Applied Intelligence (2022) 52:7218–7229
https://doi.org/10.1007/s10489-021-02819-z

Incorporating emotion for response generation in multi-turn dialogues

Yanying Mao[1] · Fei Cai[1] · Yupu Guo[1] · Honghui Chen[1]

Accepted: 2 September 2021 / Published online: 21 September 2021

Abstract

Generating semantically and emotionally context-consistent responses is key to intelligent dialogue systems. Previous works mainly refer to the context in the dialogue history to generate semantically related responses, ignoring the potential emotion in the conversation. In addition, existing methods mainly fail to consider the emotional changes of interlocutors and emotional categories simultaneously. However, emotion is crucial to reflect the interlocutor's intent. In this paper, we propose an Emotion Capture Chat Machine (ECCM) that is able to capture the explicit and underlying emotional signal in the context to generate appropriate responses. In detail, we design a hierarchical recursive encoder-decoder framework with two enhanced self-attention encoders to capture the semantic signal and emotional signal, respectively, which are then fused in the decoder to produce the response. In general, we consider the dynamic and potential information of emotion to generate the response in multi-turn dialogues in the field of both daily conversation and psychological counseling. Our experimental results on a daily Chinese conversation dataset and a psychological counseling dataset show that ECCM outperforms the state-of-the-art baselines in terms of Perplexity, Distinct-1, Distinct-2, and manual evaluation. In addition, we find that ECCM performs well for input contexts with different lengths.

Keywords Response generation · Multi-turn dialogues · Emotional response · Psychological counseling

1 Introduction

Response generation (RG) in dialogue system refers to generating an appropriate response for a given input. Its application covers various scenarios, such as intelligent assistants, in-vehicle systems, smart homes, etc., which greatly improves the quality of user experience. However, users prefer to treat the dialogue model as a spiritual partner, rather than a software for performing tasks [11]. Therefore, generating the corresponding emotional response according to user's emotional needs is the key to personalize the dialogue system.

Prior works on response generation for multi-turn dialogue system have achieved satisfied performance. For instance, [33] first utilize an attention mechanism to perform a weighted summation of the hidden layer state at each moment which is served as the input of the Intention RNN model. Serban et al. [21] further apply a hierarchical model to perform context encoding and decoding for response generation. After that, [22] introduce a Gaussian random variable in the Context RNN model to improve the diversity of generated responses. The works listed above mainly fail to consider the emotional information in multi-turn dialogues, which we argue can release user's emotional needs to help generate an anthropomorphic response. However, addressing the emotion factor in the dialogue system faces the following challenges. First of all, it is difficult to obtain the high-quality sentiment annotations in dialogue datasets as sentiment classification is a subjective and complicated task [19]. In addition, how to balance grammar and expressions of emotions in sentence is not well studied as getting semantically coherent responses is challenging enough [5].

To deal with it, [5] apply the emotional labels and intensity for word prediction which is capable of generating responses according to the determined emotional strategy. Considering emotions in a natural and coherent way, [34] propose an emotional chatting machine (ECM) for automatically generating utterances based on different emotional

✉ Fei Cai
caifei08@nudt.edu.cn

[1] Science and Technology on Information Systems Engineering Laboratory, National University of Defense Technology, Changsha City, Hunan Province, China

tags. Inspired by ECM, [29] incorporate an emotion selector to ECM that can automatically generate a unique response to the post. Besides, [35] design an integrated model with a retrieval component and a generative component to construct an open-domain emotional dialogue system. Such solution is limited to the single-turn dialogue, which ignores the previous conversational context. In multi-turn dialogues, the interlocutors' emotions will dynamically change according to the content of the conversation, which cannot be expressed simply by using the emotional labels in the above models. In addition, the emotion is typically hidden in the semantics, e.g., the rhetorical questions and irony, is not well developed in the conversation. Therefore, we argue that a good response should take into account the potential emotional needs of the interlocutor.

Thus, in this paper, we present a novel Emotion Capture Chat Machine (ECCM), which is able to capture the emotions in each utterance, and the emotions are then combined with the dialogue context to generate appropriate responses. In detail, we extend the hierarchical recursive encoder-decoder architecture [21] and divide the response generation process into three main stages. First, we apply two parallel self-attention-enhanced encoders to separately embed the semantic information and emotional information of each utterance in the context, which leads to an individual semantic vector and an emotional vector of each utterance. We then produce the semantic vector and the emotional vector of the context by aggregating the sentences in the context. Then, we combine the corresponding semantic vector and the emotional vector of the context and the current utterance to update the hidden states, which are then fused and input to the decoder for generating an emotional category vector. Finally, we integrate the emotional category vector and a semantic vector of the current utterance generated by a re-encoder, which is then sent to the decoder for generating the final response. Compared with the existing model, we add the treatment of potential emotions in ECCM, so that it can generate appropriate emotional response in multi-turn dialogue system. This feature also enables ECCM to help patients automatically in the field of psychological counseling.

We conduct comprehensive experiments on the public Douban Conversation Corpus[1] and the Emotional First Aid Dataset[2]. The former is a daily dialogue dataset and the latter is a psychological counseling dialogue dataset. The experimental results show that ECCM outperforms the state-of-the-art baselines in terms of Perplexity, Distinct-1, Distinct-2 and human evaluation, which demonstrates its effectiveness of ECCM for generating emotional responses. In addition, we find that ECCM achieves the best performance in the setting of different lengths. In summary, the contributions of this paper are listed as follows:

- To the best of our knowledge, we are the first to consider the dynamic and potential information of emotion in multi-turn dialogues to generate appropriate emotional responses.
- We propose an innovative generation model, i.e., ECCM, that follows a hierarchical recursive encoder-decoder architecture and fulfills the response generation in the form of three main stages.
- The experimental results on an emotional dialogue dataset and a daily dialogue dataset indicate that our proposal outperforms the state-of-the-art baselines in terms of Perplexity, Distinct-1, Distinct-2 and human evaluation.

  In particular, on the emotional dialogue dataset, ECCM presents a 32.5% improvement in terms of Distinct-2 over the best baseline.

2 Related work

In this section, we review related work from two angles: general response generation models and emotional response generation models.

