ECCRG: A Emotion- and Content-Controllable Response Generation Model

Hui Chen[1], Bo Wang[2(✉)], Ke Yang[1], and Yi Song[2]

[1] Hunan Seefore Information Technology Co., Ltd., Yueyang, China
yk@singhand.com
[2] College of Intelligence and Computing, Tianjin University, Tianjin, China
bo_wang@tju.edu.cn

Abstract. Most methods of emotional dialogue generation focus on how to make the generated replies express the set emotion categories, while ignoring the control over the semantic content of the replies. To this end, in this paper, we propose a emotion- and content-controllable response generation model, ECCRG. ECCRG allows for text-controlled conditions and integration into the decoding process of the language model through a self-attention layer, enabling more precise control over the content of the generated responses. We use a variety of optimization objectives including self-reconfiguration loss and adversarial learning loss to jointly train the model. Experimental results show that ECCRG can embody the set target content in the generated responses, allowing us to achieve controllability on both emotion and textual content.

Keywords: Dialogue systems · Emotional response generation · Controllable text generation

1 Introduction

Early research Partala and Surakka (2004) showed that dialogue systems capable of appropriate emotional expressions in replies can directly improve user satisfaction and make users feel more engaged. Ideally, a dialogue system with emotional intelligence can make the user experience more comfortable by means of emotional interaction, and even have the effect of psychological comfort and treatment. Some researchers have tried to make dialogue systems appear more human-like by making the system mimic human emotional expressions. Early representative work (Polzin and Waibel (2000); Skowron (2010)) used human-written rules to select responses related to specific emotions from a dialogue corpus. These rules usually need to be written by experienced experts, so such methods are difficult to scale to scenes and larger corpora containing complex, subtle emotions.

In the era of deep learning, sequence-to-sequence (Seq2seq) models have gradually been widely used in dialogue generation tasks. In early attempts to develop

H. Gao et al. (Eds.): CollaborateCom 2023, LNICST 562, pp. 115–130, 2024.
https://doi.org/10.1007/978-3-031-54528-3_7

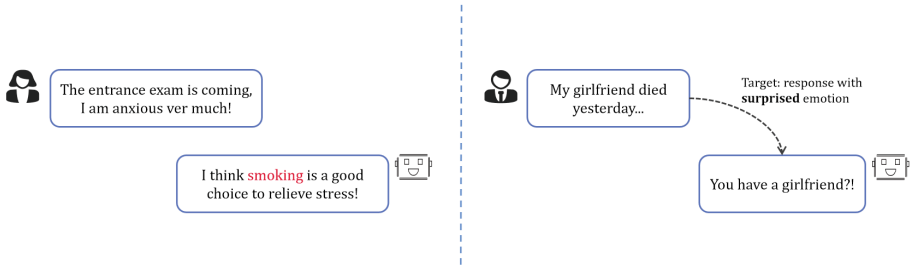

Fig. 1. A case of uncontrolled reply generation

chatbots through the Seq2seq framework, many efforts (Li et al. (2015); Gao et al. (2019)) were made to avoid boring words (like "I'm fine" and "go on") in responses, and works to make the responses more informative. But only a few works focus on the emotional expression problem in dialogue generation. Incorporating emotional expressions in a dialogue system mainly faces several challenges:

(1) Dialogue data with sufficient scale and marked emotion categories are difficult to obtain. Some work in recent years has contributed many high-quality dialogue datasets to the academic community, which are mostly manually written and annotated and limited in size. Because the task of emotion recognition in dialogue is also quite challenging, accurate emotion annotation is difficult to scale to large-scale datasets.
(2) It is difficult for us to consider emotions in a natural and coherent way. Emotional factors are only expressed on the surface of the text by emotional words in a few cases, but are often implied under the semantics of the text. Both abstraction and incorporating emotional information present difficulties.
(3) The experiments of Zhou et al. (2018) show that simply embedding emotional information into a neural network model cannot generate ideal emotional responses.

To address these challenges, Zhou et al. (2018) successfully constructed an emotion chat machine (ECM) based on the Seq2seq model, which is able to generate responses expressing that emotion based on pre-defined emotion categories. After this, there have been some studies with similar goals (Song et al. (2019); Huang et al. (2018); Asghar et al. (2018)). Zhou and Wang (2018) proposed MojiTalk, which uses a series of emoji "emoji" in the conditional variational The process of generating emotional responses is controlled within an encoder (CVAE) framework. Some studies have proposed ways to motivate models to express emotion more explicitly, for example, assigning additional probability distribution weights to related emotion words under a given emotion during decoding, or incorporating the use of emotion words into the loss function.

All of the methods mentioned above are effective in generating responses that express specific emotions. From the emotional level, most of them represent emotional categories as independent labels, such as "<happy>" or the corresponding

"emoji", and convert them into the form of embedding vectors and then integrate them into the basic generative model, or It is to integrate the emotion vector into the hidden state of the decoder to affect the decoding result, or use it as a conditional latent variable in the CVAE framework, and then splicing emotion words and general words in the decoding process to achieve the purpose of expressing emotion. This control method with emotion as a high-level attribute makes it difficult for the model to learn the specific way to express emotion from semantics. Meanwhile, the methods that try to express emotions explicitly by using a large number of emotional words are difficult to balance emotional words and other words, resulting in less fluent responses, and explicit expressions appear unnatural in many cases.

