f(z)); f(z) = tanh(Mz+b) \tag{9}$$

where el is the emotion label, latent variable z is a k dimensional vector, $\mathrm{M} \in \mathbb{R}^{d\times k}$, $\mathrm{W_e} \in \mathbb{R}^{c\times d}$ is the trainable transformation matrix and $\mathrm{b} \in \mathbb{R}^{d\times 1}$, c is the number of emotion categories. The loss function of the Emotion Supervisor is defined as:

$$loss_{es} = -\sum_{e=1}^{c} p_e * log(P(el|z)) \tag{10}$$

where p_e is a one hot vector of emotion label.

E. Topic-aware Emotional Decoder

The topic-aware emotional decoder differs from the vanilla decoder in that it takes in topic and emotion feature in decoding. In this work, we utilize a one-layer uni-directional GRU as decoder. For each time step, the output token of previous time step $\tilde{y}_{t-1}$, the latent variable z, emotion label el and the output of TCA module M_t are passed through the GRU to update its hidden state of current time step s_t:

$$s_t = \boldsymbol{G}RU(M_t, el, \tilde{y}_{t-1}, z, s_{t-1}) \tag{11}$$

We divide the words in the vocabulary into three types. The keywords are crucial for expressing core meaning. The emotional words have strong emotional polarity. And the ordinary words play a role which connects the emotion and content words to make a natural and grammatical sentence. Following [15], we use the latent variable z and the hidden state s_t to estimate the distribution over word types at each decoding step which is used to explicitly control the emotional sentence generation. The formula is as follows:

$$\rho_{e,k,o} = softmax(W_{eko} * tanh(W_{sz}[s_t;z]+b_{sz})) \tag{12}$$

where $\rho_{e,k,o} \in \mathbb{R}^3$, it can also be viewed as weights of choosing different types. We define the final generation probability as follows:

$$y_t \sim P(y_t) = \begin{bmatrix} \rho_e * P_{et}(y_t = w^e) \\ \rho_k * P_{kt}(y_t = w^k) \\ \rho_o * P_{ot}(y_t = w^o) \end{bmatrix} \tag{13}$$

where P_{et}, P_{kt} and P_{ot} are defined as the probabilities of selecting emotional words, keywords and ordinary words respectively. The probabilities of choosing words in different types are defined as:

$$P(y_{et}) = softmax(W_e * [s_t; z; el]) \tag{14}$$

$$P(y_{kt}) = softmax(W_k * [s_t; M_t]) \tag{15}$$

$$P(y_{ot}) = softmax(W_o * s_t) \tag{16}$$

As for the probability of emotion words, $P(y_{et})$ depends on the hidden state s_t, the latent variable z and emotion label el. After considering these factors comprehensively, the decoder can generate proper emotional words related to the specific emotion label. The probability $P(y_{kt})$ of selecting a keyword depends on hidden state s_t and the output of TCA module M_t. For the probability $P(y_{ot})$, we just consider the hidden state. The generation loss is based on cross-entropy:

$$loss_g = -\sum_{t=1}^{m}\log\left(P\left(y_t|y_{<t}, z, X, el\right)\right) \tag{17}$$

The overall training object function include three parts: KL divergence term D_{kl}, classification loss of the Emotion Supervisor $loss_{es}$ and generation loss of decoder $loss_g$ shown as:

$$loss = loss_{es} + loss_g + \alpha * D_{kl} \tag{18}$$

where α is set to gradually increase from 0 to 1. Following the KL cost annealing [18], we add a variable weight α to the KL loss term in the loss function at training time, which can mitigate the issue of vanishing latent variables.

IV. Experiments

A. Datasets

We used the Chinese dialogue dataset which contains 1,120,838 message-response pairs from Weibo[1]. But the dataset has no emotion labels. Thus, following [7], we built an emotion classifier to automatically annotate the emotion label for the dialogue corpus. To train the emotion classifier, we collected corpus form NLPCC2013 [2] and NLPCC2014 [3], filtered and then reserved 23,105 sentences with the manually emotion label. There are six emotion categories, including happy, disgust, sad, angry, and null, where the null label means that there is no any emotional polarity. We divided the NLPCC dataset into training, validation, and test set in a ratio of 8: 1: 1. We trained three classifiers including: Bi-LSTM [19], Self-attention [20] and BERT-based [14]. The test results are shown on Table I.

TABLE I
ACCURACY OF EMOTION CLASSIFIERS.

.

Model	Accuracy
Bi-LSTM	0.616
Self-attention	0.662
BERT-based	**0.739**

Finally, we adopted the BERT-based classifier to annotate the emotion label for responses. The basic statistics and distributions of the dialogue dataset is shown in Table II.

