SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

October 2022, Vol. 65 202102:1–202102:18
https://doi.org/10.1007/s11432-020-3356-4

An affective chatbot with controlled specific emotion expression

Chenglin JIANG[1], Chunhong ZHANG[1*], Yang JI[1], Zheng HU[1], Zhiqiang ZHAN[1] & Guanghua YANG[2]

[1]*State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing* 100876*, China;*
[2]*School of Intelligent Systems Science and Engineering, Jinan University, Zhuhai* 519070*, China*

Received 1 December 2020/Revised 9 April 2021/Accepted 12 July 2021/Published online 27 September 2022

Abstract Endowing a chatbot with the capability of specific emotion expression will significantly improve both chatbot's usability and users' satisfaction. Recently, many studies on open-domain neural emotional, conversational models (chatbots) have been conducted. However, enabling a chatbot to control what kind of emotion to respond to in conversation explicitly is still under exploration. This paper proposes a novel affective chatbot based on the sequence-to-sequence framework, responding with appropriate emotion like a human. In particular, a new module called single emotion generator is designed in the new chatbot model to address the existing issue of controlling over reacting emotion. It enables the chatbot to select the appropriate emotion for a response when interacting with users. In the decoder, an affective lexicon-based method generates emotion-awareness responses based on the specific emotion controlled by the single emotion generator. The proposed chatbot outperforms mainstream baseline algorithms for both semantic fluency and emotion consistence metrics through experimental evaluation. The experimental results also demonstrate that the new chatbot obtains the ability to control the emotion for response explicitly and responds emotionally with the specific emotion.

Keywords natural language generation, emotional chatbot, sequence-to-sequence, emotion distribution, emotion analysis

Citation Jiang C L, Zhang C H, Ji Y, et al. An affective chatbot with controlled specific emotion expression. Sci China Inf Sci, 2022, 65(10): 202102, https://doi.org/10.1007/s11432-020-3356-4

1 Introduction

Generally, humans can perceive and express emotions with language in communications, as well as they can control the specific emotion expression in various situations on their own [1, 2]. Thus, to create a chatbot capable of communicating with a user at the human level, it is necessary to equip the machine with the ability of emotion expression and place emotion in control with careful system design [3].

In earlier studies, integrating emotional manner in conversational agents has been studied primarily [4, 5]. However, the emotion policy they employed is primarily achieved through manual rules. The experts write these rules based on psychology findings or careful investigation in real conversation corpus, which makes complex emotion expression modeling difficult and limited to small-scale generation. Recently, due to the substantial development of deep learning, some researches have been conducted to construct dialog models with more "emotional" responses by using deep neural network models, such as emotional chatting machine (ECM) [6], affect-driven model [7], EmoDS [8] and CDL [9]. These models use specific emotional constraints to generate emotional responses when conversing with specific emotions. They cannot, however, control the emotion for responding like a human because the category of the responding emotion must be assigned by external decision-makers, similar to manual input by users. For example, as shown in Figure 1(a), the model aims to generate multiple responses, each of which is with one of the specific emotion categories of Sad and Happy, and external decision-makers select the two emotion

* Corresponding author (email: zhangch@bupt.edu.cn)

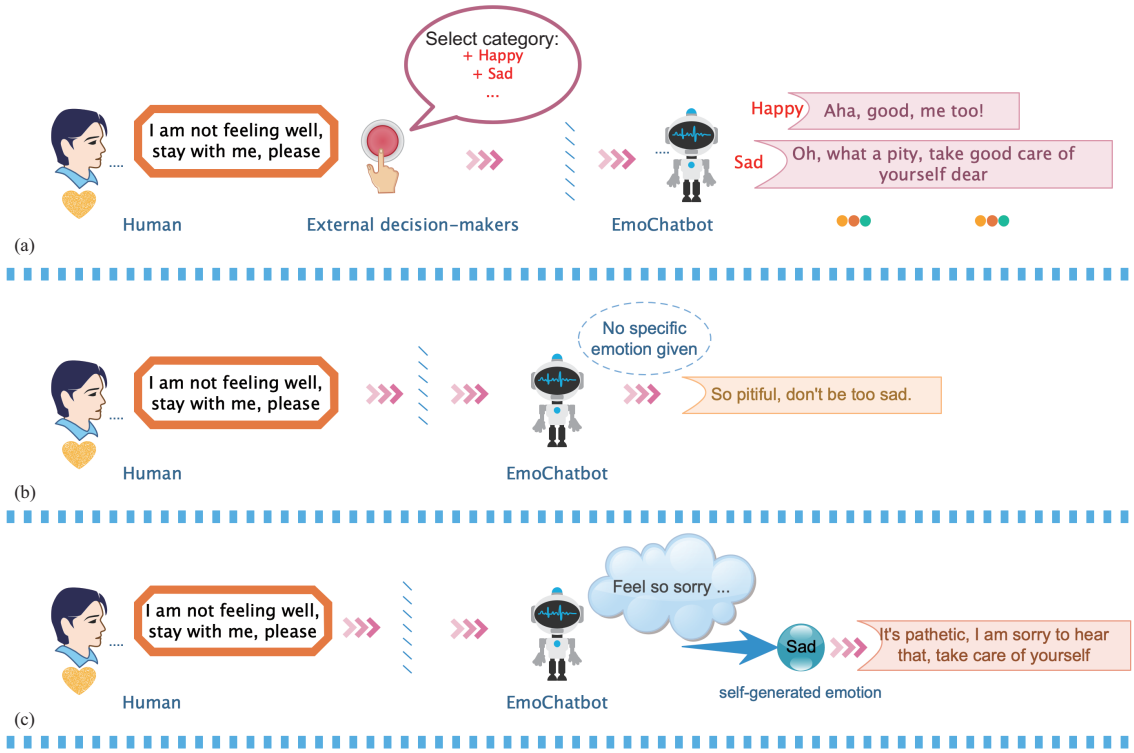

Figure 1 (Color online) (a) Affective response generation with manual given emotion in categories. The red button suggests the external control over emotion categories selection for responding. (b) Affective response generation without manual control and emotion specification. (c) Self-control affective response generation by our new presented chatbot model.

categories. Except for the above-mentioned models, some other fully data-driven models have been presented [10–12]. They can generate emotional responses from beginning to end without the intervention of external decision makers in interaction. These models, however, attempt to respond emotionally by modeling only the fine-grained emotion features in the data corpus, such as the information of various emotional words in the post and response sentences. They are unsure which high-level abstraction of the emotion expression they used in the generation. In other words, they have no control over the specific emotion category used, which may result in inappropriate emotional expression. This is also not the case in human-to-human interaction because our humans are generally aware of the concrete emotion category in expression. Thus, there is a need for explicit constraints on the control over selecting a proper emotion for a generation. An example is given in Figure 1(b), where the model produces a response without a specific emotion category defined by external decision-makers. The generated response sentence contains affect words like "pitiful," and "sorry," which suggests fine-grained emotional information, but the emotion category in usage is not specified.

To tackle the aforementioned problem, we propose a novel framework for emotional-chatbots, aiming to achieve explicit emotion control on the response generation through high-level abstraction of emotion expression (emotion category) and fine-grained emotion features (emotion words in the inferred emotion category). The corresponding example is given in Figure 1(c). When the chatbot is seeing the post of a human, it can adaptively generate a desirable response with emotion category Sad by containing corresponding emotional word 'sorry'. To fulfil the objective of adaptive emotion controlling, in addition to the conventional postencoder and responsedecoder framework of chatbot, a single emotion generator (SEG) model is proposed to bridge the adaptive emotion transition from post to response. The SEG model explicitly captures the human-like emotion mapping between post and response, which is then used by the responsedecoder to control the emotion generation of the response. Two approaches to emotion mapping for SEG are specifically designed. Since a single sentence in human conversation tends to evoke multiple emotions [13], instead of one-hot mapping, it is reasonable to model the emotion mapping of soft probability distribution over all the potential emotions to comprehensively depict the emotional nature. In contrast, SEG uses a dominant emotion classification approach to predict the dominant emotion for the response sentence, based on the observation that there is usually a dominant emotion among all the

potential emotions in real conversation. The combination of the two approaches will significantly improve emotion mapping modeling capability. Note that because most of the existed chatbot data corpus have no direct emotion distribution information for SEG learning, we therefore propose two methods, poliarity predict algorithm (PPA) and emotional distribution algorithm (EDA), together to construct emotion mapping distribution labels before training the SEG model. Based on the output of SEG, an affective weighted lexicon-based decoder is then proposed to generate a response sentence with both semantic representation and adaptive emotion.

The contributions of our algorithm are summarized as below.

• A novel emotional conversation generation framework is provided, which consists of a module named SEG, encoder, and decoder based on the traditional sequence-to-sequence model. In virtue of this framework, an autonomous chatbot can control both the selection of specific response emotion and emotion-awareness responses generation in conversation.

• A new module SEG is designed to learn the contextual affective relation between posts and responses, which can generate a desirable response emotion for one given post. In this context, the chatbot has the capability of control of emotional expression.

• To generate distribution labels for the conversation corpus with only single emotion labels, an emotional distribution labeling strategy based on PPA and EDA is proposed. Finally, convincing results from large-scale experiments on two Chinese conversation datasets and an extended emotion lexicon dictionary demonstrates the feasibility and efficiency of the proposed chatbot framework.

2 Related work

Emotion plays an important role in cognition and social behavior, and emotion has more social functions, such as eliciting people's particular response or recruiting social support [14,15]. Furthermore, the existing study also shows that emotion might be a related measure of decision-making [16]. Upon these theories, it is obvious that incorporating emotions could allow dialog models to emulate humans' conversational behavior and strengthen the emotional connection with human users.