From the semantic level, related research work only focuses on the control of emotional expression, but ignores the quality of the generated content. Dialogue generation technology is widely used in more rigorous application scenarios including legal and political court trials, and medical dialogue for mental health. Taking counseling conversations in health care as an example, users may be troubled by various psychological barriers and frustrations, such as listening to the user's conversation about test anxiety, inappropriate counseling and guidance shown in the left part in Fig. 1 is likely to have serious consequences. For open-domain dialogue systems, most of the corpus comes from chats crawled by social media platforms. Chatbots trained by such data are likely to generate "toxic" sentences. For example, in the case shown on the right part in Fig. 1, even if the emotion type "surprise" that should be expressed in the reply is specified, such a reply will obviously reduce the user's experience.

To address the above problems, we attempts to jointly constrain response generation from both the emotional level and semantic level, and control the generated text content at a finer granularity. We propose an Emotion- and Content-Controllable Response Generation model (ECCRG). ECCRG is built on a large-scale pre-trained language model, which guarantees the basic fluency of the text it generates. We add an intermediate layer into the multi-layer Transformer structure of the language model, which can incorporate control conditions in the form of text into the downstream language model through a self-attention mechanism, and thereby guide the subsequent generated content.

2 Related Work

Humans have the unique ability to perceive complex, subtle emotions and to communicate their experiences and feelings with each other through language. Existing research Partala and Surakka (2004) suggested that dialogue systems with appropriate emotional expressions in responses can directly improve user satisfaction and help increase user engagement. However, making dialogue systems more "emotional" remains a huge challenge.

In early representative work, researchers used some hand-crafted rules to select sentences associated with specific emotions from a dialogue corpus. These rules need to be written by trained experts and are difficult to scale to handle more complex and subtle emotions in large-scale corpora. In 2014, Microsoft

launched Xiaoice (Zhou et al. (2020)), a social chatbot that recognizes users' emotional needs and has empathy. It was not until Zhou et al. (2018) proposed the Emotional Chatting Machine (ECM), which used deep learning methods to build an emotion-aware dialogue system on large-scale corpus, research related to emotional dialogue generation became popular.

After this, Colombo et al. (2019) improved ECM, they used VAD lexicons to represent emotion, and improved the decoding process and loss function for the emotion factor. Song et al. (2019) pointed out that the more general way to express emotions is to express them implicitly through semantics. They used a emotion classifier co-trained with the model to guide the process of response generation to ensure the appropriate expression of specific emotions in the responses. Asghar et al. (2018) proposed an emotion-diverse beam search algorithm for decoding, and employed reinforcement learning to encourage the model to present the specified emotion at generation time. Zhong et al. (2020a) constructed an emotion-aware commonsense concept graph based on ConceptNet using emotion-annotated corpus, and then they captured the most relevant knowledge tuples under different expected emotions through mechanisms such as graph attention, and fused into the Transformer model.

In the trend of pre-trained language models (Devlin et al. (2018); Liu et al. (2019); Radford et al. (2019)) sweeping the NLP field, the research on controllable dialogue generation has also made breakthroughs. Some studies reconstruct and retrain the pre-trained model so that the generated results meet certain preset conditions. Lin et al. (2021) built a series of lightweight adaptive models for various dialogue generation needs based on the pre-trained model DialoGPT(Zhang et al. (2019)), these models allow various control for different conversational needs Conditions (including emotion, language style, etc.) for advanced control and integration.

Most of the above-mentioned methods rely on specific emotion labels to express emotions, but do not explicitly model the emotional information in the context, so that the models does not really understand the current emotional state of the dialogue and the user's intents. Meanwhile, most approaches of emotional dialogue generation focus on how to make the model more accurately express the specified emotion type, while ignoring the control over the semantic content of the responses.

3 Methodology

3.1 Problem Setup

In this paper, our task is to generate response sentences that are coherent with the dialogue history and user input and satisfy the controlled conditions, given the dialogue context and the control conditions of the emotion or content that needs to be expressed in the responses to be generated. The specific formulation of the task is as follows: Given a dialogue context $X = \{x_1, ..., x_N\}$ that may contain N utterances, where $x_i = \{w_{i,1}, w_{i,2}, ..., w_{i,l_i}\}$, x_N is the user's last input, and a control conditional text c, The model needs to generate corresponding response $Y = \{y_1, y_2, ..., y_{l_y}\}$ based on these conditions. The goal of

the generation is to keep Y semantically coherent with the dialogue history and user input, and meet the requirements of controlled conditions. Where l_i, l_y, l_c are the lengths of x_i, Y, c respectively.

In the response generation model proposed in this paper, the control condition c includes target emotion type or textual content. For the former, we input the emotion label into the control condition, then the corresponding emotional expression should be made in Y. In the latter case, the generated response Y should contain semantics consistent with the target content.

