Front. Comput. Sci.

RESEARCH ARTICLE

Generating Empathetic Responses through Emotion Tracking and Constraint Guidance

Jing Li[1], Donghong Han(✉)[1,2], Zhishuai Guo[3], Baiyou Qiao[1], Gang Wu[1]

1 School of Computer Science and Engineering, Northeastern University, Shenyang 110000, China

2 Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education, Northeastern University, Shenyang 110000, China

3 Neusoft Corporation, Shenyang 110000, China

Abstract The goal of empathetic dialogue generation is to generate relevant and fluent responses with reasonable emotional expressions, which is closer to human daily communication. However, the existing empathetic dialogue generation models ignore the continuity of parties' emotional expression in adjacent dialogue turns, resulting in inadequate emotional perception. Moreover, the emotions involved in empathetic response are flexible, it is difficult to set the specific empathetic policy. We propose the Emotion-Tracking Hierarchical Recurrent Empathetic Encoder Decoder (ETHREED) model for multi-turn dialogues to address the above problems in this paper. In order to improve the model's ability of emotion perception, we exploit the GRUs to process time sequences and design a hierarchical structure to track parties' emotions in the dialogue. For empathetic policy learning, we construct a stochastic policy network and introduce the constraint-based guided policy search method, so that the model can learn the patterns of human emotional interaction and retain the exploration characteristic at the same time. Experiments on the public Empathetic Dialogue dataset show that our model can generate more diverse and empathetic responses.

Keywords empathetic dialogue generation; hierarchical structure; stochastic policy network

Received month dd, yyyy; accepted month dd, yyyy

E-mail: xxx@xxx.xxx

1 Introduction

The goal of dialogue systems is to have effective and natural interactions between humans and machines. To achieve this goal, more and more studies on dialogue generation focus on the key factors that influence dialogues, such as commonsense knowledge, personalized information, and emotions. Empathy refers to the ability to understand the inner experience of others and perceive their emotions [1]. And the empathetic dialogue system enhances the dialogue agent's ability to perceive and express emotions [2]. Empathetic dialogue can bring machines closer to humans, moreover, the emotional feedback will make people feel understood and drive a deeper talk. Figure 1 shows examples of the empathetic dialogue when the speaker feel afraid and excited, the listener's responses express reasonable emotions and context-related contents.

Empathetic dialogue generation(EDG) task is proposed by Hannah Rashkin [3]. Since then, some empathetic dialogue datasets, such as EDOS [4] and PEC [5] have been conducted successively. Currently, the relevant research mainly focuses on two aspects: the first is to improve models' ability to perceive emotion states [6–9], and the second is to explore the empathetic dialogue policies [10–14]. For one thing, the speakers' emotions directly influence the emotions of listeners' empathetic responses. Speaker and listener are collectively called party in this paper, and the party's emotion is

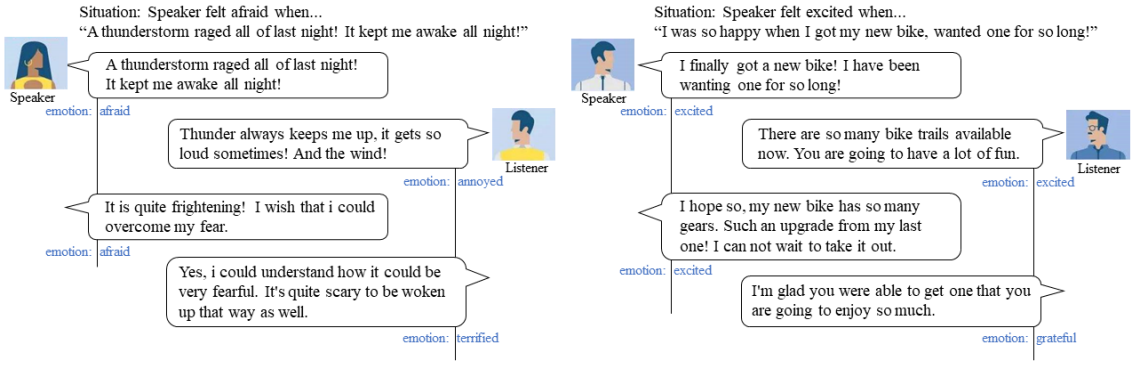

Fig. 1 Examples of the Empathetic Dialogue

often affected by many factors and continuously varied in multi-turn dialogues. It is essential and challenging to accurately capture the emotional representations. Following the example in Figure 1, parties engage in dialogue for one situation, in which parties' emotions tend to be continuous (like the speaker), or shift toward positive or negative (like the listener) depending on context. Meanwhile, the listener's (or speaker's) emotion is influenced by the speaker (or listener) and himself. Previous works promote emotion perception by external knowledge [6], emotional attention [15], emotion prediction head [16] and so on. However, above methods splice dialogue histories and utilize dialogue state embedding to distinguish parties. It handles all contextual utterances in the same way which ignores the continuity of different parties' emotional expressions in multi-turn dialogues. As a result, the model can not fully perceive the emotions of parties, which may reduce the quality of empathetic responses. For another, unlike the emotional dialogue given responses' emotion labels [17,18], the empathetic dialogue system needs to decide by itself what emotion to reply with, so the empathetic policy is indispensable. Although we treat the responses in the dataset as the golden responses, the empathetic responses are flexible, that is the appropriate response emotions and utterances are not unique. It makes the design of empathetic policies which consider the responses' flexible expression more difficult.