2.1 General response generation models

Generally, the response generation technology in dialogue system is mainly classified into three categories, i.e., retrieval-based, template-based and deep-learning-based approaches [8, 20, 23]. In recent years, profited from the development of social media, corpora of dialogue system have been enriched. Therefore, people have also begun to study dialogue models based on deep learning. Compared with other models, deep-learning based dialogue models are more flexible and creative. In detail, the task of response generation can be simplified to the input-to-output mapping problem, and the dialogue is time-series [8, 20, 23], which can be regarded as a sequence, so the Sequence-to-sequence model (Sequence-to-sequence Model, Seq2seq) [25] based on the end-to-end framework is very suitable as a model for response generation. For the first time, [27] apply Seq2seq framework in dialogue generation, both the encoder and decoder adopt the long short term memory network (Long Short Term Memory Network, LSTM) [7].

However, conversation processes are usually continuous and dynamic, and the current dialogue context, that is, historical dialogue information, needs to be referred when generating a reply [9, 30, 32]. Instead of utilizing the recurrent neural network (RNN) in encoder, [24] design

[1] The dataset is available at https://github.com/codemayq/chinese_chatbot_corpus

[2] The dataset is available at https://www.52nlp.cn/efaqa-corpus-zh

a multi-turn dialogue model with the multilayer feed-forward neural network, which encodes historical dialogue context and current utterance simultaneously. [21] propose a hierarchical neural network (Hierarchical Neural Network, HNN) to address the encoding problem of context. The model includes three modules, where Encoder-RNN is used to encode input sentences, and Context-RNN uses information at the dialogue level to encode context vectors, so as to guide Decoder-RNN to generate the final reply content. In order to obtain higher-quality responses, it is necessary to consider some common sense and background knowledge besides combining the context. Aiming at this goal, [10] contributes a dynamic memory network (Dynamic Memory Network, DMN) to leverage both historical dialogue information and existing background knowledge when generating responses. However, too much historical information and background knowledge will not only complicate calculation but also increase noise. Therefore, we use self-attention mechanism [13] to extract important historical dialogue information, and re-encode the last utterance of context as the basis for generating a semantically coherent response. Moreover, we also consider the emotional information, which is always ignored by general models.

2.2 Emotional response generation models

In recent years, some researchers have realized that emotion factors are of great significance in terms of successfully establishing human-like dialogue generation models [6, 15]. [5] propose Affect Language Model to generate text conditioned on context words and given affect categories. The method of artificially adding emotional information cannot be applied to the product, so [1] present three novel ways to incorporate emotional aspects into LSTM that improve the open-domain conversational prowess of encoder-decoder networks. Compared with LSTM, Seq2seq is more suitable for solving the problem of generating responses of different lengths [12, 29, 34]. Among them, ECM [34] is the first work that addresses the emotion factor in large-scale conversation generation. Compared with ECM, EACM (Emotion-aware Chat Machine) [29] adds an emotion selector that enables the model to automatically generate a unique affective response to the post. In E-SCBA model (Syntactically Constrained Bidirectional-Asynchronous Approach for Emotional Conversation Generation) [12], both logic and emotion of replies are improved when pre-generated emotion keywords and topic keywords are asynchronously introduced into the process of decoding. Generally, current emotional response generation methods cannot fit the multi-turn dialogues and they often ignore the changes of user's emotion. In addition, the potential emotional information that appears in context needs to be explicitly modeled for generating the most appropriate emotional responses. In contrast, we dedicate to propose a dialogue model based on multi-turn dialogues, which can capture the dynamic and potential information of emotion in the context and fuse them with semantic information into the response generation process.

3 Approach

3.1 Overview

Figure 1 presents an overview of the proposed Emotion Capture Chat Machine (ECCM) for response generation. The model consists of three main components, i.e., self-attention enhanced encoder (see Section 3.2), context encoder and fusion module (see Section 3.3), and emotional response generator (see Section 3.4).

We first explain in detail the task of response generation in multi-turn dialogues. We view a dialogue D as a sequence of $M + 1$ utterances $D = \{U_1, \cdots, U_{M+1}\}$ between two speakers. We treat D as two parts (C, W), where $C = \{U_1, \cdots, U_M\}$ represents the conversation context, and W denotes the target response U_{M+1}. Each utterance of context U_m $(m = 1, 2, \cdots, M)$ contains a sequence of tokens of arbitrary length N_m.

Thereby, the task of response generation in multi-turn dialogues is to compute the probability $P(W \mid C)$ of generating a response W based on the given conversation context C.

In this article, our task is to capture the emotion involved in the context of the conversation and fuse them into the generation process to automatically generate semantically coherent and emotionally appropriate response. First, we use two parallel self-attention enhanced encoders to process each word in utterance U_m to obtain the sentence-level semantic vector h_m^{utts} and emotional vector h_m^{utte}. Next, the context layer updates the information of a sequence of sentences to context-level, i.e., semantic vector h^{dlg} and emotional vector h^{emo}. Then they are sent into the fusion module to generate a emotion category vector e_r' for generating responses. We assume that the target response W consists of a sequence of T words, i.e., $W = (w_1, \cdots, w_T)$. Finally, the emotional response generator completes the final response word-by-word based on the hidden state x of U_M and the vector e_r':

$$w_t = arg\max P\left(w_t' \mid w_{1:t-1}, x, e_r'\right), \quad (1)$$

where w_t' is the word to be generated at step t. It should be noted that the end symbol $\langle EOS\rangle$ is added at the end of each sentence when encoding sentences. In the decoder, when the probability distribution points to the end symbol $\langle EOS\rangle$, the generation process is completed.