2.1 Emotion analysis in dialog

Emotion analysis in conversation is closely related to the emotion classification task, which can be defined as the task of classifying or predicting an emotion given a conversational sentence or contexts based on the relevant emotion representation. Most computational models of emotion have three representation categories: the dimensional approach, the discrete approach, and the appraisal approach [17]. Therefore, emotion analysis often needs the assistance of emotion categorization models and algorithms [18]. Nowadays, with the increasing interest from researchers, the related research has seen dramatic increases, and many emotion categorization algorithms are constructed to drive the system [19], such as the complex computational algorithm and wildly used neural network models. Specifically, as for the emotion analysis in dialog, in many existing kinds of research, the emotion is seen as an attribute attached to the entire sentence, which might be argued as an oversimplification, the objective is always to match the single emotion category label from ground truth [20, 21]. Furthermore, some studies have paid attention to the assumption that one sentence might involve multiple emotion categories with different intensity and also regarding different aspects [22]. Aspect-based analysis [23, 24] and emotion distribution learning (EDL) [13,25] have been produced. These approaches present the emotional distribution in the input sentence more intuitively and comprehensively to better help with the analysis of the emotional complexity and expression tendency in a sentence.

2.2 Inchoate rule-based models

Many researchers experimented with incorporating emotional elements with chatting agents in the early years and had some success. Some studies on incorporating emotional expression in conversational models found that adding a module to address emotion in conversational agents increased user satisfaction [26]. An affective listener was designed to capture the user's affective states and expression contents [27]. For decreasing the anxiety of young adults with the help of a rule-based empathic agent, woebot was constructed in [28]. However, these studies mainly depend on manual rules derived from existing psychological findings, and the final emotional responses are chosen from the pre-constructed candidate dataset.

Then, it is easy to find out that these rule-based methods are difficult to deal with ambiguous characters in emotion expression and are limited to the small-scale corpus.

2.3 Affective end-to-end conversation models without emotion specification

In recent years, with the development of researches in deep learning, a core structured prediction framework, sequence-to-sequence model [29] with neural networks, has obtained great success in many sequence-generate based studies, especially in neural dialog generation [30–32]. With a growing interest in incorporating emotion analysis into the response generation process, some studies have automatically assigned emotion ability and related characters to various dialog models rather than relying heavily on manual rules. Some studies use extra models to help incorporating the affective information or personal features of humans from corpus, like [11], which introduced an affective language model to generate emotional, conversational text in conditioned categories. Meanwhile, Refs. [33,34] had addressed the personality for coherent conversation generation in better expression. And some other studies directly depended on the fine-grained emotion information carried in the training dataset and the existed affect-rich lexicons, such as [10], which presented a novel end-to-end affect-rich neural conversational model, named ARS2S. This model incorporated rich affective knowledge by using valence-arousal-dominance (VAD) notations [35] and bias attention to make sure the final unique response is emotional. On the other hand, the aforementioned models completely follow the implicit affective features captured from the given data to generate the emotional response. The quality of the data corpus heavily influences emotional expression control, which ignores explicit control over specifying a proper emotion category for a generation. The agent is then unsure which specific emotion they are performing in one conversation, resulting in inappropriate expression and even many irrelevant responses.

2.4 Emotional response generation with specific emotion

Most recently, several more advanced models have paid attention to generating responses with specific emotions. They aimed to endow conversational models with controlled emotion category explicitly. For example, EMOTICONS [7] used continuous emotional representations and affective regularizer for words penalizing in the training stage. ECM [6] introduced the internal and external memory for emotion expression, the EmoDS [8] achieved expressing desired emotion explicitly or implicitly by lexicon-based attention and emotional classifier guidance. The CARE [36] could learn and construct commonsense-aware emotional latent concepts of the response. However, these studies partially focused on enable chatbots to express emotions with a predefined emotion category set. While interacting with these models, the specific emotion type for response generation is always defined by external decision-makers, such as manual input categories. They cannot generate one specific emotion for response by themselves, which is not the case of controlling the responding emotion trend like a human.

As illustrated, the flaws of the aforementioned models must be addressed appropriately by a large-scale "complete" emotional chatbot, where "complete" refers to one chatting machine's subjective initiative to control its emotional expression in a specific manner on its own. The purpose of this paper is to propose an affective chatbot to address the aforementioned issues. Unlike previous work on emotional conversation generation studies, the proposed chatbot addresses explicit control over the selection of responding emotion category and emotional response generation with emotional words, drawing inspiration from existing emotion analysis approaches.

3 Methodology

3.1 Model overview

First, an overview of the model is introduced. The chatbot model is constructed based on sequence-to-sequence architecture with attention [37]. As shown in Figure 2, it presents our new chatbot framework with three parts, including encoder, decoder and SEG. Encoder is responsible for providing both decoder and SEG with the semantic information representation of the given post sentence. The SEG and decoder share the same encoder. The SEG, which is shown in the lower left part in Figure 2, is designed to explicitly help the chatbot to explicitly estimate the reasonable emotion suitability for the response according to the overall conversation context. It takes the encoded post representation from encoder as input, and can predict an emotion distribution for response based on the joint learned emotional

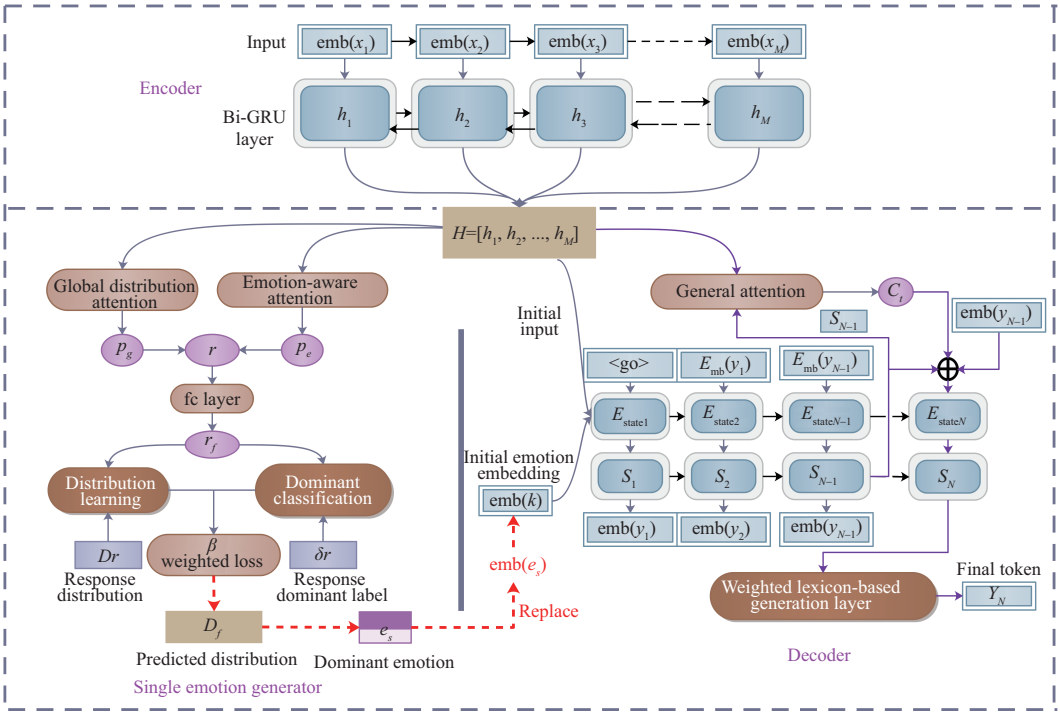

Figure 2 (Color online) Overall architecture of our chatbot framework. It contains three parts: encoder, SEG and decoder. SEG and decoder share the same encoder.

information. In the training stage, SEG needs both single dominant response emotion labels that existed in corpus and soft response emotion distribution labels that generated by designed affective distribution labeling strategy to fulfil the joint learning task. The decoder is shown in the lower right part in Figure 2, which is designed with a weighted lexicon-based generation layer. It takes the encoded post representation from encoder and a specific emotion category as input, models the internal emotion state in the decoding process, and finally generates an emotional response sentence with emotion words that are related to the input specific emotion. The input emotion category is the ground truth single emotion label for each response from same corpus as that of SEG in the training stage. SEG and decoder have no direct explicit interaction in training, but both of them update the learning state of the post representation from encoder. In the inference process, the two modules start to cooperate with each other to achieve the complete emotion expression in interaction. Specifically, the SEG is supposed to predict an emotion distribution based on the given post, and extract a proper specific emotion category for decoder from the predicted distribution, and then decoder produces an affective response. In Figure 2, the process under red dashed line suggests the interaction relation between decoder and the SEG during the inference mode.

3.2 Problem formulation

Given a post of length M: $X = (x_1, x_2, \ldots, x_M)$, the objective is to select a proper emotion category e_s for response from a set of emotion categories E, and generate an emotional response sentence of length N: $Y = (y_1, y_2, \ldots, y_N)$ that is coherent with e_s specifically. Meanwhile, $x_i \in V$ and $y_j \in V$ are words in post and response, respectively. The set E consists of all specific emotion categories that the chatbot can express, and in this paper, $E = \{\text{Like}, \text{Disgust}, \text{Angry}, \text{Happy}, \text{Sad}, \text{Other}\}$, aligning to previous studies [38, 39]. Note that the limited emotion categories in the set E is just for illustration in this paper and the categories can be extended to more fine-grained emotion categories. Noticeably, $V = V_n \cup V_e$ is the vocabulary used for generation, where $V_n \cap V_e = \emptyset$. V_n is a vocabulary of neutral words where V_e is an emotion lexicon. Furthermore, the lexicon V_e is divided into several subsets V_e^k, each of which stores the words associated with an emotion category k in E.