3.2 Model Architecture

ECCRG is built on a large-scale pre-trained language model. This kind of language model has comprehensively learned language knowledge (such as semantics, syntax, grammar, etc.), commonsense knowledge, and specialized knowledge from a large-scale corpus in the pre-training stage. Specifically in the experiments of ECCRG, we use the pre-trained parameters of GPT-2 (Radford et al. (2019)) to initialize the language model, denoted as LM, which ensures that the text generated by ECCRG has basic fluency and diversity sex. GPT-2 generates text in an auto-regressive manner, which can be expressed as:

$$p(x_t, ..., x_l | x_1, ..., x_{t-1}) = \prod_{i=t}^{l} p(x_i | x_1, ..., x_{i-1}), \tag{1}$$

where $t-1$ is the length of the input sequence, and l is the maximum length of the sequence generated by the language model. Due to the auto-regressive generation manner of GPT-2, the words generated in the current step will be added to the end of the original input sequence to form a new sequence as the input sequence generated in the next step. For step t, that is, the first word x_t generated in the GPT-2 generation process, it is sampled from the probability distribution of the output of the language model $o_t = \mathrm{LM}(x_{:t-1})$. For the reply generation task, we take user utterances as the raw input sequence for GPT-2.

Figure 2 shows the overall architecture of ECCRG. The GPT-2 language model is composed of stacked multiple layers, each layer is a Transformer block. We add an intermediate layer based on the self-attention mechanism in the middle of the multi-layer structure, which separates the language model into two parts. We call the middle layer ECC-layer, the upstream language model is denoted as LM_A, and the downstream part is denoted as LM_B. LM_A extracts features from the embedded representation of the input sequence and outputs its current hidden state $h_{:t-1}$:

$$h_{:t-1} = \mathrm{LM}_A(x_{:t-1}). \tag{2}$$

LM_B takes $h_{:t-1}$ as input and outputs the predicted probability distribution o_t of the language model:

$$o_t = \mathrm{LM}_B(h_{:t-1}). \tag{3}$$

Equations (2) and (3) can be combined as:

$$o_t = \mathrm{LM}(x_{:t-1}) = \mathrm{LM}_B(\mathrm{LM}_A(x_{:t-1})) = \mathrm{LM}_B(h_{:t-1}). \tag{4}$$

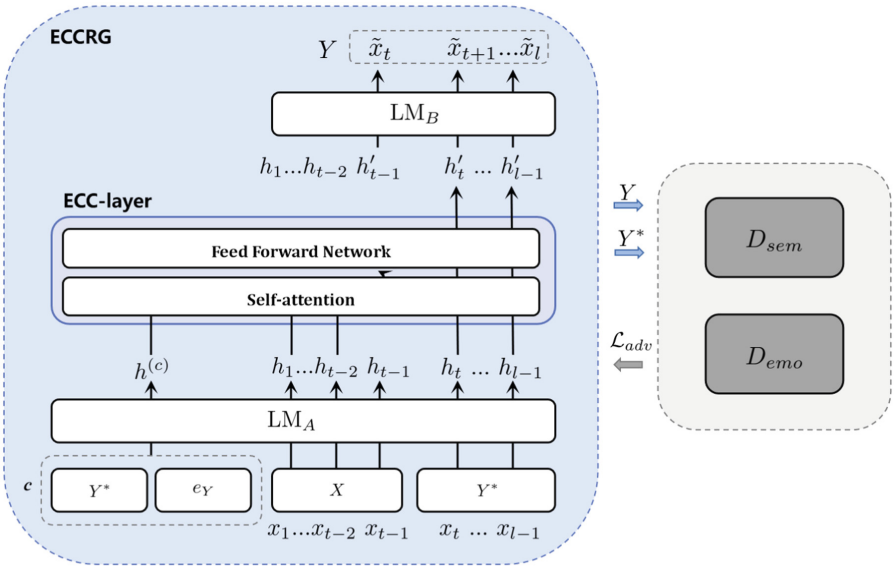

Fig. 2. Model structure of ECCRG

It can be seen from (4) that the probability distribution o of the next word generated by LM_B is affected by the hidden state h of LM_A. Using this effect, we integrate the controlled condition sequence c with the original hidden state h in ECC-layer to obtain a new hidden state h':

$$h'_{:t-1} = \mathrm{ECC}(h^{(c)}, h_{:t-1}), \tag{5}$$

where $h^{(c)}$ is the hidden state obtained by LM_A with the embedded representation of the control condition sequence as input, that is, $h^{(c)} = \mathrm{LM}_A(c)$.