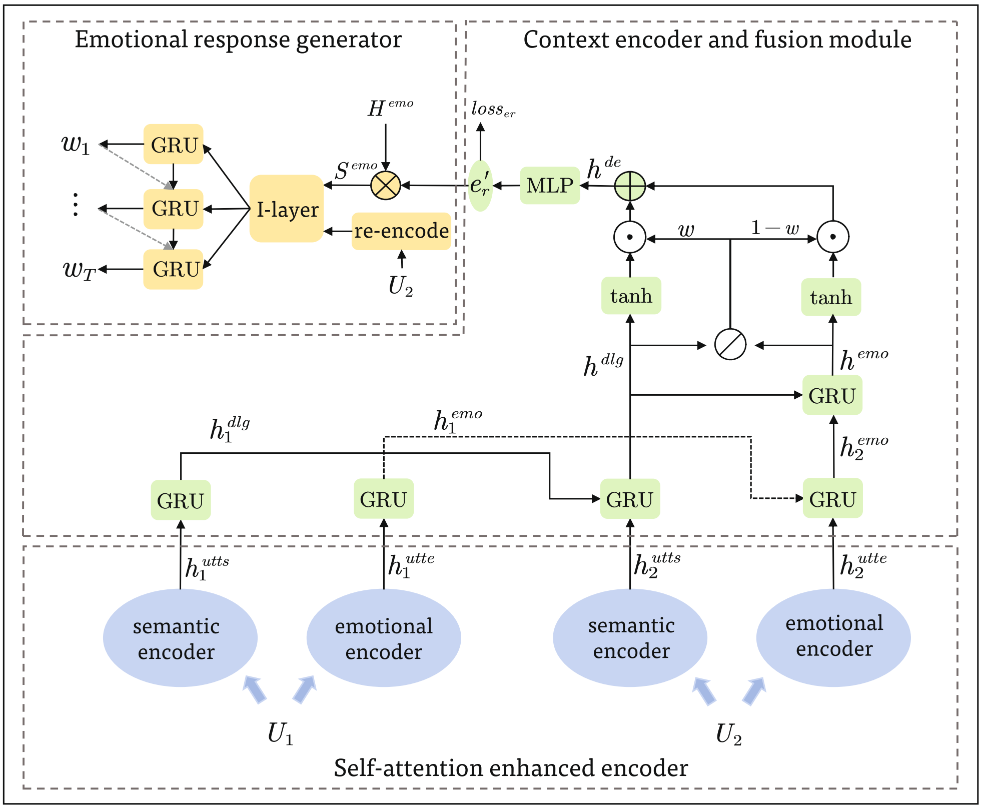


Fig. 1 Overview of the Emotion Capture Chat Machine (ECCM)

3.2 Self-attention enhanced encoder

In order to retain emotional and semantic information to the greatest extent, we use an emotion encoder and a semantic encoder in the encoder RNN independently. The emotional and semantic information of the utterance are encoded separately, and are then sent to the context RNN. Specifically, we employ a GRU network in both encoders to extract the semantics $w_{m,n}^s$ and the emotion $w_{m,n}^e$ from the utterance $U_m = \{w_{m,1}, \cdots, w_{m,N_m}\}$, and the words in each sentence are entered one by one in order. Here, $w_{m,n}^s$ and $w_{m,n}^e$ respectively embed the semantic and emotional information of the word.

Figure 2 takes an input utterance U_m as an example to illustrate how these two encoders work. For simplicity, we detail the emotion encoder in Fig. 2b as it is structurally identical to the semantic encoder. We input a sequence of word embedding $w_{m,n}^e$, and the hidden representations $h_m^e = (h_{m,1}^e, h_{m,2}^e, \cdots, h_{m,N_m}^e)$ are produced by applying a GRU as follows.

$$h_{m,n}^e = GRU\left(h_{m,n-1}^e, w_{m,n}^e\right), \tag{2}$$

where $h_{m,n}^e$ is the hidden state of n-th word in the utterance U_m.

Then, to enhance the expressive ability of the feature vectors, we apply the self-attention mechanism [2, 13] to enable the emotion encoder to capture the emotional change in the sentences. Here, the emotional hidden state h_m^{utte} is defined as:

$$h_m^{utte} = \sum_{n=1}^{N_m} a_{m,n}^e h_{m,n}^e, \tag{3}$$

where $a_{m,n}^e$ is the weight of hidden state $h_{m,n}^e$ calculated by:

$$a_{m,n}^e = softmax\left(V_a tanh\left(W_a\left(h_{m,n}^e\right)^T\right)\right), \tag{4}$$

where we utilize a multi-layer perceptron with a softmax layer to process $h_{m,n}^e$ and let all the weights sum up to 1; V_a and W_a are the weight matrices.

As simply using the explicit emotion information is not enough to simulate the emotional interactions between humans, we consider the crucial implicit emotion hidden in the semantics contained in the conversations. For example, considering the input is “I am so sad that my dog got lost”, the interlocutor needs to be comforted rather than immersed in sadness incessantly. Obviously, the semantic meaning of context plays a vital role. Therefore, we not only consider the emotion of context, but take into account its semantic information by an independent encoder (i.e., semantic encoder) for response generation. Similarly, we can get a weighted sum of hidden states as h_m^{utts}.

$$h_m^{utts} = \sum_{n=1}^{N_m} a_{m,n}^s h_{m,n}^s, \tag{5}$$

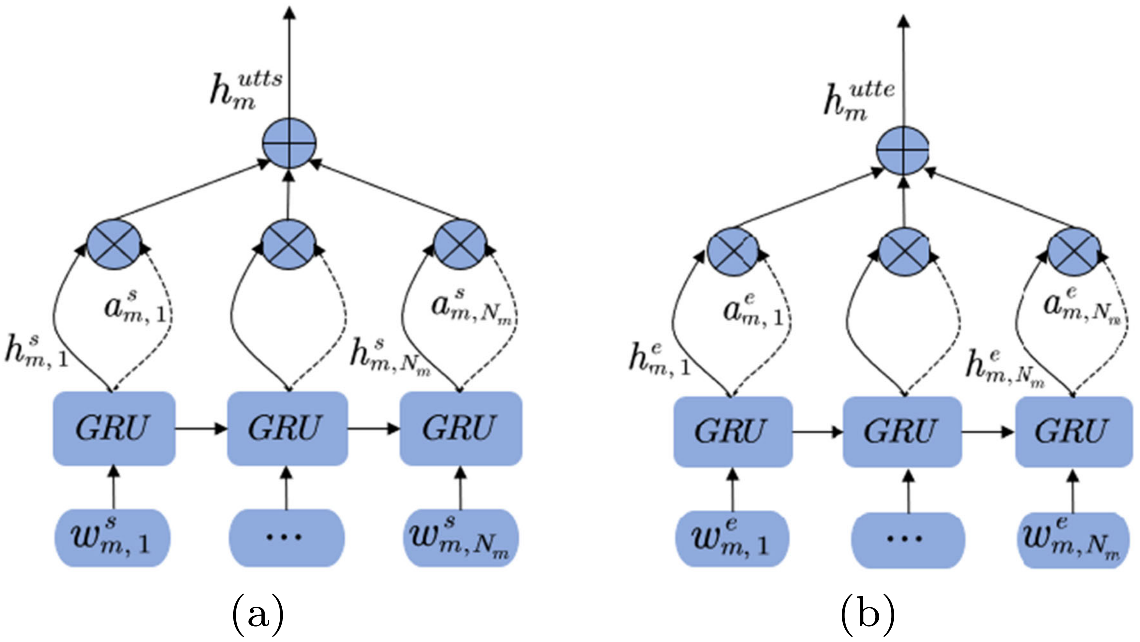


Fig. 2 Details of the self-attention Enhanced Encoder

Considering that the emotion information of a given input utterance may not be monotonous, we use the emotion embedding to the emotion encoder and expect it to fully express different aspects of emotion. In particular, according to [16, 26], we apply the sentiment-specific word embedding *SSWE* for emotion representation and the *word2vec* embedding for semantic representation in our paper. It should be noted that we use *SSWE* and *word2vec* because they are classic and efficient in the word embedding. And in the training process we also find that the training results of the *word2vec* method are the best and the training speed of the *word2vec* is the fastest.