3.3 Affective chatbot model construction

As illustrated above, the presented chatbot model contains three modules, which achieves response emotion category selection and emotional response generation. In this subsection, further construction details

of these three parts: encoder, SEG and decoder (affective weighted lexicon-based decoder) are given in following subsections, respectively.

3.3.1 *Encoder*

Encoder provides the information in post sentences by transforming the post sentences into vector representations. We use bidirectional gated recurrent unit (Bi-GRU) network [40] as the encoder in chatbot, and use the hidden state vectors as the representations of an input post, $X = (x_1, x_2, \ldots, x_M)$. Formally, the hidden states of encoder are computed as follows:

$$\begin{aligned} \overrightarrow{h_i} &= \mathrm{GRU}_{\mathrm{forward}}\left(\mathrm{emb}(x_i), \overrightarrow{h_{i-1}}\right), \\ \overleftarrow{h_i} &= \mathrm{GRU}_{\mathrm{backward}}\left(\mathrm{emb}(x_i), \overleftarrow{h_{i+1}}\right), \end{aligned} \tag{1}$$

where $i = 1, 2, \ldots, M$, $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ are the i-th hidden states of forward and backward GRUs, respectively. With respect to $\mathrm{emb}(x_i) \in \mathbb{R}^K$, it is the K-dimensional word embedding vector corresponding to x_i in the post. In this paper, we use pre-trained word embeddings (details in Subsection 4.1) that incorporate in the emotional information of words. The final hidden state representation by the two GRUs is the concatenation of $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$, namely h_i.

3.3.2 *SEG*

As illustrated, the SEG is designed to enable the chatbot to decide what kind of emotion is suitable for responding by itself. SEG learns the contextual affective relations between posts and responses in training stage, following the strategy of joint learning [25] to integrate the two approaches, EDL and emotion category classification. While in inference mode, the generator could predict an emotion distribution for response from the given post, and extract the final responding emotion e_s from the predicted distribution. Formally, the emotion distribution is a representation set of all emotions intensities that are evoked in a single sentence T based on predefined emotions set and it is usually formed as $D_T = \{d_T^j\}_{j=1}^z$, where $\sum_{j=1}^z d_T^j = 1$, d_T^j denotes the proportion of the j-th emotion category in pre-defined emotion set E, and z is the size of this set. The distribution can help us take all possible emotions influences into consideration, and e_s is the extracted one with maximal probability in the predicted emotional distribution, which is defined as dominant emotion.

Concretely, in order to realize the contextual emotional relation mapping and make sure trustful emotion distribution prediction, we innovatively use the ground truth emotion labels of response in data corpus as pseudo labels for the corresponding posts. Then, the objective of EDL is to map each post sentence to its corresponding response emotion distribution labels. In the emotion category classification, each post sentence is classified to the ground truth dominant emotion labels of the response. Through the joint training of both two tasks, the chatbot could learn and use the contextual emotional relations to predict a proper emotion distribution for response based on the given input post. In detail, as shown in Figure 2, the SEG shares the same encoder with decoder. It takes the vector matrix H as input, which consists of all input hidden states. Then we apply the idea of making prediction by focusing on creating high-quality latent representation with contextual and emotional features [41] to learn a distribution representation for the post. The attention mechanism for relation classification tasks [42] is adopted to capture both the global and specific emotional features of post. The representation p_g with global features of the post is formulated by a weighted sum of these output vectors:

$$p_g = H\alpha_g^{\mathrm{T}}, \quad \alpha_g = \mathrm{softmax}(u_g^{\mathrm{T}} e_g), \tag{2}$$

$$e_g = \tanh\left(W_g^{\mathrm{T}} H + b\right), \tag{3}$$

where $H \in \mathbb{R}^{K \times M}$, K is the dimension of the word vectors, α represents the attention weight that is determined by activation function e_g to compute on all encoded representations in H of the input post, which shows global attention on all the words in the post. The dimensions of α_g, u_g, p_g are M, K, K separately. And the representation p_e with specific affective features of the post is formed by leaving attention on the words in emotion lexicon V_e and the modifiers in the post to capture all explicit affective

features. We still follow the computing process of p_g to form p_e:

$$p_e = H\alpha_e^{\mathrm{T}}, \quad \alpha_e = \mathrm{softmax}(w_e^{\mathrm{T}} e_a), \tag{4}$$

$$e_a = \tanh\left(W_a^{\mathrm{T}}\left[G_e H\right] + b\right), \quad G_e = \eta_{\mathrm{intens}} \times \mathrm{Extract}\left[X\right], \tag{5}$$

where $\mathrm{Extract}\left[X\right]$ denotes an index-select filter matrix (with 0 or 1 inside) constructed based on whether there is an emotion relative word or not in the post X that belongs to V_e. Besides, η_{intens} is a weight matrix (with 1 or specific weight value inside), which is constructed based on whether there is an intensifier word or not before emotion words and it has the same shape with $\mathrm{Extract}\left[X\right]$. In detail, we use an intensity lexicon I_l that contains words under four-level general degree: 'Extreme', 'Very', 'More' and 'Ish', the intensity decays from left to right and we set different scores from 4 to 1 for each modifier with different intensities definitely, once the word before one emotion word is in I_l. The value of corresponding element in η_{intens}, which has the same index as this emotion word in $\mathrm{Extract}\left[X\right]$, is assigned with the related intensity weight score, and if there is no intensifier word before one emotion word, the value always equals to 1. $[G_e H]$ denotes the reconstructed emotion-aware representation, in which the original no-emotional word features in H are filtered. The concatenation of p_g and p_e is used as the latent representation of the post $r = [p_g; p_e]$ to form the final distribution representation r_f of post used for the tasks in joint learning:

$$r_f = \mathrm{softmax}(\mathrm{fc}(\tanh(r))), \tag{6}$$

where r_f contains the activation values of the results through a fully connected layer fc() based on r. Meanwhile, it is defined by $r_f = \{d_X^j\}_{j=1}^z$ to approximate the form of post emotion distribution. Note that z is the number of emotion categories in E.

After obtaining the distribution representation of the post sentence, we use EDL to model the contextual emotional relation of each post and corresponding response. As illustrated, we map each post sentence to its corresponding response's emotion distribution to achieve EDL. The KL loss in [13] is used to measure the distance between the predicted and true distributions. The true labels are the generated soft emotion distribution labels of response sentences (details in Subsection 3.4), defined by $D_{(Y)} = \{d_Y^j\}_{j=1}^z$. The KL loss is

$$\mathrm{KL}(\Theta_1) = -\sum_{j=1}^{z} d_Y^j \ln d_X^j, \tag{7}$$

where $d_Y^j \in D_Y$ and $d_X^j \in r_f$. Note that there is a dominant label in distribution which always dominates the final expression of responding [25]. We learn EDL with dominant emotion classification simultaneously to improve the dominant emotion accuracy in predicted distribution. We use cross-entropy loss to formulate this emotion classification:

$$E_{\mathrm{domin}}(\Theta_1) = -\frac{1}{z}\sum_{j=1}^{z} \widehat{\delta}_j \ln d_X^j + \lambda \left\|\Theta_1\right\|^2, \tag{8}$$

where $\widehat{\delta}$ is the one-hot represented dominant emotion label of response. Meanwhile, the L2 regularization with a hyper-parameter λ is used to avoid over-fitting. The overall joint learning loss of SEG is combined by loss functions of the two tasks with different weights finally as follows:

$$L(\Theta_1) = \beta KL(\Theta_1) + (1-\beta)E_{\mathrm{domin}}(\Theta_1), \tag{9}$$

where $\beta \in [0, 1]$ and the weight controls the importance of two losses in training. While in the inference, SEG can predict a response emotion distribution $\widehat{D}_{(Y)} = \{\widehat{d}_Y^j\}_{j=1}^z$, and choose the emotion category in E as the dominant responding emotion e_s, which has the highest proportion $\widehat{d}_Y^j$ in $\widehat{D}_{(Y)}$.

3.3.3 *Affective weighted lexicon-based decoder*

In the chatbot model, the decoder contains a uni-directional GRU network and the emotional generation is based on the pre-defined emotion categories in the set E, formally. Considering that the emotion presented in dialogs is not absolutely static according to the psychological findings in [43, 44]: emotional responses are relatively short lived and involve changes, and the emotion is in a dynamic situation in expression [45], we model the internal emotion state decay in each decoding process inspired by ECM [6],

in order to capture the emotion dynamics in generation. In detail, the emotion state decay suggests that at each decoding step the state decays by a certain amount and once the decoding process is completed, where the emotion state should decay to zero indicating the emotion is expressed completely. In formulation, the emotion state E^j_{state} is computed and updated by a read r_j and write gate r_w:

$$r_j = \text{sigmoid}\left(W_r\left[c_t; \text{emb}\left(y_{j-1}\right); s_{t-1}\right]\right), \quad w_j = \text{sigmoid}\left(W_w s_t\right), \tag{10}$$

where $j = 1, 2, \ldots, N$, s_0 equals to the last hidden state h_M of encoder, and $[\cdot;\cdot]$ denotes the operation that concatenates the vectors separated with semicolons. $\text{emb}(y_{j-1})$ is the K-dimensional word embedding of y_{j-1}. As for c_j, it is the weighted sum of all input hidden states $[h_1, h_2, \ldots, h_M]$ from encoder of j-step, and the related weights α are computed by the attention mechanism [46]:

$$c_j = \sum_{j=i}^{M} \alpha_{ij} h_j, \quad \alpha_{ij} = \frac{\exp\left(e_{ij}\right)}{\sum_{m=1}^{M} \exp\left(e_{mj}\right)}, \quad e_{ij} = v_a^{\mathrm{T}} \tanh\left(W_a s_{j-1} + U_a h_j\right), \tag{11}$$

where e_{ij} is the alignment model for attention score computing, which scores how well the inputs around position i and the output at position j match, and v_a, W_a, U_a are the weighted matrices. Then, E^j_{state} can be formulated as

$$E^{r,j}_{\text{state}} = r_j \times E^j_{\text{state}}, \quad E^{j+1}_{\text{state}} = w_j \times E^j_{\text{state}}, \tag{12}$$

where "$\times$" is element-wise multiplication, $E^{r,j}_{\text{state}}$ denotes the emotion state vector for decoding computing and then updates the $(j+1)$-th emotion state by r_w. Finally, decoder GRU will update its state s_j conditioned on $E^{r,j}_{\text{state}}$, the previous hidden state s_{j-1}, the context vector c_j, and the previously decoded word y_{j-1}, which records the emotion state decay successfully.

$$s_j = \text{GRU}_{\text{decoder}}([E^{r,j}_{\text{state}}; \text{emb}(y_{j-1}); c_j], s_{j-1}). \tag{13}$$

E_{state} is initialized as $\text{emb}(k)$. It is the embedding of the specific emotion category k in E, which is the ground truth single emotion label that is carried in data corpus during training process, and in the inference, k will be replaced by the dominant emotion e_s that is selected by SEG.