Specifically, ECC-layer integrates $h^{(c)}$ and h through self-attention, first from $h_{:t-1}$ obtains $Q \in \mathbb{R}^{(t-1)\times d}$, $K \in \mathbb{R}^{(t-1)\times d}$ and $V \in \mathbb{R}^{(t-1)\times d}$ of the original hidden state through linear transformation, where d is the dimension of the hidden state. Similarly, from $h^{(c)}$ we can obtain $K^{(c)}, V^{(c)} \in \mathbb{R}^{l_c \times d}$, where l_c is the length of the control condition sequence, then calculate the output of self-attention $A \in \mathbb{R}^{(t-1)\times d}$:

$$\begin{aligned} K' &= [K^{(c)}; K] \in \mathbb{R}^{(l_c+t-1)\times d}, \\ V' &= [V^{(c)}; V] \in \mathbb{R}^{(l_c+t-1)\times d}, \\ A &= \mathrm{softmax}(QK'^{\mathrm{T}})V' \in \mathbb{R}^{(t-1)\times d}, \end{aligned} \tag{6}$$

where $A = \{a_1, a_2, ..., a_{t-1}\}$, a_{t-1} is converted by a fully connected layer to get $h'_{t-1} = \ mathrmFFN(a_{t-1})$, which is the new hidden state integrating control condition information obtained through ECC-layer at the current step. By concatenating h'_{t-1} with the original hidden state of the previous step and

sending them to LM_B, new probability distribution $\tilde{o}_t$ is conditioned on c:

$$\tilde{o}_t = \mathrm{LM}_B([h_{:t-2}; h'_{t-1}]), \tag{7}$$

The $\tilde{x}_t$ sampled by $\tilde{o}_t$ is the controlled generation result affected by the control condition c.

We fine-tune the pre-trained GPT-2 on the Reddit dataset. In the subsequent training process of ECC-layer, we fixed the parameters of the language model and only updated the parameters of ECC-layer.

3.3 Training of ECC-Layer

The training goal of ECCRG is to make the responses it generates controllable both at emotional level and content level.

At emotional level, although we incorporate emotion information into the input embedding of each token, it is difficult for the model to learn the difference in sampling space under different emotion types. Therefore, like other work on emotional response generation, we need to inject the target emotion label into the control condition of ECC-layer to explicitly guide the emotional expression in response generation.

At content level, due to the diversity of natural language and the randomness of sampling in the decoding process of auto-regressive models, even given sufficient dialogue context information, there may be a huge number of candidate responses that can maintain semantic coherence with the context. In other words, the probability of the model generating ground-truth is very low given only the dialogue history and no target response.

For these two problems, we design multiple loss functions as optimization targets for ECC-layer training.

Self-reconstruction Loss. In order to make the responses generated by the model close to the text content in the controlled condition, inspired by Chan et al. (2020), we adopt their proposed self-reconstruction loss to concatenate the context and the ground-truth response as the input sequence, and with the ground-truth response as the control condition.

Specifically, we concatenate the dialogue context X and the ground-truth response Y^* into a new input sequence $X' = [X; Y^*]$ and denote it as $X' = \{x_1, ..., x_{t-1}, x_t, , ..., x_l\}$, where l is the total length of the sequence, $x_{:t-1}$ is the original context, $x_{t:l}$ is the original response. Meanwhile, we use Y^* as control condition, denoted as $c = Y^*$. Same as Eq. (2), we first get the intermediate hidden states of X' and c from LM_A:

$$h_{:l} = \mathrm{LM}_A(x_{:l}),\ h_{l_c}^{(c)} = \mathrm{LM}_A(c) = \mathrm{LM}_A(x_{t:l}), \tag{8}$$

where $l_c = l - t + 1$ is the sequence length of c. Similar to Eq. (5), ECC-layer fuses $h_{l_c}^{(c)}$ with the original hidden state $h_{t:l}$ of the reply generation stage to obtain the conditioned hidden state:

$$h'_i = \mathrm{ECC}(h_{:l_c}^{(c)}; h_i),\ \text{where } i \in [t-1, l]. \tag{9}$$

Similar to Eq. (7), the original hidden states of the context is concatenated with the conditioned hidden states of the response transformed by ECC-layer, and passed into LM_B to produce the LM logits:

$$\tilde{o}_{i+1} = \mathrm{LM}_B([h_{:t-2}; h'_{t-1:i}]), p(\tilde{x}_{i+1}|x_{:i}, c) = \mathrm{softmax}(\tilde{o}_{i+1}), \quad \text{where } i \in [t-1, l]. \tag{10}$$

The self-reconstruction loss $\mathcal{L}_{recon}$ is defined as the cross-entropy loss between the generated and the ground-truth response:

$$\mathcal{L}_{recon} = -\sum_{i=t}^{l} \log p(x_i|\{x_1, x_2, ..., x_{i-1}\}, (c = x_{t:l})). \tag{11}$$

Emotion Controlled Loss. In order to make the decoding process adapted by the model under the condition of the specified emotion type still generate syntactically and syntactically fluent responses, we inject the emotion type corresponding to the real response into the ECC layer as a controlled condition. For the emotion category e_{Y^*} marked by the real reply Y^*, we refer to the conceptual interpretation in English and simply convert it into a short sentence describing the emotion, for example, for the emotion type "admiration", describe it as "I feel a state of admiration and pride.", still denoted as $c = e_{Y^*}$. Similar to the self-reconstruction loss, we define the emotion controlled loss $\mathcal{L}_{emo}$ as:

$$\mathcal{L}_{recon} = -\sum_{i=t}^{l} \log p(x_i|\{x_1, x_2, ..., x_{i-1}\}, (c = e_{Y^*})). \tag{12}$$

It should be noted that in the training and testing stages, when the self-reconstruction loss and the emotion controlled loss are activated at the same time, we concatenate the corresponding control conditions to the self-reconstruction loss and the emotion-controlled loss and inject it into ECC-layer, namely $c = [Y^*; e_{Y^*}]$, to evaluate the performance of ECCRG under the condition on emotion and content variables.