3.3 Context encoder and fusion module

3.3.1 Context encoder

After encoding the utterance to generate h_m^{utts} and h_m^{utte}, they are then fed into the context encoder to model the sequence of dialogue turns to produce the dialogue context hidden representation h^{dlg} and h^{emo}, which corresponds to the semantic signal and the emotional signal, respectively, as

$$h_m^{dlg} = GRU\left(h_m^{utts}\right), \qquad (m = 1), \tag{6}$$

$$h_m^{dlg} = GRU\left(h_{m-1}^{dlg}, h_m^{utts}\right), \quad (1 < m \le M), \tag{7}$$

$$h^{dlg} = h_M^{dlg}. \tag{8}$$

In practice, verbal expressions are not all straight-forward, and it is inevitable to encounter challenging contexts such as rhetorical question and irony. In this case, sentiment-specific vocabulary is insufficient for the model to perceive the exact emotion of the sentence, or even completely deviate. For instance, considering that we input “Do you really think quarreling is a happy thing?”, the chatbot will answer “Yes”, but it is evident that the speaker is waiting for mollifying by the semantic analysis. So in order to capture the emotion hidden in the semantics, we embed the emotion of the context h^{emo} constrained to h^{dlg}:

$$h_m^{emo} = GRU\left(h_m^{utte}\right), \qquad (m = 1), \tag{9}$$

$$h_m^{emo} = GRU\left(h_{m-1}^{emo}, h_m^{utte}\right), \quad (1 < m \le M), \tag{10}$$

$$h^{emo} = GRU\left(h_M^{emo}, h^{dlg}\right), \tag{11}$$

3.3.2 Fusion module

Next, following [29], we construct a fusion network to balance the contribution derived from the semantic and emotional information, and send the mixed information to a selection network for determining the response emotion e_r. The obtained h^{dlg} and h^{emo} are first concatenated and then fed to a sigmoid layer to calculate the weight of semantics:

$$w = \sigma\left(\left[h^{dlg}; h^{emo}\right]\right), \tag{12}$$

$$h_\star^{dlg} = \tanh\left(h^{dlg}\right), \tag{13}$$

$$h_\star^{emo} = \tanh\left(h^{emo}\right), \tag{14}$$

$$h^{de} = w \odot h_\star^{dlg} + (1 - w) \odot h_\star^{emo}, \tag{15}$$

where σ is an activation function, $\odot$ means an element-wise multiplication, h^{de} represents the weighted sum of the contextual semantic information and the emotional information. Then we feed h^{de} into a selection network to yield an emotional category vector for generator as:

$$e_r' = \sigma\left(W_r h^{de} + b\right), \tag{16}$$

$$loss_{er} = -e_r \log\left(e_r'\right), \tag{17}$$

where e_r' is the ultimate emotional-inclined vector for generating response produced by the selection network. Note that e_r is the multi-hot representation of response emotion category. We get e_r' by sending h^{de} into MLP and mapping it to the probability distribution on the emotion categories. In summary, this pivotal module links the encoder and generator, which determines the best emotion category to response generation by taking into account the semantic and emotional information transmitted by the encoder.

3.4 Emotional response generator

Considering to construct a generator that is capable of blending the affective part with semantic part, we first convert e_r' as follows:

$$S^{emo} = H^{emo} e_r', \tag{18}$$

where the matrix H^{emo} represents the underlying emotional factors derived from previous work [18]. As semantic deviation exists in the generation of multi-turn dialogues, following [28], we re-encode the latest input $U_M = \{w_{M,1}, \cdots, w_{M,N_M}\}$ to get a sequence of semantic hidden states $x = (x_1, x_2, \cdots, x_N)$ to I-layer for generating the context vector c_t and the current hidden state s_t. The specific calculation process of the I-layer is described as follows:

$$s_t = GRU\left(W_4\left[s_{t-1}; c_{t-1}\right], \left[w_{t-1}; S^{emo}\right]\right), \tag{19}$$

$$z_t^n = v^T \tanh\left(W_1 x_n + W_2 s_t + W_3 S^{emo}\right), \tag{20}$$

$$\alpha_t^n = softmax\left(z_t^n\right), \tag{21}$$

$$c_t = \sum_{n=1}^{N} \alpha_t^n x_n, \tag{22}$$

where the W_1, W_2, W_3 and W_4 are learnable parameters. To obtain the important semantics, we re-weight each hidden states x_n by applying an attention mechanism and then adding the key emotion representation S^{emo}. The context vector c_t contains the pivotal information at step t, which makes our model perform well in decoding long sentences. We resort to a GRU network in our generator where w_{t-1} stands for the word generated at the previous step t-1, i.e., $U_{M+1} = W = \{w_1, w_2, \cdots, w_T\}$. To train our model, the final loss function is a weighted sum of the semantic loss and the emotion loss defined as:

$$loss_{sr} = -\sum_{t=1}^{T} \log P\left(w_t | w_1, w_2, \cdots, w_{t-1}, c_t, S^{emo}\right), \tag{23}$$

$$loss\left(\theta\right) = \alpha loss_{sr} + (1-\alpha)\, loss_{er}, \tag{24}$$

where θ denotes the parameter set and α is a balance factor.

4 Experiments

In this section, we detail our experimental setup. We focus on three research questions as follows:

RQ1: Can our proposed Emotion Capture Chat Machine (ECCM) perform better than the baselines for response generation?

RQ2: Are the responses generated by our model anthropomorphic?

RQ3: How is the impact of the context length on the model performance for response generation?

4.1 Datasets

We conduct experiments on two public conversational datasets, i.e., the Douban Conversation Corpus [31] and the Emotional First Aid Dataset. Douban Conversation Corpus contains 1.1 million dyadic dialogues (i.e., conversation between two persons) longer than two turns from Douban group which is a popular social networking service in China. Emotional First Aid Dataset, which is the first open QA corpus in the field of psychological consultation, includes 20,000 pieces of psychological consultation data and is the largest publicly available Chinese psychological consultation dialogue material. The datasets are rich in content, with detailed annotations such as discourse status (negative or positive) and interlocutor information (consultant or client).