In order to ensure a reasonable emotional response generation, we further make explicit constraint on the generated words based on existed emotional lexicon, inspired by the fact that the emotion words are distinct with neutral words in a sentence [47]. Meanwhile, considering that the frequency of emotional words appearance is lower but dominant in the expression [7], we adopt the idea of inverse token frequency (itf) weight [48] to further balance the frequency difference, and try to stand out the emotion words in generated response. Finally, we design an affective weighted lexicon-based generation layer to balance the proper choice between specific emotion word and other generic word generation in each decoding step. To achieve this weighted generating layer, we set up constraints on the generation probability. In the generating process, the decoder always generates a token $\widehat{y_j}$ by sampling from the output probability distribution ϕ_j computed from the decoder's state s_j, which is formed as follows:

$$\begin{aligned} \widehat{y_j} \sim \phi_j &= P(y_j \mid y_1, y_2, \ldots, y_{j-1}, c_j, E^{r,j}_{\text{state}}) \\ &= \text{softmax}(W_o s_j). \end{aligned} \tag{14}$$

To assign frequency constraints to the generic words, we first split the original generation probability distribution ϕ_j into a generation probability distribution ϕ_k over all the emotional words w_e in current specific emotion lexicon V^k_e and a generation probability distribution ϕ_g over all other generic words w_g in V except the emotion words in V^k_e. The formulation of ϕ_k and ϕ_j is given as

$$\phi_{k_j} = \phi_j \times \text{Extract}[V], \quad \phi_{g_j} = \phi_j - \phi_{k_j}, \tag{15}$$

where $\text{Extract}\,[V]$ denotes an index-select filter matrix (with 0 or 1 inside) constructed based on whether there is an emotion word or not in vocabulary V that belongs to V^k_e. The "$-$" denotes the filter operation to set values in ϕ_j as zero, and the indexes of these changed values in ϕ_j are same as the non-zero value indexes in ϕ_k. Then, the probability distribution ϕ_g over all generic words is obtained. Next, the itf weight is added to ϕ_g with the purpose of adding an explicit constraint on the generic words probabilities, by

which the emotional features may stand out better and also help with content diversity improvement. Thus at each step j, we first need to refine the ϕ_{g_j} with itf weight as follows:

$$\phi_{\mathrm{fr}_j} = \mathrm{fr}_j \phi_{g_j} = \frac{1}{\mathrm{fre}(y_{\mathrm{tar}_j})^{\gamma}} \phi_{g_j}, \tag{16}$$

where ϕ_{fr_j} suggests the regenerated probability distribution over all w_g, fr_j is the global weight calculated based on the frequency of the target token y_j at time t, and γ is a hyper-parameter that controls the weight of frequency influence. Then, we explicitly model emotion expressions by estimating ϕ_{fr_j} with emotional probabilities ϕ_{k_j}:

$$\eta_t = \mathrm{sigmoid}(W_{\eta}^{\mathrm{T}} s_j), \tag{17}$$

$$\widehat{y_j} \sim \phi_{\widehat{y_j}} = \begin{Bmatrix} \eta_j \phi_{k_j} \\ (1-\eta_j)\, \phi_{fr_j} \end{Bmatrix}, \tag{18}$$

where s_j is the output state at time-step j and $\eta \in [0,1]$ is a type selector to control the weight of generating an emotional or a neutral word, and W_{w_e}, W_{w_n} and W_{η} are trainable parameters. $\phi_{\widehat{y_j}}$ is the final word decoding distribution which is a concatenation of ϕ_{fr_t}, ϕ_{k_t} after weighted computation.

3.3.4 *Loss function of affective response generation*

In this subsection, all cost functions used in end-to-end affective response generation part would be given as supposed. The first loss function is formulated as

$$L_1(\Theta_2) = -\sum_{t=1}^{N} p_j \log(\widehat{y_j}) \tag{19}$$

which is the conversational objective function of sequence-to-sequence model. It is used to minimize the cross-entropy error between the gold distribution p_j and the predicted one y_j during the chatbot training stage. To measure the explicit selecting process in the lexicon-based weighted generating layer, we design another loss function as follows:

$$L_2(\Theta_3) = -\sum_{j=1}^{N} \mu_j \log(\eta_j) \tag{20}$$

which is used to supervise the probability of selecting the proper words at the right time steps. Note that η_t is the probability of choosing an emotion word or a generic word and $\mu_t \in \{0,1\}$ is the true choice of a specific emotion word or a generic word in Y while decoding. With respect to the emotional state capturing loss, we just use the emotion state at the last step N E_{state}^{N} to compute the final decay results, noting that we expect the value reaching to zero in order to ensure the emotion expressed completely:

$$L_3(\Theta_4) = \left\| E_{\mathrm{state}}^{N} \right\|, \tag{21}$$

where $\|\cdot\|$ denotes the operation on computing norm of vector. Finally, the complete training cost function of our affective response generation approach is the combination of above functions:

$$J(\Theta_{2,3,4}) = L_1(\Theta_2) + L_2(\Theta_3) + L_3(\Theta_4). \tag{22}$$

Finally, we present the overall training objective of the chatbot model, which is combined with the loss functions of SEG and decoder in the training stage:

$$L_{\mathrm{overall}}(\Theta) = L(\Theta_1) + J(\Theta_{2,3,4}). \tag{23}$$

3.4 Affective distribution labeling strategy

It is known that in the training stage of SEG, we need the emotion distribution labels of the response sentences in the data corpus. However, the most existed large-scale dialog corpus only provides single emotion labels for the responses, so we leverage a new context-dependent strategy to generate soft emotion distribution labels for response sentences before training. The proposed strategy includes two algorithms,

Algorithm 1 Emotional distribution algorithm

Input: Y: response sentence; e_d: dominant emotion label; V_e, P_l, N_l: lexicon; ε: dominant weight;
Output: D_Y: affective distribution; $D = \{d_Y^j\}_{j=1}^z$, where $\sum_{j=1}^z d_Y^j = 1$.

Initial: PPA(Y) $\Rightarrow$ p_{e_d}: polarity; q_m and q_{sd}: allocation ratios with p_{e_d}; e: emotion in Y; l: list of categories in Y on E;
$e =$ none $\parallel$ $e \notin E$ $\parallel$ only $e_d \Rightarrow d_Y^j = 1$ when $j = \text{idx}(e_d)$, otherwise $d_Y^j = 0$;
len(l) $\geqslant 2$ and $e_d \in l$:
(a) Same polar $\Rightarrow d_Y^{j=\text{idx}(e_d)} = \varepsilon$, other $e \in l \Rightarrow d_Y^{j=\text{idx}(e)} = (1-\varepsilon)\text{frequency}_{(e_{\text{words}})}$;
(b) Opposite polar $\Rightarrow \diamond$len(l) = 2: same as (a) but set $d_Y^{j=\text{idx}(e_d)} = q_m$ $\diamond$len(l) $> 2 \Rightarrow d_Y^{j=\text{idx}(e_d)} = \varepsilon$, other $e \in l$ with $p_{e_d} \Rightarrow d_Y^{j=\text{idx}(e)} = q_{\text{sd}}(1-\varepsilon)\text{frequency}_{(e_{\text{words}})}$, opposite $p_{e_d} \Rightarrow d_Y^{j=\text{idx}(e)} = (1-q_{\text{sd}})(1-\varepsilon)\text{frequency}_{(e_{\text{words}})}$
len(l) $\geqslant 2$ but $e_d \notin l \Rightarrow$ add e_d to l and repeat step 3, but keep $d_Y^{j=\text{idx}(e_d)} \equiv \varepsilon$;
return D_Y.

PPA and EDA, and the computation process is inspired by the lexicon-based conversion strategy [25].