Adversarial Loss. In order to motivate the model to generate emotional expressions in replies and make replies closer to real text, we introduce two discriminator models to provide additional signals for the training of ECC-layers, namely the semantic discriminator D_{sem} and the emotion discriminator D_{emo}. Both discriminators are classifier structures constructed based on convolutional neural networks.

The semantic discriminator measures the semantic distance between the generated responses Y and the ground-truth Y^*. In order to avoid the discriminator loss cannot be back-propagated to ECC-layer caused by discrete sampling in the decoding process, We do not use the sampled token id sequence, instead send the probability distribution obtained from LM_B to LM_A to obtain the embedded representation of the hidden layer. Based on this we define the semantic adversarial loss $\mathcal{L}_{adv}^{sem}$:

$$\mathcal{L}_{adv}^{sem} = \log \mathrm{D}_{sem}(\mathrm{LM}_A(Y^*)) + \log(1 - \mathrm{D}_{sem}(\mathrm{LM}_A(Y))). \tag{13}$$

The emotion discriminator determines whether the generated response expresses the specified emotion type. We also define the emotion-based adversarial loss $\mathcal{L}_{adv}^{emo}$:

$$\mathcal{L}_{adv}^{emo} = \log p_{\theta_{\mathrm{D}_{emo}}}(e_{Y^*}|\mathrm{LM}_A(Y^*)) - \log p_{\theta_{\mathrm{D}_{emo}}}(e_{Y^*}|\mathrm{LM}_A(Y))), \tag{14}$$

Table 1. Statistics for the Reddit dataset

Dataset	#Dial.	Emotion label Statistics of Responses						
		Admiration	Approval	Caring	Joy	Sadness	Surprise	Neutral
Train	146,451	67,738	10,166	7544	18,919	6887	21,033	14,164
Valid	18,961	8904	1347	993	2321	886	2654	1856
Test	20,137	9606	1320	1056	2540	815	2898	1902

where $\theta_{\mathrm{D}_{emo}}$ is the parameter of D_{emo}.

In adversarial learning, the training objective of ECC-layer is to minimize the adversarial losses $\mathcal{L}_{adv}^{sem}$ and $\mathcal{L}_{adv}^{emo}$, while the training objectives of the semantic discriminator and the emotion discriminator are to maximize $\mathcal{L}_{adv}^{sem}$ and $\mathcal{L}_{adv}^{emo}$ respectively. ECCRG is trained in an end-to-end manner with two discriminators.

Training. ECCRG still needs to perform basic auto-regressive generation without setting any control variables. We define the auto-regressive loss $\mathcal{L}_{ar}$:

$$\mathcal{L}_{ar} = -\sum_{i=t}^{l} \log p(x_i|\{x_1, ..., x_{i-1}\}). \tag{15}$$

The total optimization objective of training is to minimize the sum of all the above loss functions:

$$\mathcal{L} = \lambda_{ar}\mathcal{L}_{ar} + \lambda_{recon}\mathcal{L}_{recon} + \lambda_{emo}\mathcal{L}_{emo} + \lambda_{adv}^{sem}\mathcal{L}_{adv}^{sem} + \lambda_{adv}^{emo}\mathcal{L}_{adv}^{emo}, \tag{16}$$

where λ is the weight factor that balances the influence of each part of the losses.

4 Experimental Setup

4.1 Dataset

Zhong et al. (2020b) proposed a large-scale dialogue dataset scraped from the social forum Reddit. The dialogues in this dataset come from two partitions Happy and Offmychest where users communicate emotionally. This chapter uses the Reddit dataset to train and evaluate ECCRG, mainly considering that the response sentences in this dataset are shown to contain distinct emotional expressions and empathy. In addition, we use the emotion annotation on the dataset by Zheng et al. (2021), they use a BERT-based classifier to label 8 emotion types

and neutral for each utterance. The accuracy on the testset is 65.8%. To alleviate the problem of unbalance distribution of emotion types, we filter out dialogues containing utterances labeld as "anger" or "fear". The statistics of the dataset are shown in Table 1. For each dialogue, we keep the last two sentences as context and response, respectively. We set the maximum number of token per utterance to 30.

4.2 Experimental Settings

We initialize the language model in ECCRG using the model configuration and parameters of GPT2-medium. We take the first 6 layers of the language model as LM_A, and the last 18 layers as LM_B.

During training, we first fine-tune the GPT-2 on the trainset for 10 epochs. Then, the parameters of the language model are fixed, ECC-layer and the two discriminators are updated for 10 epochs with a batch size of 4. We use the Adam optimizer for the above training process, with the learning rate set to $5e-5$. For each part of the weight factor λ in the loss function, we set all of them to 1.0. To obtain the generated samples for ECCRG, we sample the next word from the probability distribution obtained by the downstream language model using kernel sampling algorithm with p set to 0.9.