As the raw text does not have emotional tags, we refer to the previous work [34] to append the emotion labels to the responses in the training sets. Following [34], we set six emotion categories, i.e., *(happy, sad, angry, disgust, fear, and other)*, where *other* represents the absence of any emotional information. It is worth noting that the emotion categories which are not common in daily conversations are removed, and the category "fear" is retained because it appears frequently in the Emotional First Aid Dataset. Then we extract the former three-turn dialogues between two interlocutors for our experiments where the third utterance is the target response U_{m+1}. Finally, we obtain 486,975 triples from Douban Conversation Corpus and 17,954 triples from Emotional First Aid Dataset. In addition, for each dataset, 200 triples corresponding to each emotion are randomly selected for validation and test, leading to 1,200 triples for each partition. The remainder is used for training. The statistics of the datasets we use are shown in Table 1. We find that the ratio of emotional dialogue data in the Emotional First Aid Dataset is higher than that in the Douban Conversation Corpus.

4.2 Baselines

We compare the performance of our model against three state-of-the-art baselines.

HRED [21], a hierarchical encoder-decoder model that introduces an additional context encoder to model the interactive structure of multi-turn dialogues[3].

Emo-HRED [14], a fully data driven chat-oriented dialogue system that can dynamically mimic affective human interactions by considering the emotion information at the context level[4]. After training with a dialogues corpus that contains positive-emotion eliciting responses, this model produces responses with more positive emotion.

EACM [29], an emotion-aware chat machine that can produce emotional responses based on a single utterance[5]. Due to the lack of context, there is a significant decrease in semantic coherence and syntactic accuracy. Therefore, for comparison, we use the previous utterances, i.e., the context, as the input of EACM.

4.3 Evaluation metric

As emotional influence cannot be evaluated automatically, we perform objective evaluation and human evaluation to verify the effectiveness of our model. Following [14, 17], we first use three metrices, i.e., Perplexity, Distinct-1 and Distinct-2, to objectively evaluate the performance of the proposed method and the baselines. Then, we use a human evaluation to measure the semantic coherence and emotional appropriateness of the generated responses.

Objective evaluation

(1) Perplexity explicitly measures the model's ability to account for the syntactic structure of the dialogue and the syntactic structure of each utterance [21].

The smaller the Perplexity is, the better the language model performs. The mathematical definition of Perplexity is as follows:

$$Perplexity = \sqrt[N]{\prod_{i=1}^{N} \frac{1}{p(w_i|w_1 w_2 \cdots w_{i-1})}}, \tag{25}$$

where $p(w_i)$ is the probability of the i-th word. The first word is $p(w_1|w_0)$, and w_0 is a placeholder.

(2) Distinct-1 and Distinct-2 are often used to measure the responses in terms of diversity by computing the number of distinct unigrams and bigrams in the generated responses, respectively.

A higher Distinct score indicates the more abundant vocabulary contained in the response. The mathematical definitions of Distinct-1 and Distinct-2 are as follows:

$$Distinct-1 = \frac{count(distinct_{w_i \in R}(w_i))}{count(all_{w_i \in R}(w_i))}, \tag{26}$$

$$Distinct-2 = \frac{count(distinct_{w_i w_{i+1} \in R}(w_i w_{i+1}))}{count(all_{w_i w_{i+1} \in R}(w_i w_{i+1}))}, \tag{27}$$

where, R represents all generated results on the test set, $distinct()$ indicates that all repetitions are removed, and $all()$ represents all results, $count$ represents the number of statistics.

Human evaluation We randomly select 100 triples from the full test set and ask two human judgers to compare the responses based on two criteria, i.e., the semantic coherence and the emotional appropriateness, which corresponds to that whether the response is syntactically smooth as well as logically contextual and whether the emotion in response conforms to the context, respectively. In order to ensure the validity of the results, these two judgers did not know which model the response was generated from during the evaluation process. We divide the evaluation index into five levels from 5 to 1, representing *strongly agree, agree, not necessarily, disagree, strongly disagree*, respectively.

4.4 Implementation details

In our model, we use the GRU network with 256 hidden units for encoding and decoding, and the parameters are different in each layer. The embedding size of word and emotion category are both set to 200, which are randomly initialized. We choose the 40,000 most common words from Douban Conversation Corpus as our vocabulary, and all out of vocabulary words are replaced with a special token *UNK*. In order to generate different responses, we use the beam search in the decoding process with the beam size set to 20. We adopt the stochastic gradient descent (SGD) algorithm [3] with mini-batch during training, and the batch size and learning rate are set to 128 and 0.5, respectively. Following

[3] https://github.com/hsgodhia/hred

[4] https://github.com/nlsskysn/emoHRED

[5] https://github.com/CCIIPLab/EACM

Table 1 The statistics of Douban conversation corpus and emotional first aid dataset

Dataset	Types of emotion						Validation	Test	Overall
	happy	sad	angry	disgust	fear	other			
Douban Conversation Corpus	99,186	74,522	62,091	63,251	11,904	173,621	1,200	1,200	486,975
Emotional First Aid Dataset	1,073	6,147	1,537	935	4,473	1,389	1,200	1,200	17,954

[34], we leverage a well-trained Seq2seq model to get the initialization parameters and thus the training speed is accelerated. The setting process about the balance factor α is as follows: we first observe the order of magnitude when the two parts of the loss function are close to convergence and set the benchmark weight. For example, a part of the loss function is about 0.2, and the other part is about 0.02, then the baseline weight is set to 1:10. Then on this basis, the weight is subjected to a fine adjustment to achieve the best results. To ensure a fair comparison, we use the best tuning methods reported in their articles for the baselines.

5 Results and discussion

5.1 Overall performance

To answer **RQ1**, we investigate the accuracy and diversity of the responses generated by ECCM and the baselines in terms of Perplexity; Distinct-1 and Distinct-2. The test results on the Douban Conversation Corpus and Emotional First Aid Dataset are shown in Table 2. Generally, our method ECCM achieves the best performance in terms of all metrics on both datasets, which indicates its effectiveness on generating the emotional responses in multi-turn dialogues. Particularly, the improvement of ECCM over the best performing baseline in terms of Perplexity is statistically significant at level $p < 0.05$. Next, we zoom in on the results on the Douban Conversation Corpus. We find that the Perplexity scores of Emo-HRED, ECAM, and ECCM are significantly better than that of HRED, which can be due to that the addition of emotional information is beneficial to improve the syntactic structure of responses. In addition, Emo-HRED performs significantly better than EACM in terms of Perplexity, indicating the responses generated by Emo-HRED are more semantically fluent and contextual than EACM. It means that adding the context information can make the responses generated by the corresponding model more reasonable. We can obtain similar results on the Emotional First Aid Dataset. It is worth mentioning that all models perform better on Emotional First Aid Dataset than the Douban Conversation Corpus. This may be attributed to the fact that richer semantic and emotional information are contained in the Emotional First Aid Dataset than that in the Douban Conversation Corpus as a relatively large percentage of high-quality emotional dialogues is observed on the Emotional First Aid Dataset.