Formally, in EDA, which is summarized as Algorithm 1, the input is the response sentence Y with its ground truth dominant emotion label e_d in the corpus, desired emotion-awareness lexicons, V_e, P_l and N_l, and an manual assigned proportion weight for e_d. The output is the emotion distribution of response, defined by $D_{(Y)} = \{d_Y^j\}_{j=1}^z$. In detail, both the emotion lexicon V_e and positive and negative word lexicon P_l and N_l (construction in Subsection 4.1.2), are utilized to map the words from each sentence to the specific emotion and polarity. At the same time, we calculate the intensity probability d_j of the j-th emotion according to the mapped emotions in l if there exist other emotional words except for the ones with dominant emotions. Elsewise, we just use the one-hot distribution for the sentences. The probabilities are finally normalized to the emotion distributions. To incorporate the contextual interaction influence in computation, we cover the intensity of modifiers and polarity changing in expression based on the related affect words to further modify the probabilities of the dominant label e_d ($d_T^{j=\text{idx}(e_d)}$ set as ε) and others ($\sum d_T^{j\neq\text{idx}(e_d)} = 1-\varepsilon$), where idx() suggests the index of the related emotion category in set E. Specifically, the polarity p_{e_d} of e_d is firstly recorded as the primary one. And the probability only should be modified in two conditions when the mapped emotions are in different polarities: (1) if there are only two types of emotions in l including e_d, the value of ε is replaced by the general polarity weight ratio q_m; (2) if there are more than two emotions in l, the value of $1-\varepsilon$ should be further divided by the polarity deviation ratio q_{sd}, and should compute each probability with divided results of corresponding polarity. Both q_m and q_{sd} are computed from the results of PPA.

In PPA, we use extra two lexicons: D_l containing deny words like "not" and an intensity lexicon I_l that contains words under four-level general modifier class $R = \{A : \text{Extreme}, B : \text{Very}, C : \text{More}, D : \text{Ish}\}$, the intensity decays from left to right in R and we set different score S for each modifier with different intensity in R definitely, formed as $S = s_{i(i=A,\ldots,D\in R)}$. The initial score of each word in sentence is set to zero which will be increased by one if the word in V_e. The score is multiplied by intensity score S of related modifier in R, noting that we only consider the influence of modifiers before the current emotion word to ensure reliability. We also judge the appearance time of a denial word d before emotion word, current emotional score should be reversed to negative value when odd number of times appears, and this matches the truth that single deny introduces opposite (e.g., "not good" equals "bad" in expression). Finally, we obtain $q_m = \text{sum}(p_{e_d})/(\text{sum}(p_{e_d}) + \text{sum}(\text{opposite}))$ and $q_{\text{sd}} = \text{sd}(p_{e_d})/(\text{sd}(p_{e_d}) + \text{sd}(\text{opposite}))$ by sum ((sum()) and standard deviation (sd()) results over word score values in bio-polarity. We summarize PPA's procedure in Algorithm 2. In order to help researchers get a better understanding of the proposed algorithm, we open source the lexicons and clean code of our algorithms as EmotionDisEDAPPA[1)].

Algorithm 2 Polarity predict algorithm

Input: Y: response sentence; V_e, P_l, N_l, D_l, I_l: lexicon; S: manual assigned intensity score;
Output: L_{PPA}: a list of sum, mean and sd values of polarity score;

Initial: score(t) $= 0, y \in Y = \langle y_1, y_2, \ldots, y_N \rangle$; i_r presents $i \in I_l$ with intensity $r \in R$;
for each $y \in Y$ **do**
  if $y \in V_e \diamond i_r$ before $y \diamond d \in D_l$ appears odd times **then**
    $\Rightarrow$ Increase score(y) with 1 $\diamond$ Modify score(y) by $\times s_{i_r}$ $\diamond$ Reverse score(y) negative; $\cdots \triangleright$ (emotional $\diamond$ Modifier $\diamond$ Reverse).
  end if
end for
return set of score(y): S_y;
whether y in P_l or N_l: add S_y, mean S_y, sd $S_y \Rightarrow$ score(p), score(n) $\triangleright$ in sum/mean/sd form;
return L_{PPA}.

[1)] https://github.com/Jiangchenglin521/EmotionDisEDAPPA.

4 Experiments

4.1 Data preparation

4.1.1 *Conversation data corpus*

Two emotional conversation datasets, ESTC [6] and NLPCC2017[2)], are used to evaluate the performance of the proposed emotional chatbot. However, the original datasets are not well aligned with the requirement of our algorithm. Thus, we conduct three kinds of refinement to the ESTC and NLPCC2017 datasets. First, both datasets are re-formated as one-to-one dialog pairs, with each post having only a single response with the ground-true emotional label. The emotional label refers to the specific emotion category in E. Second, considering that the emotional annotation accuracy in the original ESTC is just 0.623, to reduce the noise of the dataset, we train an emotion classifier based on attention Bi-GRU to re-annotate the emotions of the dialog in ESTC with higher accuracy up to 0.89. The classifier is trained using the NLPCC2013[3)] dataset including 13252 samples and 5418 samples from NLPCC2017 dataset. Finally, we use our affective distribution labeling strategy to generate a soft emotion distribution label for each response sentence. The results show that the final dataset used for the experiment combines the refined ESTC and NLPCC2017 and consists of a total of 2538558 post-response pairs with response single emotion labels and distribution labels. Inspired by the experiment setup in [8], we randomly split our dataset into training/validation/test sets with the ratio of $9 : 0.5 : 0.5$.

4.1.2 *Emotional lexicon construction*

We use the combination of existing linguistic resources to construct the required emotion lexicon for our experiments. As for emotional words lexicon V_e, we extend the ECM external dictionary with emotional lexicon ontology in DLUT [47] to get a high-quality affective lexicon, since ECM dictionary only has a small number of words in several emotion classes which will limit its usage efficiency. For example, the emotion class "Angry" of the ECM dictionary only contains 39 words. We extract all the desired emotional words of major class in our categories from DLUT, and the rest words are added to the class "Other.". Finally, after combination, the new emotion lexicon contains 6 major categories and a total of 63279 words, much larger than the ECM emotion dictionary. The number of words in the extended emotion lexicon V_e is shown in Table 1. Next, we just divide V_e by polarity label in DLUT to get the lexicons of two polarities: P_l and N_l, where $P_l \cap N_l = \emptyset$.

4.1.3 *Affective embedding*

Since traditional word vectors do not explicitly consider the emotional features of words, and the outstanding performance of VAD notations is limited to English so far, we adopt a universal methodology in [49] to fit emotional information into the pre-trained classical word vectors. In our experiment, Tencent AI Lab Embedding Corpus [50] is used as the original training embedding model for our re-training. This corpus was chosen because it provides over 8 million Chinese word vector representations with a 200-dimension size of each vector that covers many absent words in other open-source word embedding corpora, such as idioms, maintains freshness, and improves accuracy. Furthermore, the Tencent Chinese embedding corpus saves non-topic words, such as stop words, negatory words and intensity words, increasing the general applicability in various scenarios. Specifically, we use our emotion lexicon and the experimental rules introduced in [47] to generate emotional word pairs and follow the positive and negative similarity constraints in [45] to do the objective training over generated emotional word pairs. Then we get the new embedding with a 200-dimension size of each vector. The compared results with original embedding are given in Figure 3(a). It is clear that the emotional similarity between two opposite emotion categories decreased and there is an obvious increase of in-category mutual similarity in emotion categories. These results demonstrate that training constraints are useful. The implementation of this Chinese emotional word embedding refinement method is open sourced as ChEmoEmb[4)].

4.2 Training details

We use a two-layer bidirectional GRU for the encoder and a uni-directional GRU for the decoder in our chatbot with 256 hidden units in each layer. We set vocabulary size as 40000. The maximum sequence

[2)] http://coai.cs.tsinghua.edu.cn/hml/challenge2017/.
[3)] https://github.com/Jiangchenglin521/nlpcc2013data.
[4)] https://github.com/Jiangchenglin521/ChEmoEmb.

Table 1 Statistics of the new emotion lexicon V_e

Class	Like	Sad	Disgust	Angry	Happy	Other
New lexicon	11107	23147	10282	388	1967	37221

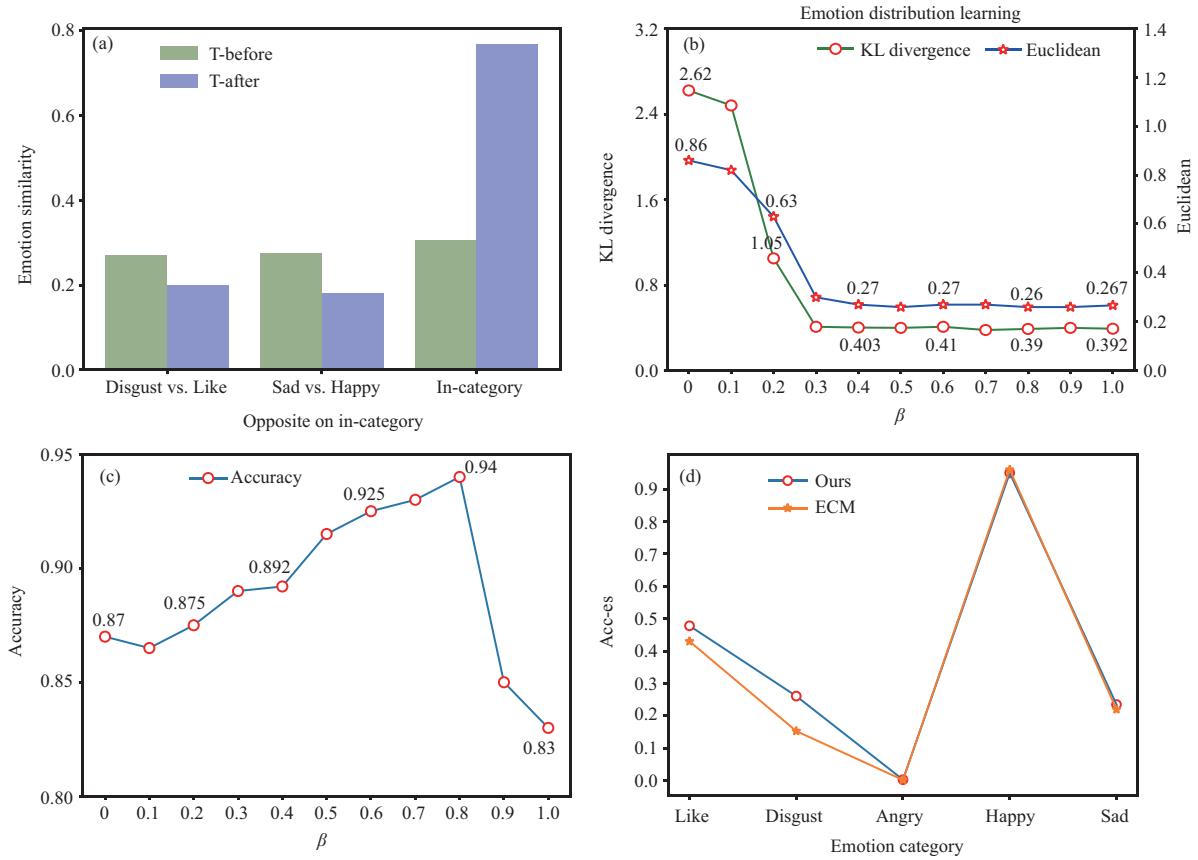

Figure 3 (Color online) (a) Embedding training results. (b) and (c) Effect of β in joint training between EDL and dominant emotion classification. (d) Validate test results comparing with original ECM.