4.3 Baselines

We compare ECCRG with the following methods in our experiments:

Seq2seq-emo. On the basis of the Seq2seq model, the emotion label of the target reply is encoded into an embedding vector, which is used as an additional input to the decoder.

ECM (Zhou et al. (2018)) is the first model to generate emotional responses on large-scale dialogue datasets. We implemented it based on code published by the open source community.

DialoGPT (Zhang et al. (2019)) is a large-scale dialogue generation model based on GPT-2, trained on more than 147M dialogues captured from Reddit. We fine-tune DialoGPT for 5 epochs with the Reddit dataset.

4.4 Automatic Metrics

(1) **Perplexity** is a gram-based method to measure the strengths and weaknesses of language probability models.
(2) **BLEU** (Papineni et al. (2002)) measures how close the model-generated text is to the ground-truth by how much the n-gram phrases overlap. Some studies have pointed out that BLEU is not suitable for evaluating dialogue generation tasks because its results are less correlated with human evaluations. In this paper, we evaluate the effect of controllable generation by calculating the distance between the generated responses and the ground-truths by BLEU-1 and BLEU-2.

(3) **Distinct** (Li et al. (2015)) evaluate the diversity of generated responses based on n-gram counts. We use Distinct metrics under uni-gram and bi-gram, denoted as Dist-1 and Dist-2, respectively.
(4) **Emotion Accuracy Rate (Emo-acc)**: We use a emotion classifier based on RoBERTa (Liu et al. (2019)) to evaluate the emotion accuracy the generated responses.

Table 2. Experiment results on Reddit dataset

Model	Content Similarity		Fluency	Diversity		Emotion
	BLEU-2 ↑	BLEU-4 ↑	PPL ↓	Dist-1 ↑	Dist-2 ↑	Emo-acc ↑
Seq2seq-emo	6.86	1.95	**55.7**	0.038	0.142	60.4
ECM	7.33	1.97	63.8	0.042	0.172	63.5
DialoGPT	12.73	3.23	58.7	**0.065**	**0.247**	54.1
ECCRG	19.26	6.44	65.8	0.053	0.216	**69.7**
w/o $\mathcal{L}_{recon}$	14.67	3.82	65.2	0.059	0.233	67.2
w/o $\mathcal{L}_{emo}$	16.44	5.72	68.6	0.052	0.210	64.1
w/o $\mathcal{L}_{adv}^{emo}$	19.53	6.61	64.9	0.049	0.208	65.9
w/o $\mathcal{L}_{adv}^{sem}$	**19.77**	**6.80**	64.4	0.050	0.210	68.6

4.5 Human Evaluation Metrics

We recruited 5 volunteers with good English language skill to manually evaluate the dialogues generated by the model. We sampled 20 dialogues for each emotion type from dialogues generated by ECCRG and two other baseline models for emotion controllable response generation, and also sampled 120 dialogues from DialoGPT-generated dialogues, with volunteers from three Dimensions score the quality of model generation. These three dimensions are: (1) **fluency** measures whether the response is natural and fluent; (2) **relevance** measures whether the response is semantically coherent with the context; (3) **emotion quality** measures whether the response accurately expressed the specified emotion type. Volunteers were asked to rate 1 to 5 on each of the three dimensions according to which responses were made: 1-unacceptable, 3-moderate, 5-very excellent, and 2 and 4 for transitions of uncertainty.

5 Results and Analysis

5.1 Automatic Evaluation

Comparison with Baselines. The upper part of Table 2 shows the experimental results of ECCRG with baseline methods on automatic evaluation metrics. BLEU-1 and BLEU-2 represent the similarity of the generated response to the ground-truth, and also reflect the impact of ECC-layer fused with control

Table 3. Manual evaluation results

Model	Fluency	Relevance	Emotion Quality
Seq2seq-emo	3.47	2.93	2.79
ECM	3.17	3.38	2.98
DialoGPT	**3.68**	3.53	–
ECCRG	3.45	**3.82**	**3.36**

conditions; PPL represents the fluency of the generated utterances. From the comparative experimental results, it can be seen that ECCRG is significantly outperform than other baselines on BLEU and emotion accuracy. In terms of distinct indicators, the score of ECCRG is lower compared with the advanced model DialoGPT, but it still shows advantages in comparison with Seq2seq-emo and ECM. In general, ECCRG generates responses under the influence of control conditions, which reduces the search space during the decoding process to a certain extent, the generated responses are significantly closer to the ground-truth, and can express the specified emotions more accurately. Considering our motivation to generate models in a controlled range, it should be explainable at the expense of some fluency and variety, a problem that also arises in other controllable text generation methods.