In addition, in order to make the comparison between the models more intuitive, we count the number of parameters, model size and training time of different models. As shown in Table 3, we can find that the smallest model is EACM with the minimum number of parameters and the shortest training time. This indicates that EACM is the simplest model, possibly because it is the only one applied to the single-turn conversation and does not compute context-level information. The other three models are applied to multi-turn dialogue. Obviously, Emo-HRED is more complex than HRED, and ECCM is more complex than Emo-HRED. This

Table 2 Objective evaluations of different response generation models

Dataset	Model	Perplexity	Distinct-1	Distinct-2
Douban Conversation	HRED	187.46	0.0204	0.0537
	Emo-HRED	87.11	0.0335	0.0923
	EACM	103.72	0.0496	0.1771
	ECCM	**69.27***	**0.0643***	**0.2346**
Emotional First Aid	HRED	191.28	0.0185	0.0365
	Emo-HRED	79.73	0.0371	0.0992
	EACM	112.51	0.0554	0.2004
	ECCM	**65.88***	**0.0696**	**0.2894**

The results produced by the best baseline and the best performing model in each column are underlined and boldfaced, respectively; statistical significance of pairwise differences of ECCM against the best baseline (*) is determined by t-test ($p < 0.05$)

Table 3 The complexities of models

Model	Number of parameters	Size of model	Training time
HRED	4.1M	3.91MB	6.2h
Emo-HRED	4.5M	4.29MB	6.7h
EACM	3.9M	3.72MB	5.9h
ECCM	4.8M	4.58MB	7.5h

Training time refers to the training speed of each model on the Douban Conversation Corpus

can be attributed to the addition of emotional information to ECCM and Emo-HRED. Among them, Emo-HRED divides emotions only into positive and negative, while ECCM divides emotions into specific categories. Although the complexity of ECCM is the highest, it can be inclusive according to experimental results. For example, in terms of number of parameters ECCM is 23.1%, 17.1%, and 6.7% more than EACM; HRED and Emo-HRED respectively; but in terms of Perplexity index ECCM is 33.2%, 63.0%, and 20.5% more optimised than EACM, HRED, and Emo-HRED, respectively.

5.2 Human evaluation

To answer **RQ2**, we focus on a comparison between ECCM and the baselines in terms of semantic coherence and emotional appropriateness. We present the results in Table 4 and observe that EACM performs the worst in terms of semantic coherence, as it is difficult to learn the contextual information. In contrast, HRED performs the worst in terms of emotional appropriateness as it does not take the emotional information into account. For the baselines, Emo-HRED performs best in terms of semantic coherence, EACM performs best in terms of emotional appropriateness. Apparently, ECCM gets the highest scores in terms of both metrics, which indicates that the responses generated by ECCM are more anthropomorphic than the baselines. This may be due to the ability of ECCM to separately encode the emotional and semantic information of multi-turn dialogues that is fused into the decoding process. Accordingly, ECCM benefits from injecting the emotion signal into the response generation process.

Table 4 Human evaluation

Model	Semantic coherence	Emotional appropriateness
HRED	3.07	2.24
Emo-HRED	3.46	3.36
EACM	2.97	3.51
ECCM	**3.69**	**3.73**

5.3 Impact of context length

To answer **RQ3**, we analyze the performance of ECCM, Emo-HRED and HRED on the test sets by varying the context lengths. Here, we only present the experimental results on the Douban Conversation Corpus as similar phenomena can be found on the Emotional First Aid Dataset. As shown in Table 5, we divide the test set into four categories according to the context length. The majority of the tests have a context length between 20 and 60 words, which allows the model to learn enough contextual information for generating responses. We use the objective evaluation criteria (i.e., Perplexity, Distinct-1, and Distinct-2) to measure the model performance, and the results are shown in Fig. 3.

It can be found from Fig. 3 that ECCM outperforms the baselines in terms of all metrics at each context length, which confirms the robustness of ECCM across samples with different context lengths. In particular, ECCM performs prominently when the context is longer than 60 words. However, ECCM and the baselines perform similarly in terms of Perplexity when the context length is smaller than 20. This indicates that all the models discussed have a similar ability to deal with short contexts. In addition, as the context length increases, the scores of all models monotonously increase. This means that when the length of the context increases, it is increasingly hard for dialogue machines to balance the accuracy and diversity in the process of generating responses. It can be attributed to the fact that long contexts not only bring rich information to evaluate the comprehending capacity of dialogue machines, but inject noise to affect the generation process.

5.4 Case study

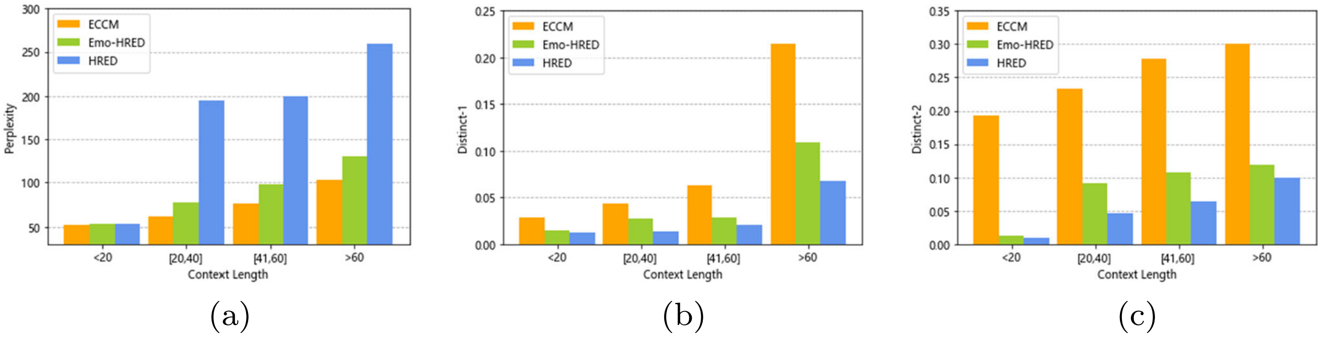
To clearly illustrate the effectiveness of our proposal, we conduct a case study by selecting three cases from the datasets we use in our experiments; see Fig. 4. Case 1 is taken from Douban Conversation Corpus; Case 2 and Case 3 are selected from Emotional First Aid Dataset.