TABLE II
THE STATISTICS OF DATASET

Type	The number of sentence pairs	The emotion distribution
Training	1,097,010	like:14.07% ; null:23.26%;
Validation	11,194	sad:11.23% ; disgust:18.50%
Test	11,194	happy:24.56% ; angry:8.37%

B. Experiment Setting

We use single layer GRU with 256 cells as the encoder and decoder. We apply an existing word vectors file[4] to construct the topic words embedding table and the initial embedding of words. The vocabulary size is 40,000 and the batch size was set to 128. In the entire vocabulary, there are 5768 emotional words, 10,000 keywords, and 24,232 ordinary words. We collected the emotional words based on the the existing emotional dictionary[5]. The keywords were obtained from the dialogue corpus by the tool of keyword extraction such as TextRank. We adopted the Stochastic Gradient Descent algorithm to optimize our model and we set the learning rate to 0.1. The dimension of the latent variable z is 128. In addition, the code of BTM is available at this website[6]. We collected 1,000,000 message-response pairs from STC dataset [21] as the corpus to train the topic model. Each message-response pair is regarded as a short document. We run Gibbs sampling with 1,000 iterations to ensure that the BTM can reach a state of convergence and set the parameters of the topic number $K = 81$, hyperparameters $\alpha = 0.05$, $\beta = 0.01$.

C. Baselines

We regarded the following modules as baselines which were implemented with the settings provided in the original papers and the same dataset with our model.

- **Seq2Seq [6]**: This model is a standard dialogue generation model that evolved from Neural Machine Translation. The Seq2Seq learning framework with recurrent neural networks (RNNs) has been successfully used to build chatbots.
- **ECM [7]**: It's the first work that proposes to address the emotion factor in large-scale conversation generation. This model has three emotion mechanisms: Emotion embedding, Internal Memory, External Memory. The code has been released by [7].
- **ERG**: We also build an Emotion Response Generation model without the topic commonsense aware module. We can better analyse TCA module's importance for topic relevance from the experimental results.

V. Evaluation

A. Automatic Evaluation:

The following metrics are used to automatically evaluate the generated responses and model performance: The **BLEU** score is used to approximate the overlap between generated responses and target responses. We adopt the **perplexity** [22] to evaluate whether the generated responses are fluent and grammatical. Three **embedding-based metrics** including average, greedy and extrema [23] which are used to evaluate the semantic similarity between the generated responses and the targets. Besides these generic metrics, we got the **emotion accuracy** of generated responses with the help of the BERT-based emotion classifier. The emotion accuracy is calculated as follows:

$$acc_e = \frac{n_m}{n_a}; \tag{19}$$

where n_m is the matched number of predicted emotion labels and expected labels, n_a is the total number of test samples.

In terms of **topic relevance**, we trained a classifier that judges whether two sentences are topic related. We collected another one million message-response pairs from Weibo. The original message-response pairs were annotated positive samples and mismatched dialogue pairs were regarded as negative samples. We train the classifier by fine-tuning the BERT [14] model. The accuracy of the topic relevance classifier is **0.89**. Then the topic classifier is used to determine whether it is topic-related between the message and generated response. The formula of topic relevance score is shown as :

$$tr_{score} = \frac{n_p}{n_a}; \tag{20}$$

[1] https://weibo.com/ (A Chinese social platform)
[2] http://tcci.ccf.org.cn/conference/2013/
[3] http://tcci.ccf.org.cn/conference/2014/
[4] https://github.com/Embedding/Chinese-Word-Vectors
[5] https://github.com/ZaneMuir/DLUT-Emotionontology
[6] https://github.com/xiaohuiyan/BTM

TABLE III
RESULTS OF THE AUTOMATIC EVALUATION

	BLEU				Fluency	Relevance			Emotion	Topic
Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	perplexity	Emb Average	Externa	Greedy	emotion acc	topic relevance
Seq2Seq	0.0893	0.0206	0.0035	0.00169	83.89	0.531	0.360	0.396	-	0.2713
ECM	0.0912	0.0196	0.0040	0.00189	75.58	0.711	0.470	0.580	0.6511	0.3611
ERG	0.0937	0.0196	0.0061	0.00246	65.84	0.733	0.485	0.598	**0.7210**	0.4857
TERG	**0.0999**	**0.0234**	**0.0067**	**0.00320**	**63.95**	**0.786**	**0.588**	**0.638**	0.7109	**0.6083**

where n_p represents the number of positive labels predicted by the topic classifier, n_a is the total number of test samples.