size is set to 25 for post and response. We adopt inference repetition suppressor in ITF [48] with $\lambda = 0.2$ to avoid the word repetition in predicted responses. Then, we assign an integer score from 1 to 4 for the intensity score S of corresponding modifier classes in R to represent the intensities for our strategy. We also set $\varepsilon = 0.6$ for a dominant label. The stochastic gradient descent algorithm is chosen as the optimized algorithm with an auto-descent learning rate (lr). Initial lr is set to 0.5 with the decay factor of 0.98. We set $\gamma = 0.2$ in the weighted generating layer. The proposed chatbot is implemented in Tensorflow[5)], and it costs one week for 150-epoch training on a Titan XP GPU server.

4.3 β-balance test

As illustrated, β is a hyper-parameter used to control the weight of EDL and emotion classification loss (ECL) in joint learning. We just train our chatbot with each value in the zone $\{0, 0.1, 0.2, \ldots, 0.9, 1\}$ respectively, and choose the value with the best performance. To evaluate the performance of EDL, the similarity between a predicted distribution and the truth one is measured by the KL divergence and Euclidean, and emotion accuracy is used for ECL measurement to calculate the proportion of true predictions. The effect of β is shown in Figure 3. In both Figures 3(b) and (c), $\beta = 0$ suggests only using ECL in training SEG, and it can be clear seen from Figure 3(c) that if not using EDL, the dominant emotion accuracy is relatively low, which means that only using dominant emotion classification cannot reach the best performance. When β increases from 0 to 0.5, as can be seen in Figure 3(b), the performance of EDL increases obviously then in steady tendency till 1, and as shown in Figure 3(c), ECL performance shows obvious increase before 0.8, obtaining the best accuracy score at 0.94. These results show that

[5)] https://www.tensorflow.org/.

Table 2 Overall objective evaluation results with baselines

Model	Perplexity	BLEU	EACC	Acc-w	Acc-es
S2SA (2015)	58.04	0.10	–	0.36	0.24
ECM (2018)	76.15	0.08	–	0.43	0.353
ARS2S (2019)	50.262	0.11	–	0.37	0.257
CCM (2019)	76.15	0.08	0.89	0.38	0.349
Ours	**54.74**	**0.12**	**0.94**	**0.48**	**0.402**

ECL, which is jointly used with ECL, surely improves dominant emotion accuracy. When $\beta = 0.8$, the performance achieves the most favorable state, representing the joint learning reaches a balance. After 0.8, accuracy drops quickly due to the heavy weight of distribution loss that cannot preserve the dominant emotion. At last, $\beta = 1$ means that only using EDL in training SEG, in Figure 3(b), the accuracy score is only 0.83, which is lower than any other accuracy result, though the distribution learning gains the steady and good performance. This result proves that only using EDL cannot make SEG perform at its best, either. So, we finally choose $\beta = 0.8$ in experiment.

4.4 Baselines

We compare our chatbot with the following baseline models. First of all, we need one traditional chatting model to show the affect efficiency and expressive of the compared emotional models, so we choose S2SA, the standard sequence-to-sequence model with attention [13], which holds no emotional constraints. Then, according to the features of our new chatbot model and the illustration in related work, it is clear that we aim to achieve and evaluate the complete expression efficiency of a presented framework for emotional conversation model, rather than combine the existed method to defeat all state-of-art models directly. We just choose another two typical and problem-related emotional conversation models as the baselines. One is ECM [6], which represents the one that only has emotional expression in a controlled manner but cannot decide specific emotion upon its characters, and another one is the ARS2S [10], an affect rich model based on sequence-to-sequence model implemented with our Chinese emotional embedding. ARS2S represents another kind of existed affect conversation model, which can directly respond with emotion on its own, but cannot specify the type and rationality of the emotion in use. Furthermore, to stand out the learning efficiency and affect collaboration of the SEG in our chatbot model, we also implement a model named CCM, inspired by the MECS [2]. CCM contains an independent emotion classifier within the ECM framework. The classifier is pre-trained and directly used in ECM response generation to predict an emotion from a post-sentence. Furthermore, in addition to presenting experiments comparing the performance of different models, we conduct the necessary experiments to evaluate our proposed distribution labeling strategy for the corpus with only single emotion labels. In detail, we evaluate our EDA and PPA algorithms against CICS [25], implication constraint CICS (CICS+ic) [51] and EDA without PPA algorithm (EDA-p).

4.5 Automatic evaluation

4.5.1 *Metrics*

Perplexity, a wildly used metric in existing conversation models, is applied to evaluate grammatical accuracy and content relevance in our model. And the smoothed BLEU score [52] is adopted to give a supplement of the model evaluation in a way for considering lacking general accepted outstanding automatic metrics in the conversation generation field. As for emotion evaluation, we design three metrics: emotion accuracy (EACC), the ratio of accurate emotion expression of response comparing with the predicted emotion category used in generation; Acc-w, the percentage of generated responses that contain words in E with corresponding category in V_e; Acc-es, the percentage of responses that contain affect words with corresponding emotional specific categories which reject the label "Other," since this unspecific category may introduce uncertain influence in accuracy.

4.5.2 *Results and analysis*

Firstly, we utilize the model-level measurement to evaluate the overall emotional and semantic performance of our emotional chatbot among baselines. The results are given in Table 2. As can be seen, S2SA performs rather poorly on nearly all emotion metrics, primarily because it does not consider any

Table 3 Specific emotion word proportion score (Acc-es) with ground truth score as reference

Model	Acc-es1	Acc-es2	Acc-es3	Acc-es4	Acc-es5
Ground Truth	0.582	0.340	0.130	1.000	0.254
ECM	0.430	0.153	0.002	0.960	0.210
Ours	**0.478**	**0.261**	**0.003**	**0.950**	**0.235**

affective factor and tends to generate generic responses. Nevertheless, our chatbot achieves significant improvements on Acc-w and Acc-es over ECM, ARS2S and CCM, indicating that our chatbot can generate responses with better emotional expression. Besides, by comparing the Acc-w and Acc-es scores, we could get that label "Other" has truly introduced obvious influence in the final score. The highest EACC score of our chatbot shows that the SEG can produce an accurate emotion based on the joint learning of contextual affective relation in conversations, rather than making direct label classification like CCM model. One note is that S2SA, ARS2S, and ECM cannot predict the desired emotion used in generation, so they have no EACC score. In addition to emotional performance, our chatbot also gets a lower perplexity score and the highest BLEU score, denoting that our model successfully achieves an efficient balance between content and emotion expression in generation.

After discussing the models' overall performance, we give the specific assessment on the positive efficiency of our improvement in structure. As is known that we present an affective weighted lexicon-based decoder, it adopts the idea of inverse token frequency to stand out the selection of low-frequency emotional words in response generation, which can be seen as an improving version of ECM decoder. We then compare the specific emotional performance between ECM and our chatbot. We use the same emotion lexicon with ECM to exclude the influence of emotion dictionary size. The results are presented in Figure 3(d) with related values in Table 3. Meanwhile, we use the metric Acc-es to measure the relative performance, where a higher proportion score suggests the generated response is more likely to hold corresponding emotional words and the ground truth score is the true Acc-es score that was calculated from the validated test data. The index from 1 to 5 after Acc-es in Table 3 refers to the specific emotion category in E in order. We can get that the improvement in the decoder obtains a good income, which can recall more affect words in response than ECM, especially in categories: 1-"Like" and 2-"Disgust". With respect to the SEG in our chatbot, there should be a note that except for EACC score and β-balance test in Subsection 4.3, we further depend on human level evaluation for its performance measurement, since there is no appropriate auto metric that is suitable for the SEG evaluation to our best knowledge.

4.6 Human evaluation

We use human evaluation on the quality of generated emotion and responses to better model performance. Besides, as emphasized above, evaluation on the efficiency of SEG also depends on the human satisfaction of generated emotion for expression. Therefore, we further add a model in evaluation, which is presented as Ours-S. This model can be seen as a version of our proposed model, which removes the SEG structure. This kind of ablation experiment can verify whether SEG has played a role in rationally perceiving and selecting specific emotions for a response.

4.6.1 *Model performance*

We randomly selected 100 posts from the testing set and generated responses from all baseline models compared. To better understand the quality of generated emotion and responses, we gave posts, emotion categories, and related randomized responses to three human annotators to score from the content and emotion levels. The following are the evaluation criteria that we employ.

(1) Content level.

- 0: A response has either serious grammar error or completely irrelevant to the post.
- 1: A response has correct grammar but is not very natural or is too universal.
- 2: A response has correct grammar and also is appropriate and natural to a post.