Case 1 In this case, the input content is ordinary daily conversation without obvious emotional expression. Results

Table 5 Ratio of samples with different context length in the test set of Douban Conversation Corpus

Context length	<20	[20,40]	[41,60]	>60
Ratio	10.37%	41.63%	39.48%	8.52%

The context length equals the number of the input words



(a) (b) (c)

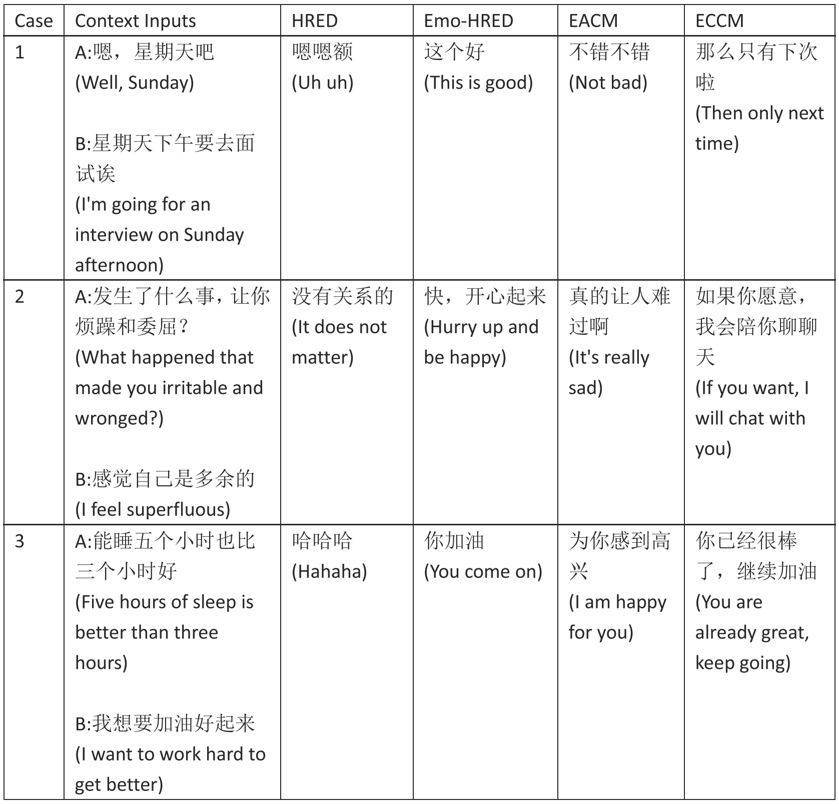
Fig. 3 Experimental results of different context lengths

generated by HRED and EACM are meaningless; the response produced by Emo-HRED is out of context; only EACM generates a suitable response. From the aspect of emotion, ECCM appropriately expresses regret about not being able to date, even if there are no obvious emotional messages in context. This obviously demonstrates that our method has the ability to learn potential emotional information from the conversations.

Case 2 The input in this case is a conversation in psychological counseling, and the client's emotion is negative. Clearly, the responses generated by other models except HRED are contextual. In detail, the response produced by EACM has a strong emotional (sad) resonance with the consultant, which is missing in Emo-HRED in contrast. Furthermore, the emotional expression of ECCM is considered to be more appropriate, while comforting interlocutor can also avoid the proliferation of negative emotions, which is very practical in psychological counseling.

Case 3 In this example, the client's mood is changed from negative to positive. We see that HRED fails to generate a complete sentence; on the other hand, EACM fails to achieve the semantic coherence. The generated responses by Emo-HRED and ECCM have a good balance between semantics and emotion, however, the response from ECCM is more attractive than Emo-HRED.

Fig. 4 Case study



Case	Context Inputs	HRED	Emo-HRED	EACM	ECCM
1	A:嗯，星期天吧 (Well, Sunday) B:星期天下午要去面试诶 (I'm going for an interview on Sunday afternoon)	嗯嗯额 (Uh uh)	这个好 (This is good)	不错不错 (Not bad)	那么只有下次啦 (Then only next time)
2	A:发生了什么事，让你烦躁和委屈？ (What happened that made you irritable and wronged?) B:感觉自己是多余的 (I feel superfluous)	没有关系的 (It does not matter)	快，开心起来 (Hurry up and be happy)	真的让人难过啊 (It's really sad)	如果你愿意，我会陪你聊聊天 (If you want, I will chat with you)
3	A:能睡五个小时也比三个小时好 (Five hours of sleep is better than three hours) B:我想要加油好起来 (I want to work hard to get better)	哈哈哈 (Hahaha)	你加油 (You come on)	为你感到高兴 (I am happy for you)	你已经很棒了，继续加油 (You are already great, keep going)

6 Conclusions and future work

In this paper, we propose the Emotion Capture Chat Machine (ECCM) to generate the emotional responses in multi-turn dialogues. To the best of our knowledge, we are the first to consider the dynamic and potential information of emotion in multi-turn dialogues to generate appropriate emotional response. In particular, we utilize two parallel self-attention-enhanced encoders to separately embed the semantic and emotional information of each utterance in the context. Experimental results on two publicly available datasets show the effectiveness of our model, presenting that ECCM outperforms the baselines at both semantic-level and emotional-level, in terms of Perplexity, Distinct-1, Distinct-2, and human evaluations. In addition, our model is robust for context with different lengths. As to future work, it is necessary to consider long dialogue history to generate context-specific responses [14]. Besides, we would like to explore the topic-based emotional response generation and expand the applicable scenarios of chat machines [4].

Acknowledgements We are especially grateful to participants for completing the human evaluation.