B. Human Evaluation

We recruited six volunteers who are well-educated native speakers of Chinese to score the test results of our TERG and baselines. We randomly sampled 200 messages and generated responses in the test set. Following [24], we designed two evaluation strategies. **Pointwise evaluation**: Three volunteers rated the generated responses from the perspective of fluency, topic relevance and emotion expression accuracy. A graded assessment scale was used to score the generated responses, where 0=very terrible, 1= bad, 2=borderline, 3=not bad, 4= good, 5=surprised. **Pairwise evaluation**: The remained three volunteers evaluated whether the responses generated by our model are better than the baselines, where 1=better, 0=equal, -1=worse. If they could not understand both replies, they were asked to choose "equal". The source of generated responses is blind for volunteers and the final scores are average scores. By this way, we can comprehensively evaluate the results generated by different models.

C. Evaluation Results and Analysis

Our model shows substantial improvements against baseline methods in terms of perplexity, bleu score and manual evaluation. Table III report evaluation results on automatic metrics. The lower perplexity indicates that our model has the ability to generate more fluency responses and the bleu score of our model is much higher than the ECM and Seq2Seq, which indicates responses generated by our model are closer to the ground truth. Since dialogue generation is an open-ended problem, scores in the tasks are typically much lower than those observed in machine translation. In terms of semantic and topic relevance, our model yielded a significant performance boost. As we can see, after removing the TCA module, the topic relevance score decreased significantly. The results verify that introduction of TCA model is particularly useful in generation topic-related responses. In addition, the emotion accuracy of ERG is a little higher than the TERG. The reason of the slightly lowness on the emotion accuracy may be that the addition of our external topic commonsense. So the model is slightly biased towards the capture of topic information. From the overall results, the TERG model is better than the ERG.

Human evaluation results are shown in Table IV and V. The pointwise evaluation results show the TERG model yields the best score in all metrics. Agreements to measure inter-rater consistency among three annotators were calculated with the Fleiss's kappa [25]. The Fleiss's kappa for fluency, topic relevance and emotion accuracy is respectively 0.46,0.41,0.49, showing moderate annotator agreement. In the pairwise annotation protocol, the scores larger than 0 indicates our model outperforms its competitors.

TABLE IV
THE POINTWISE HUMAN EVALUATION RESULTS

Model	Fluency	Topic relevance	Emotion accuracy
Seq2Seq	2.8432	2.0522	-
ECM	2.6757	2.0943	2.5277
ERG	2.9772	2.3030	3.0025
TERG	**3.1919**	**3.3863**	**3.0833**

TABLE V
THE PAIRWISE HUMAN EVALUATION RESULTS

Model	score
TERG vs Seq2Seq	0.5327
TERG vs ECM	0.4826
TERG vs ERG	0.2587

D. Topic Relevance Analysis

In this section, we will further analyze the topic relevance between the real dialogue corpus and generated dialogues using the unsupervised topic model BTM. Specifically, we randomly selected 5,000 message-response pairs from the test set and took the same messages as inputs to generate the responses by baseline and our models. Then we adopted a statistical algorithm which is shown in Algorithm 1 to calculate the topic relevance score. From the line chart of results in Figure 4, we can intuitively find that the topic relevance of the real-life dialogue corpus is much higher than the dialogue generated by the baseline models. And the scores of our model are very close to or even higher than the scores of real conversations. The reason why the score is higher than the real dialogue 's is that the corpus is collected from Weibo rather than the real-word dialogue. Weibo users do not always use standard grammar or spellings, and frequently use colloquial language. Thus, it's significant to introduce the topic commonsense knowledge for a dialogue system. In this work, the topic commonsense is in the form of a series of topic words that are closely related to conversation. Some specific examples of topic commonsense are shown in Figure 5.

Algorithm 1 Topic relevance score calculation

Given: Topic model T
Input: Test pairs

range n [1:15]
for each message-response pair **do**
  T assign topic distribution for message and response
  sort and select top n topics M for message
  sort and select top n topics R for response
  count=0
  if $M \bigcap R \neq \varnothing$ **then**
    count++;
score=count / size of test pairs
return score

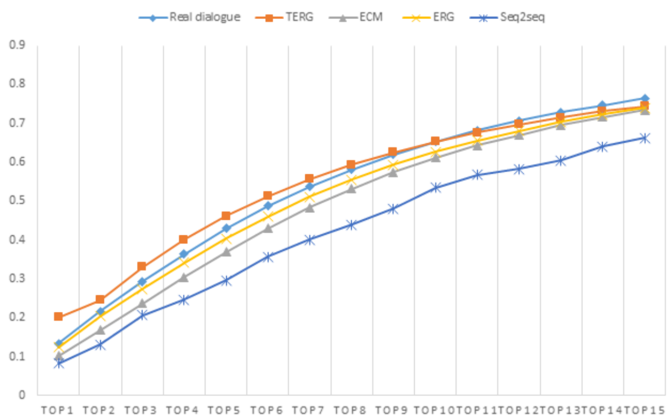

Fig. 4. The column chart of dialogue's topic relevance score .