(2) Emotion level-accuracy.

- 0: The emotional expression of a response does not match the given emotion category.
- 1: The emotional expression of a response agrees with the given emotion category.

(3) Emotion level-satisfaction.

- 0: The given emotion is not very expressive and used rarely in general conditions by considering the express content and relation with post.

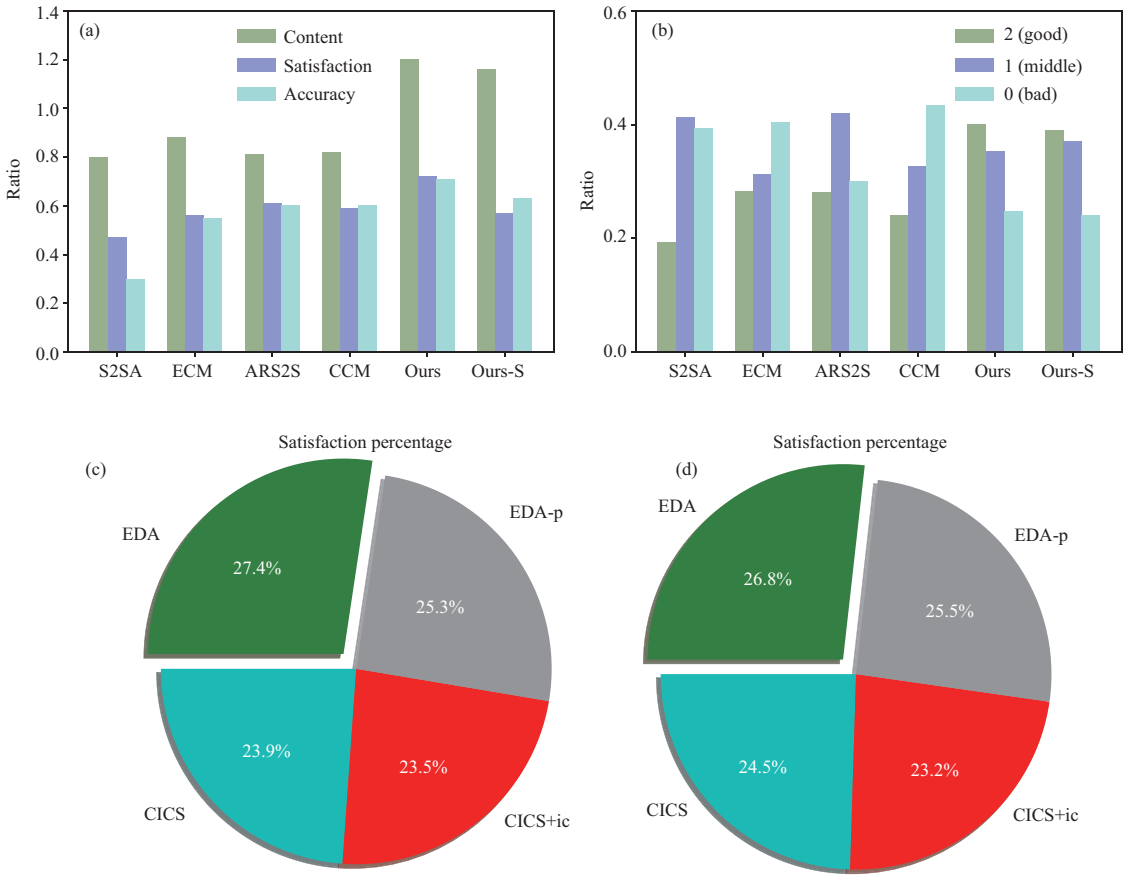

Figure 4 (Color online) Human evaluation on overall model performance and distribution labeling strategy performance. (a) Score on content and emotion performance. (b) Ratio of each score in content level evaluation, "0", "1" and "2" refer to the score of the "Content Level" indicator. (c) Distribution satisfaction of conversation corpus. (d) Distribution satisfaction of hotel comments corpus.

- 1: The given emotion is suitable and expressive for response generation.

Specifically, the first two policies, content level and emotion accuracy mainly focus on the semantic and emotional performance of the model. As for the final one, it concentrates more on the user satisfaction on the generated emotion category by the designed SEG. To make a convincing illustration, we first calculate the Fleiss' Kappa coefficient [53] to measure inter-rater consistency among three annotators. The results are 0.563, 0.624, and 0.550 for content, emotion appropriateness, and accuracy, indicating "Moderate" agreement, "Substantial" agreement, and "Moderate" agreement, respectively. These agreement degrees show the score from the three annotators is valid. The score results from three annotators are presented in Figure 4(a) and (b). As for the content level evaluation, it can be seen from Figure 4(b) that our chatbot gains the most ratings of score 2 and the lowest number of 0. As shown in Figure 4(a), our chatbot overall has the highest score 1.2 in content level. These results suggest that our chatbot obtains significant improvement on content relevance over baseline models. And in the measurement of emotion level, for the emotional satisfaction evaluation, our chatbot has the highest score at 0.72 in Figure 4(a), which suggests that our generated emotion is acceptable and more favorable than baselines. S2SA and ECM get lower scores than others, where the reason may be that they ignore modelling contextual emotional information in conversations explicitly. Note that ARS2S's score is higher than other baselines, which is also constructed without emotion specification. One possible explanation is that it generates responses via an emotional attention mechanism with affective word embedding that incorporates more emotional information. As for the performance of our proposed model without SEG (Ours-S), it obtains a lower score in satisfaction measurement than the complete model. This result shows that SEG can efficiently control the appropriate emotion in conversation. Furthermore, by comparing CCM with ours, we can obtain that the joint learning method in our affective generator works on producing a more acceptable response emotion due to our outperformance. As for emotional accuracy, our chatbot still receives the best score at 0.71, implying the designed affective decoder is more effective in generating a response with

Table 4 Sampled responses in testing set[a)]

Translated post	Model	Emotion	Translated response	Avg score
The 3D promotional video at the front theater is better than this film!	S2SA	–	Aha, me too	0
	ECM	Disgust	You think like this, but you don't know it	0.83
	ARS2S	–	A good film just like this	0.67
	CCM	Disgust	This year, prices have risen too high.	0.17
	Ours	like	**OK, fine, but I like niche performance**	**1**
Aha, Huang Ying covered this song? Its lyrics I heard when I was a child!	S2SA	–	Yeah, like it very much	0.33
	ECM	Like	Ok, fine	0.17
	ARS2S	–	Haha, heard before	0.67
	CCM	Happy	Haha, yes, me too	0.67
	Ours	like	**Yes, it's true, I love his voice either**	**1**
I thought you have met each other before	S2SA	–	Haha, you too	0.17
	ECM	Happy	Haha, I am on purpose	0.67
	ARS2S	–	I met too	0.33
	CCM	Angry	You don't come, How dare I	0.83
	Ours	Happy	**Haha, me too, I think the same to you**	**0.83**
I weighted that day, I grew fat! I'm self-abased	S2SA	–	Haha, yes	0.33
	ECM	Sad	I'm already old	0.17
	ARS2S	–	Haha, it's inferiority	0.67
	CCM	Sad	I have been tortured by you.	0.33
	Ours	Sad	**I think so, it's sad I am too thin now**	**0.83**

a) Avg score shows the user satisfaction with both emotional appropriateness and content relevance.

the specific emotion categories.

4.6.2 *Algorithm evaluation*

To ensure the efficiency of our distribution labeling strategy, we specifically measure the quality of the soft emotional distribution label generated by our strategy among existing methods. We randomly sample 100 responses and comments with labels from the conversation test set and a hotel comment corpus[6)] that contains long affective comments of hotels, respectively, since the conversation sequence is short with limit affect. For the assigned dominant weight ε in each approach, we set it all as 0.6. Three annotators score on generated distributions from all compared methods following our designed protocol.

Distribution satisfaction.

- 0: The weight of each emotion in distribution is allocated without clear distinction.
- 1: The weight of each emotion in distribution is allocated well to the real emotional intensity expressed in a given sentence.

In this evaluation process, the Fleiss' Kappa value is 0.71, indicating "Substantial" agreement among annotators. The results in Figures 4(c) and (d) show that our EDA algorithm outperforms other methods on distribution satisfaction, even in the conversation data with low affective information. Besides, it gains the highest satisfactory proportion. In detail, considering the lower scores of CICS and CICS+ic, we could get the necessity of the context emotional influence in one sentence. The intensity and polarity always depend on multiple elements, such as several negative words or the intensity words, not only the emotional words. Besides, EDA-p that ignores the polarity features introduces little satisfaction than EDA, indicating that our PPA algorithm provides effective polarity influence for distribution labeling in EDA, where PPA properly guarantees the appropriate emotional tendency of our algorithm.

4.7 Case study

To present a more intuitive view of our chatbot's performance, we provide some sampled generation examples over the testing set (in domain) in Table 4 to show the better emotion expressing ability of our chatbot upon assigned characters. In Table 4, the values in the last column cover average scores from the aforementioned human evaluation protocols with the same weight, where score "1" refers to the best score. And the column emotion presents the specific emotion used for response generation. In detail, when doing generation, S2SA and ARS2S have no specific emotion input. ECM uses the ground-truth

[6)] https://www.aitechclub.com/data-detail?data id=29.