Ablation Study. The lower part of Table 2 shows the comparative experimental results of ECCRG after ablation of partially optimization targets. From the comparison results with the full model, we make the following analysis:

(1) Without using the self-reconstruction loss $\mathcal{L}_{recon}$, the most significant change that can be observed is the reduction of the BLEU value, and the emotional accuracy rate is also reduced to a certain extent, which indicates that the self-reconstruction Loss is a key part of the controlled generation of the model, which can significantly improve the similarity between the response generated by the model and the specified text content, and it is also helpful for emotion expression when the text content and target emotion in the control conditions are consistent.
(2) Without using the emotion controlled loss $\mathcal{L}_{emo}$, we find that the performance of the model on all metrics is getting worse. Different from $\mathcal{L}_{recon}$, $\mathcal{L}_{emo}$ fuses emotion labels with language models. In this case, the generation of the model needs to consider the fluency of the text content, meanwhile, the emotion information is used as additional knowledge to enrich the diversity of the generated responses.
(3) Without using the adversarial learning loss, the performance of the model on BLEU is improved, which may be because the model is affected by more $\mathcal{L}_{recon}$. Except that $\mathcal{L}_{adv}^{emo}$ is helpful for emotion accuracy, the overall adversarial loss did not bring about the impact we expected. Without using the semantic adversarial loss $\mathcal{W}henL_{adv}^{sem}$, the model achieves the best performance on fluency. This shows that it is difficult for the model to process the adversarial loss signal returned by the discriminator. In order to weigh the adversarial loss, the quality of the generated text is reduced.

5.2 Manual Evaluation

Table 3 shows the results of manual evaluation. We used the Fleiss' Kappa score to measure the consistency of ratings among 5 volunteers, in which the Fleiss' Kappa scores for fluency, relevance, and emotion quality were 0.51, 0.46, and 0.57, respectively, indicating that the three evaluation dimensions of human evaluation results Both achieved "moderate consistency".

The results of manual evaluation show that, compared with the baseline method, ECCRG shows the best effect in both relevancy and sentiment quality, especially in relevancy, which shows the controllability of ECCRG in content An advantageous effect is provided. In terms of fluency, both ECCRG and another emotion-controllable response generation model, ECM, are lower than GPT-2 and Seq2seq-emo, and our model favors the target emotion in generating results at the expense of fluency. Another possible reason for the lower fluency than the Seq2seq model is that our model generates longer replies, and the semantic inconsistency in responses is caused by the instability of multiple optimization objectives during training.

Table 4. Samples of emotion-controllable response generated by ECM and ECCRG

	What do you think of Starbucks?	
ECM	Admiration	It's a nice **place**!
	Approval	It is cool to be kind!
	Caring	I think it's good for you.
	Joy	Big smiles, good for you!
	Sadness	I'm sorry to hear that.
	Surprise	I'm *not sure* why it's not being done.
ECCRG	Admiration	I'm not a *fan* of Starbucks, but I do love their **coffee**.
	Approval	I'm *with you*. I'm going to try and get a cup of **coffee** every day.
	Caring	I don't know, but I' m sure it's good for your **wallet**.
	Joy	It's the best. I'm a **coffee** drinker, and I love it!
	Sadness	I don 't hate Starbucks. I just don't like the way they **serve coffee**.
	Surprise	I'm *not sure*, but it's just a **coffee place** , *isn't it*?
	Rock music always makes me feel alive.	
ECM	Admiration	Wow, you look great!
	Approval	I hope you enjoy it.
	Caring	Good for you man!
	Joy	So happy for you man!
	Sadness	This music makes me crying.
	Surprise	What music is this?
ECCRG	Admiration	Awesome! I'm a *big fan* of **the Beatles**.
	Approval	Good luck dude, I'm pretty sure that's a quote from the **movie**!
	Caring	I wish you guys are doing well!
	Joy	I'm glad I'm not the only one who *feels this way*.
	Sadness	I'm not sure if you're serious or not, but I *don't really listen* to it.
	Surprise	Because it's so **metal**?

5.3 Case Study

For emotion control, we convert the specified emotion labels into text adding to the control condition in the manner described in Sect. 3.3. For content control, we add the target text directly to the control condition.

Table 5. Emotion- and content-controllable response samples generated by ECCRG

Do you like cats or dogs?	
Content: I like dogs	I like dogs, but I'm not a fan of cats
Content: I like cats	I like cats because they're cute and they're not that bad
Content: I like cats Emotion: Joy	I love cats so much! I've had a few cats that were very comfortable
Content: I like cats Emotion: Sadness	I like cats but they 're loud and annoying

Table 4 presents two set of cases where ECM and ECCRG generate responses under emotion control conditions. From the comparison of the generation results of ECCRG and ECM in the table, we can first intuitively observe that the utterances generated by ECCRG are longer, which makes the responses contain richer information and reflects better diversity. More emotional words (underlined words in the table) are used in the responses generated by ECM to express emotions explicitly, but too much reliance on emotional words makes the semantic content of the sentence relatively simple, and it's also difficult to capture the subtle differences between positive emotions like "joy" and "admiration". In addition to a small amount of emotional words, the responses generated by ECCRG can also express the specified emotions through combinations of general words (words in italics in the table), for example, "approval" is expressed by "I'm with you". Since ECCRG is built based on pre-trained language model, which makes the quality of its generation benefit from the linguistic and commonsense knowledge accumulated by the pre-trained model, additional concept words are used in the responses (bold in the table) such as "coffee" in relation to "Starbucks" and "the Beatles" in relation to "rock music".