E. Case Study

The presented test results' examples in Figure 6 show that our model can generate more informative responses which are related to the given messages. In the first example, to generate a response with 'like' emotion, TERG no longer uses the word 'like', to our surprise, it uses the word 'darling'. Furthermore, the generated responses by our model are all related to the topic of eggs. In the second example, our dialogue system can get the words "monsters, Sailor Moon" in the generated sentences which are related to the entity word "Ultraman" due to the addition of the TCA module.

VI. Conclusion and Future Work

In this paper, we have constructed a novel response generation model for chatbots by introducing the latent variable and fusion of multi-attention. Our model shows substantial improvements against several baseline methods in both automatic and manually evaluation. Our work has important implications for the design of chatbots. An excellent chatbot should be able to perceive, understand, and express different emotions like a human. It is a crucial step to generate emotional responses in the process. The purpose of chatbots is to have more rounds of conversations with their users and to accomplish some ambiguous goals, such as: filling the user's boring time, emotional companionship, and encouraging users. It's a useful step to introducing external topic common sense in response generation.

Message	Topic commonsense
我的摄影技术很棒哦! My photography skills are great!	**技能，摄影， 摄影师，录音**， **照片，** 呵呵， 不错， 老师， 谢谢， **手机， 效果， 拍照**， 漂亮， **技术，** 感觉， 好看， 喜欢， **相机，** 可惜， **专业， 视频，电视** **skill, photography, photographer, recording, photo,** hehe, good, teacher, thank you ,**mobile phone, effect, photo,** beautiful, **technology,** good looking , like, **camera,** pity, **professional**, **video**, **TV**
看来 博主就是钢琴高手。 It seems that the blogger is a piano master.	**高手** 小时候 **风琴 管风琴** 妙手 **博主 学琴 钢琴 巨匠** 喜欢 **贝多芬** 老师 好听 不错 学校 大学 **音乐** 学习 **专业** 快乐 **声音 艺术** **master**, childhood, **organ, pipe organ**, master, **blogger, piano, piano master**, like, **Beethoven**, good teacher, good school, university, **music**, learning, **professional**, happy, **sound**, **art**

Fig. 5. Examples of message & topic commonsense

Although our dialogue system performs well in the many evaluation metrics, there are still some problems to solve in the future. The dialogue system was trained on the corpus from the Weibo platform. There is a lot of noise data in the conversational corpus, which leads to the weak logical relationship between the generated responses and messages. In the future, we will explore how to enhance the logical association among the conversations. How to model the emotion and topic attributes to capture their transformations in the multiple rounds dialogue system.

VII. Acknowledgment

This research is funded by the Science and Technology Commission of Shanghai Municipality (No. 18511105502) and Xiaoi Research.

References

[1] H. Prendinger and M. Ishizuka, "The empathic companion: A character-based interface that addresses users'affective states," *Applied Artificial Intelligence*, vol. 19, no. 3-4, pp. 267–285, 2005.
[2] H. Prendinger, H. Dohi, H. Wang, S. Mayer, and M. Ishizuka, "Empathic embodied interfaces: Addressing users' affective state," in *Tutorial and Research Workshop on Affective Dialogue Systems*. Springer, 2004, pp. 53–64.
[3] T. S. Polzin and A. Waibel, "Emotion-sensitive human-computer interfaces," in *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.
[4] M. Skowron, "Affect listeners: Acquisition of affective states by means of conversational systems," in *Development of Multimodal Interfaces: Active Listening and Synchrony*. Springer, 2010, pp. 169–181.
[5] H. Wang, Z. Lu, H. Li, and E. Chen, "A dataset for research on short-text conversations," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 2013, pp. 935–945.
[6] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.
[7] H. Zhou, M. Huang, T. Zhang, X. Zhu, and B. Liu, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
[8] N. Asghar, P. Poupart, J. Hoey, X. Jiang, and L. Mou, "Affective neural response generation," in *European Conference on Information Retrieval*. Springer, 2018, pp. 154–166.