Table 5 Sampled responses of our chatbot in out-domain set

Translated post	Gnerated emotion	Model	Translated response
I really want to eat cakes, I haven't eaten for a long time	Happy	**Ours**	Aha, sure, I want to eat either!
You're too bad, I don't want to be with you	Angry	**Ours**	You're a also a bad man !
Tell the truth, you really make me feel sick	Sad	**Ours**	but I am a very miserable person.
I am ill, so sick and uncomfortable ...	Sad	**Ours**	Me too, I am sorry and sad

label, and CCM classifies an emotion from the post as the input, and ours uses the emotion generated by SEG. As a result, our chatbot can respond to the desired emotion with the highest average satisfaction score. There is also a surprising finding: even when responses are with no obvious affective words, our response still shows great affective satisfaction. Furthermore, to give an observable view of our chatbot's robustness, we also give several examples on out-domain set in Table 5, where the posts in this set are coming from real scenes and never appear in the training set.

5 Conclusion and future work

We discussed the issue of "incomplete" emotional expression in this paper. In most existing large-scale emotional conversation generation models, "incomplete" expression refers to the case where emotional chat agents are unable to control the specific emotion manner explicitly and fully on their own. Then, we present a new generating framework to address the "incomplete" expression problem. Concretely, we construct an effective emotional chatbot. It has a joint learning-based module SEG to select a proper emotion category for response generation. A weighted lexicon-based layer is further designed in the decoder to generate final affective responses with emotion words related to the specific emotion given by SEG. Extensive experimental results show that our new chatbot performs favorably on both content coherence and user satisfaction against other emotional dialog models. It successfully achieves complete control over both high-level emotion expression categories and fine-grained emotion features in response generation.

Considering that it is the first step toward creating an emotional chatbot in the new form, further explorations are still desired. Therefore, in our future work, we will explore the explicit emotional expression mode in real world interaction based on contextual information in multiturn chatting data.

Acknowledgements This work was supported in part by National Key Research and Development Program of China (Grant No. 2019YFF0302601).

References

1 Huang C, Zaiane O R. Generating responses expressing emotion in an open-domain dialogue system. In: Proceedings of International Conference on Internet Science, 2019. 100–112
2 Zhang R, Wang Z Y, Mai D C. Building emotional conversation systems using multi-task Seq2Seq learning. In: Proceedings of Natural Language Processing and Chinese Computing, 2018. 612–621
3 Picard R W. Affective Computing. Cambridge: The MIT Press, 1997
4 Ochs M, Pelachaud C, Sadek D. An empathic virtual dialog agent to improve human-machine interaction. In: Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, 2008. 89–96
5 Prendinger H, Ishizuka M. The empathic companion: a character-based interface that addresses users' affective states. Appl Artif Intell, 2005, 19: 267–285
6 Zhou H, Huang M L, Zhang T Y, et al. Emotional chatting machine: emotional conversation generation with internal and external memory. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, 2018. 730–738
7 Colombo P, Witon W, Modi A, et al. Affect-driven dialog generation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019. 3734–3743
8 Song Z Q, Zheng X Q, Liu L, et al. Generating responses with a specific emotion in dialog. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 3685–3695
9 Shen L, Feng Y. CDL: curriculum dual learning for emotion-controllable response generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020. 556–566
10 Zhong P X, Wang D, Miao C Y. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In: Proceedings of the AAAI Conference on Artifficial Intelligence, 2019. 33: 7492–7500
11 Sayan G, Mathieu C, Eugene L, et al. Affect-LM: a neural language model for customizable affective text generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, 2017. 634–642
12 Lin Z J, Xu P, Winata G I, et al. CAiRE: an end-to-end empathetic chatbot. In: Proceedings of the AAAI Conference on Artifficial Intelligence, 2020. 34: 13622–13623
13 Gao B B, Xing C, Xie C W, et al. Deep label distribution learning with label ambiguity. IEEE Trans Image Process, 2017, 26: 2825–2838
14 Marsella S, Gratch J. Computationally modeling human emotion. Commun ACM, 2014, 57: 56–67
15 Ma Y K, Nguyen K L, Xing F Z, et al. A survey on empathetic dialogue systems. Inf Fusion, 2020, 64: 50–70

16 Busemeyer J R, Dimperio E, Jessup R K. Integrating emotional processes into decision-making models. In: Integrated Models of Cognitive Systems. Oxford Scholarship Online, 2007. 213–229
17 McTear M F, Callejas Z, Griol D. The Conversational Interface. Berlin: Springer Publishing Company, Incorporated, 2016
18 Wang Z X, Ho S B, Cambria E. A review of emotion sensing: categorization models and algorithms. Multimed Tools Appl, 2020, 79: 35553–35582
19 Savery R, Weinberg G. A survey of robotics and emotion: classifications and models of emotional interaction. In: Proceedings of the 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN), 2020. 986–993
20 Soleymani M, Lichtenauer J, Pun T, et al. A multimodal database for affect recognition and implicit tagging. IEEE Trans Affective Comput, 2012, 3: 42–55
21 Cambria E, Fu J, Bisio F, et al. AffectiveSpace 2: Enabling Affective Intuition for Concept-level Sentiment Analysis. Austin: AAAI Press, 2015. 508–514
22 Lee G G, Kim H K, Jeong M, et al. Natural Language Dialog Systems and Intelligent Assistants. Berlin: Springer Publishing Company, Incorporated, 2015
23 Ma Y K, Peng H Y, Khan T, et al. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. Cogn Comput, 2018, 10: 639–650
24 Peng H Y, Ma Y K, Li Y, et al. Learning multi-grained aspect target sequence for Chinese sentiment analysis. Knowledge-Based Syst, 2018, 148: 167–176
25 Zhang Y X, Fu J M, She D Y, et al. Text emotion distribution learning via multi-task convolutional neural network. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018. 4595–4601
26 Prendinger H, Mori J, Ishizuka M. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. Int J Human-Comput Studies, 2005, 62: 231–245
27 Marcin S. Affect listeners: acquisition of affective states by means of conversational systems. In: Development of Multimodal Interfaces: Active Listening and Synchrony. Berlin: Springer, 2010. 169–181
28 Fitzpatrick K K, Darcy A, Vierhile M. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): a randomized controlled trial. JMIR Ment Health, 2017, 4: e19
29 Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems, 2014. 3104–3112
30 Dziri N, Kamalloo E, Mathewson K W, et al. Augmenting neural response generation with context-aware topical attention. In: Proceedings of the 1st Workshop on NLP for Conversational AI, 2019. 18–31
31 Hancock B, Bordes A, Mazare P-E, et al. Learning from dialogue after deployment: feed yourself, chatbot! In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019. 3667–3684
32 Wu Y, Wei F R, Huang S H, et al. Response generation by context-aware prototype editing. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2019. 7281–7288
33 Qian Q, Huang M L, Zhao H Z, et al. Assigning personality/profile to a chatting machine for coherent conversation generation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence Main track, 2018. 4279–4285
34 Zhong P X, Zhang C, Wang H, et al. Towards persona-based empathetic conversational models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020. 6556–6566
35 Mehrabian A. Pleasure-arousal-dominance: a general framework for describing and measuring individual differences in Temperament. Curr Psychol, 1996, 14: 261–292
36 Zhong P X, Wang D, Li P F, et al. CARE: commonsense-aware emotional response generation with latent concepts. In: Proceedings of the AAAI Conference on Artificial Intelligence, 2021. 35: 14577–14585
37 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In: Proceedings of the International Conference on Learning Representations, 2015. 1–15
38 Poria S, Majumder N, Mihalcea R, et al. Emotion recognition in conversation: research challenges, datasets, and recent advances. IEEE Access, 2019, 7: 100943–100953
39 Rolls E T, Ekman P, Perrett D I, et al. Facial expressions of emotion: an old controversy and new findings. Phil Trans Roy Soc Lond B, 1992, 335: 63–69
40 Cho K, Merrienboer B V, Bahdanau D, et al. On the properties of neural machine translation: encoder-decoder approaches. In: Proceedings of the 8th Workshop on Syntax, Semantics and Structure in Statistical Translation, 2014. 103–111
41 Seyeditabari A, Tabari N, Gholizadeh S, et al. Emotion detection in text: focusing on latent representation. 2019. ArXiv:1907.09369
42 Zhou P, Shi W, Tian J, et al. Attention-based bidirectional long short-term memory networks for relation classification. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, 2016. 207–212
43 Gross J J. The emerging field of emotion regulation: an integrative review. Rev General Psychol, 1998, 2: 271–299
44 Hochschild A R. Emotion work, feeling rules, and social structure. Am J Sociol, 1979, 85: 551–575
45 Alam F, Danieli M, Riccardi G. Annotating and modeling empathy in spoken conversations. Comput Speech Language, 2018, 50: 40–61
46 Luong M T, Pham H, Manning C D. Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, 2015. 1412–1421
47 Xu L H, Lin H F, Pan Y, et al. Constructing the affective lexicon ontology (in Chinese). J China Soc Sci Tech Inf, 2008, 2: 180–185
48 Nakamura R, Sudoh K, Yoshino K, et al. Another diversity-promoting objective function for neural dialogue generation. In: Proceedings of AAAI 2019 Workshop on Reasoning and Learning for Human-Machine Dialogues (DEEP-DIAL 2019), 2019
49 Seyeditabari A, Tabari N, Gholizadeh S, et al. Emotional embeddings: refining word embeddings to capture emotional content of words. 2019. ArXiv:1906.00112
50 Song Y, Shi S M, Li J, et al. Directional skip-gram: explicitly distinguishing left and right context for word embeddings. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018. 2: 175–180
51 Yang J F, She D Y, Sun M. Joint image emotion classification and distribution learning via deep convolutional neural network. In: Proceedings of the 26th International Joint Conference on Artificial Intelligence Main track, 2017. 3266–3272
52 Chen B X, Cherry C. A systematic comparison of smoothing techniques for sentence-level BLEU. In: Proceedings of the 9th Workshop on Statistical Machine Translation, Baltimore, 2014. 362–367
53 Fleiss J L. Measuring nominal scale agreement among many raters. Psychol Bull, 1971, 76: 378–382