Table 5 presents a case where ECCRG generates responses given specified emotion and content. Overall, ECCRG is able to well reflect the content conditions in the generated responses, which means that we can adjust the semantics of the responses generated by the dialogue system by modifying the textual content. But we can also see from the table that when the given emotion and content control conditions are inconsistent, for example, when the emotion condition is specified as "sadness" and the target content is specified as "I like cats" at the same time, the response semantics generated by ECCRG are not coherence, which is a problem that should be further solved in practical applications.

6 Conclusion

In this paper, we propose a emotion- and content-controlled response generation model ECCRG, which integrates a text-type control condition into the language model's intermediate hidden state by adding a self-attention layer to the pre-trained language model, to achieve the purpose of making the language model generation results controllable at a fine-grained level. We separately write emotion type and textual content into the control condition text, enabling the responses generated by ECCRG to express specific emotions, or to include semantic conditioned on content. We fine-tune the pre-trained language model and train ECC-layer on the Reddit dataset. The experimental results compared with multiple baseline methods show that ECCRG, given both emotion and content conditions, generates responses that are closer to the ground-truths and have better emotion accuracy. However, the experimental results also show that ECCRG is insufficient in balancing the natural expressions of emotion and semantics, which will also be a worthwhile research direction in our future work.

References

Asghar, N., Poupart, P., Hoey, J., Jiang, X., Mou, L.: Affective neural response generation. In: Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A. (eds.) ECIR 2018. LNCS, vol. 10772, pp. 154–166. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-76941-7_12

Chan, A., Ong, Y.-S., Pung, B., Zhang, A., Fu, J.: CoCon: a self-supervised approach for controlled text generation. In: International Conference on Learning Representations (2020)

Colombo, P., Witon, W., Modi, A., Kennedy, J., Kapadia, M.: Affect-driven dialog generation. arXiv preprint arXiv:1904.02793 (2019)

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)

Gao, J., Bi, W., Liu, X., Li, J., Shi, S.: Generating multiple diverse responses for short-text conversation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 6383–6390 (2019)

Huang, C., Zaiane, O.R., Trabelsi, A., Dziri, N.: Automatic dialogue generation with expressed emotions. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 49–54 (2018)

Li, J., Galley, M., Brockett, C., Gao, J., Dolan, B.: A diversity-promoting objective function for neural conversation models. arXiv preprint arXiv:1510.03055 (2015)

Lin, Z., Madotto, A., Bang, Y., Fung, P.: The adapter-bot: all-in-one controllable conversational model. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, pp. 16081–16083 (2021)

Liu, Y., et al.: Roberta: a robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

Papineni, K., Roukos, S., Ward, T., Zhu, W.-J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pp. 311–318 (2002)

Partala, T., Surakka, V.: The effects of affective interventions in human-computer interaction. Interact. Comput. **16**(2), 295–309 (2004)

Polzin, T.S., Waibel, A.: Emotion-sensitive human-computer interfaces. In: ISCA Tutorial and Research Workshop (ITRW) on Speech and Emotion (2000)

Radford, A., et al.: Language models are unsupervised multitask learners. OpenAI Blog **1**(8), 9 (2019)

Skowron, M.: Affect listeners: acquisition of affective states by means of conversational systems. In: Esposito, A., Campbell, N., Vogel, C., Hussain, A., Nijholt, A. (eds.) Development of Multimodal Interfaces: Active Listening and Synchrony. LNCS, vol. 5967, pp. 169–181. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12397-9_14

Song, Z., Zheng, X., Liu, L., Xu, M., Huang, X.-J.: Generating responses with a specific emotion in dialog. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3685–3695 (2019)

Zhang, Y., et al.: Dialogpt: large-scale generative pre-training for conversational response generation. arXiv preprint arXiv:1911.00536 (2019)

Zheng, C., Liu, Y., Chen, W., Leng, Y., Huang, M.: Comae: a multi-factor hierarchical framework for empathetic response generation. In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 813–824 (2021)

Zhong, P., Wang, D., Li, P., Zhang, C., Wang, H., Miao, C.: Care: commonsense-aware emotional response generation with latent concepts. arXiv preprint arXiv:2012.08377 (2020a)

Zhong, P., Zhang, C., Wang, H., Liu, Y., Miao, C.: Towards persona-based empathetic conversational models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 6556–6566 (2020b)

Zhou, H., Huang, M., Zhang, T., Zhu, X., Liu, B.: Emotional chatting machine: Emotional conversation generation with internal and external memory. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 32 (2018)

Zhou, L., Gao, J., Li, D., Shum, H.-Y.: The design and implementation of xiaoice, an empathetic social chatbot. Comput. Linguist. **46**(1), 53–93 (2020)

Zhou, X., Wang, W.Y.: Mojitalk: generating emotional responses at scale. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1128–1137 (2018)