Message	Model	E-label	Generated Responses
很营养！鸡蛋煎得不错。 Very nutritious! Eggs are fried well .	Seq2seq	\	哈哈，谢谢，谢谢。[Haha，thanks !]
	ECM	null	是啊，是我的菜。[Yeah, it's my dishes.]
		like	是啊，我喜欢的菜。[Yeah, I like the dish.]
		sad	我还不知道我是什么东西。[I still don't know what I am.]
		disgust	那是我的菜，好吃营养。[That is my dish, delicious and nutrition.]
		angry	是吗？是我的菜？[Really? Is it my dish?]
		happy	哈哈，我的手艺。[Haha, it's my craft.]
	ERG	null	是的，我做的。[Yeah，I made it.]
		like	我喜欢吃的，味道不错。[I like to eat, it tastes good.]
		sad	可惜没吃到，我不知道吃什么。[Unfortunately, I didn't eat it. I don't know what to eat.]
		disgust	你吃的太浪费了。[You are wasting too much.]
		angry	你怎么不说这是我的？[Why didn't you say it is mine?]
		happy	哈哈，我也是。[Haha, me too.]
	TERG	null	是的， 鸡蛋煎的是兴趣， 是一个女孩的品味 。[Yeah, frying eggs are an interest,which can also reflect the taste of a girl]
		like	亲爱的,你是想吃鸡蛋吗？ [Dariling, do you want to eat eggs?]
		sad	惭愧,我是想吃煎鸡蛋,可不敢吃了。[I want to eat fried eggs, but I dare not eat them.]
		disgust	傻瓜，鸡蛋难吃。[Fool, eggs are tasted badly.]
		angry	我没耐心等你吃鸡蛋。[I have no patience waiting for you to eat eggs.]
		happy	哈哈，是啊，我想吃鸡蛋。[Haha, yeah, I want to eat eggs.]
我喜欢看奥特曼。 I like to watch a cartoon called Ultraman	TERG	null	我也想看，但是我觉得我还没看完。[I also want to watch it, but I haven't finished watching it yet.]
		like	我喜欢看奥特曼打怪兽。[I like to watch the section that Ultraman beats monsters.]
		sad	我也想看动画片，可是我没看过。[I also want to watch the cartoon too, but I have not seen it.]
		disgust	我靠，我还以为是美少女战士。[Damned, I had expected that it it was a Sailor Moon.]
		happy	哈哈，我想知道你是奥特曼吗？[Haha, I want to know whether you are an Ultraman?]

Fig. 6. Examples of responses generated by different models. The red marked words contain the emotional properties and the blue marked words are keywords which are used to express the core meaning of sentences.

[9] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proceedings of the 22nd international conference on World Wide Web*. ACM, 2013, pp. 1445–1456.

[10] R. Zhang, Z. Wang, and D. Mai, "Building emotional conversation systems using multi-task seq2seq learning," in *National CCF Conference on Natural Language Processing and Chinese Computing*. Springer, 2017, pp. 612–621.

[11] Z. Song, X. Zheng, L. Liu, M. Xu, and X.-J. Huang, "Generating responses with a specific emotion in dialog," in *Proceedings of the 57th Conference of the Association for Computational Linguistics*, 2019, pp. 3685–3695.

[12] P. Ekman, W. V. Friesen, and P. Ellsworth, *Emotion in the Human Face: Guide-lines for Research and an Integration of Findings: Guidelines for Research and an Integration of Findings*. Pergamon, 1972.

[13] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *Advances in neural information processing systems*, 2015, pp. 3483–3491.

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[15] P. Ke, J. Guan, M. Huang, and X. Zhu, "Generating informative responses with controlled sentence function," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 1499–1508.

[16] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[17] C. Xing, W. Wu, Y. Wu, J. Liu, Y. Huang, M. Zhou, and W.-Y. Ma, "Topic aware neural response generation," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[18] S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio, "Generating sentences from a continuous space," *Computer Science*, 2015.

[19] A. Graves, S. Fernández, and J. Schmidhuber, "Bidirectional lstm networks for improved phoneme classification and recognition," in *International Conference on Artificial Neural Networks*. Springer, 2005, pp. 799–804.

[20] Z. Lin, M. Feng, C. N. D. Santos, Y. Mo, and Y. Bengio, "A structured self-attentive sentence embedding," 2017.

[21] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[22] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[23] C.-W. Liu, R. Lowe, I. V. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," *arXiv preprint arXiv:1603.08023*, 2016.

[24] L. Mou, Y. Song, R. Yan, G. Li, L. Zhang, and Z. Jin, "Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation," *arXiv preprint arXiv:1607.00970*, 2016.

[25] J. L. Fleiss and J. Cohen, "The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability," *Educational and psychological measurement*, vol. 33, no. 3, pp. 613–619, 1973.