Compliance with Ethical Standards

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

References

1. Asghar N, Poupart P, Hoey J et al (2018) Affective neural response generation. In: Advances in Information Retrieval - 40th European Conference on Research, pp 154–166. https://doi.org/10.1007/978-3-319-76941-7_12
2. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations. arXiv:1409.0473
3. Bottou L, Curtis FE, Nocedal J (2018) Optimization methods for large-scale machine learning. SIAM Rev 60:223–311. https://doi.org/10.1137/16M1080173
4. Dziri N, Kamalloo E, Mathewson KW et al (2018) Augmenting neural response generation with context-aware topical attention. CoRR arXiv:1811.01063. https://doi.org/10.18653/v1/w19-4103
5. Ghosh S, Chollet M, Laksana E et al (2017) Affect-lm: Neural language model for customizable affective text generation. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp 634–642. https://doi.org/10.18653/v1/P17-1059
6. Gonċalves VP, Costa EP, Valejo A et al (2017) Enhancing intelligence in multimodal emotion assessments. Appl Intell 46:470–486. https://doi.org/10.1007/s10489-016-0842-7
7. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput:1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
8. Ji Z, Lu Z, Li H (2014) An information retrieval approach to short text conversation. CoRR arXiv:1408.6988
9. Khandelwal U, He H, Qi P, Jurafsky D (2018) Sharp nearby, fuzzy far away: How neural language models use context. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, pp 284–294. https://doi.org/10.18653/v1/P18-1027
10. Kumar A, Irsoy O, Ondruska P et al (2016) Ask me anything: Dynamic memory networks for natural language processing. In: Proceedings of the 33nd International Conference on Machine Learning, pp 1378–1387. arXiv:1506.07285
11. Li H, Wen G (2019) Sample awareness-based personalized facial expression recognition. Appl Intell 49:2956–2969. https://doi.org/10.1007/s10489-019-01427-2
12. Li J, Sun X (2018) A syntactically constrained bidirectional-asynchronous approach for emotional conversation generation. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp 678–683. https://doi.org/10.18653/v1/d18-1071
13. Lin Z, Feng M, dos Santos CN et al (2017) A structured self-attentive sentence embedding. In: 5th International Conference on Learning Representations. https://openreview.net/forum?id=BJC_jUqxe
14. Lubis N, Sakti S, Yoshino K et al (2018) Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp 5293–5300. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16317
15. Mayer JD, Salovey P (1997) What is emotional intelligence? Emotional Development and Emotional Intelligence, pp 3–31. https://psycnet.apa.org/record/1997-08644-001
16. Mikolov T, Chen K, Corrado G et al (2013) Efficient estimation of word representations in vector space. In: 1st International Conference on Learning Representations. http://arxiv.org/abs/1301.3781
17. Pietquin O, Hastie HF (2013) A survey on metrics for the evaluation of user simulations. Knowl Eng Rev 28:59–73. https://doi.org/10.1017/S0269888912000343
18. Plutchik R (1980) A general psychoevolutionary theory of emotion. In: Theories of emotion, pp 3–33. https://doi.org/10.1016/C2013-0-11313-X
19. Poria S, Majumder N, Mihalcea R et al (2019) Emotion recognition in conversation: Research challenges, datasets, and recent advances. IEEE Access 7:100943–100953. https://doi.org/10.1109/ACCESS.2019.2929050
20. Ritter A, Cherry C, Dolan WB (2011) Data-driven response generation in social media. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp 583–593. https://aclanthology.org/D11-1054/
21. Serban IV, Sordoni A, Bengio Y et al (2016) Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, pp 3776–3784. http://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/view/11957
22. Serban IV, Sordoni A, Lowe R, et al. (2017) A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, pp 3295–3301, http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14567
23. Shang L, Lu Z, Li H (2015) Neural responding machine for short-text conversation. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, pp 1577–1586. https://doi.org/10.3115/v1/p15-1152
24. Sordoni A, Galley M, Auli M et al (2015) A neural network approach to context-sensitive generation of conversational responses. In: The 2015 Conference of the North American Chapter of the Association for Computational

Linguistics, Human Language Technologies, pp 196–205. https://doi.org/10.3115/v1/n15-1020

25. Sutskever I, Vinyals O, Le QV (2014) Sequence to sequence learning with neural networks. CoRR arXiv:1409.3215
26. Tang D, Wei F, Yang N et al (2014) Learning sentiment-specific word embedding for twitter sentiment classification. In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pp 1555–1565. https://doi.org/10.3115/v1/p14-1146
27. Vinyals O, Le QV (2015) A neural conversational model. CoRR arXiv:1506.05869
28. Vinyals O, Kaiser L, Koo T et al (2015) Grammar as a foreign language. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing System, pp 2773–2781. arXiv:1412.7449
29. Wei W, Liu J, Mao X et al (2019) Emotion-aware chat machine: Automatic emotional response generation for human-like emotional interaction. In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management, pp 1401–1410. https://doi.org/10.1145/3357384.3357937
30. Wu X, Du Z, Guo Y et al (2019) Hierarchical attention based long short-term memory for chinese lyric generation. Appl Intell 49:44–52. https://doi.org/10.1007/s10489-018-1206-2
31. Wu Y, Wu W, Xing C et al (2017) Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, pp 496–505. https://doi.org/10.18653/v1/P17-1046
32. Xing C, Wu Y, Wu W et al (2018) Hierarchical recurrent attention network for response generation. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp 5610–5617 https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16510
33. Yao K, Zweig G, Peng B (2015) Attention with intention for a neural network conversation model. CoRR arXiV:1510.08565
34. Zhou H, Huang M, Zhang T et al (2018) Emotional chatting machine: Emotional conversation generation with internal and external memory. In: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, pp 730–739. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/16455
35. Zhuang Y, Wang X, Zhang H et al (2017) An ensemble approach to conversation generation. In: Natural Language Processing and Chinese Computing - 6th CCF International Conference, pp 51–62. https://doi.org/10.1007/978-3-319-73618-1_5

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

Yanying Mao is now pursuing the master degree in management science and engineering at the National University of Defense Technology, Hunan, China. She received her B.S. degree in Marine Technology from Dalian University of Technology, Liaoning, China, in 2019.

Fei Cai is an assistant professor at the National University of Defense Technology, Changsha, China. He got his Doctor degree on Computer Science from the University of Amsterdam under the supervision of Prof. Maarten de Rijke. His research interests include information retrieval and query formulation. He has several papers published in SIGIR, CIKM, FnTIR, TOIS, TKDE, etc. In addition, he serves as a PC member for CIKM and WSDM as well as a reviewer for SIGIR, WWW, WSDM, CIKM, TKDE, IPM, JASIST, etc.

Yupu Guo is now pursuing the doctor degree in management science and engineering at the National University of Defense Technology, Hunan, China. He received the M.S. degree in information system engineering from the National University of Defense Technology, Hunan, China, in 2021.

Honghui Chen is a professor in the National University of Defense Technology, Hunan, China. He got his Ph.D. degree in Operational Research from the National University of Defense Technology, Hunan, China, in 2007.