To address the above challenges, we investigate the multi-turn dialogue characteristics. In order to be more sensitive to the parties' emotions and improve responses' performance, we propose the Emotion-Tracking Hierarchical Recurrent Empathetic Encoder Decoder (ETHREED) model based on the typical Hierarchical Recurrent Encoder Decoder (HRED) [19]. First of all, inspired by the dialogue emotion recognition model DialogueRNN [20], we introduce four Gated Recurrent Units (GRUs) [21] to track the global state, the party state, the emotional representation and the content representation in dialogues respectively. The global GRU tracks context to obtain the global state containing dialogue history information. The party GRU updates one party state based on the all information relevant to the current utterance and another party state. The emotion GRU extracts two parties' emotional representation respectively. The content GRU obtains the dialogue content of current turn based on the global state to guide response generation and mitigate the impact of emotion perception errors on the generation task. In this way, we exploit the hierarchical recurrent structure to better model the interaction between the parties and track the emotions of them concurrently. Secondly, we consider that sharing others' emotion states in empathy can be regarded as the transfer of emotion between humans, and the empathetic policy is this transition process between the speaker's emotion and the listener's emotion. For the policy's flexible, we construct a stochastic policy network to learn this process. The policy network predicts the listener's emotional representation based on the speaker's emotional representation. We hope to learn the empathetic policy from the data directly instead of setting the specific reward function, and retain the reinforcement learning's exploratory, so we exploit the constraint-based guided policy search (GPS) method [22] combining the ideas of reinforcement learning and imitation learning. Specifically, we treat the listener's true response emotion distribution as the constraint to guide the prediction of the listener's emotional representation. At last, in order to maintain good contextual relevance, we conduct a pointer generation network, in which we use an emotional attention mechanism to dynamically incorporate the predicted listener's emotional representation at each step of the decoding process.

In summary, our contributions conclude:

1. We propose a novel EDG model ETHREED based on the HRED structure. Our model conducts four different GRUs to comprehensively obtain the emotional representation of the parties and content representation of the dialogue, which improves the emotional perception ability of the model.

2. We construct a stochastic policy network to learn the empathetic policy, and utilize the constraint-based guided policy search method to optimize it. This preserves the flexibility of reinforcement learning when training on real data, which makes the generated responses more diverse.

3. The experimental results on the Empathetic Dialogue dataset show that the responses generated by ETHREED are more diverse and empathetic.

2 Related work

In the study of EDG, models can be divided into lightweight models [10–12] and large language models (LLM) [16, 23–26]. Lightweight models can support and complement emerging techniques like LLMs, with easier training conditions and faster training. The excellent performance of LLMs comes largely from a large amount of training data and a large number of model parameters, so our method is based on lightweight model. The Recurrent Neural Network (RNN) [27] and Transformer [28] are typical encoder structures. The previous works based on transformer [6, 10–12, 15] often splice the dialogue histories and use the attention to attract required feature from the whole context. It's not easy to extract the utterance features of different parties from the long context feature. Apart from the transformer, HRED proposed by Serban et al. [19] extract the feature of every utterance and transmits information through the hierarchical RNN structure. Subsequently, Serban et al. [29] improve HRED by adding a Gaussian random variable to promote the diversity of responses. Our model focuses on modeling the constant interaction between parties and the temporal features of their emotion flow in the dialogue. The RNN-based units are advantageous in handling temporal features and have better results in extracting local features. Therefore, our model is based on HRED.

Improving the model's ability to perceive emotions can help generate more empathetic responses. Li et al. [6] combine coarse-grained dialogue-level emotions and fine-grained token-level emotions to perceive the speaker's emotion states. Li et al. [15] construct an emotional context graph by multi-type knowledge and use emotional cross attention to learn emotional signals. Shen er al. [8] conduct emotion consensus in dialogue and use unpaired data to train. Zandie et al. [16] add two additional tasks, the next utterance prediction and the next emotion prediction, to GPT2 [30] to generate empathetic responses. Lubis [31] proposes Emo-HRED which adds another RNN to HRED to obtain emotional representation for emotion recognition and as emotion bias during decoding. Previous studies mostly utilize the dialogue state embedding to simply distinguish different parties and extract the emotional representation from the whole context. They ignore the connection between parties and the continuous emotional changes in successive dialogue turns. In this paper, we exploit GRUs to model the interaction of two party states and track their emotion flows in the dialogue which can make model more sensitive to parties' emotions.

Many studies explore the policy for the speaker's current situation and emotion to make the responses more empathetic. Without the response's emotion label for training, Lin et al. [10] propose the MoEL which contains multi-decoders for different emotions and softly combines the output. Majumder et al. [11] perform emotion mimicry for positive and negative emotions, and combine different representations for generation. Gao et al. [12] utilize the gated attention to focus on emotion causes while generating responses. Xie et al. [32] firstly consider modeling emotion regulating intents in EDG. Saha et al. [33] incorporate emotion and intent to an hierarchical transformer network. Chen et al. [2] use CVAE [34] to acquire diverse response intents, and combine explicit and implicit intents to guide generation. In addition, since the empathetic dialogue is an interactive process, the reinforcement learning can be applied to it. Shin et al. [35] propose looking ahead method, setting the reward function according to the speaker's emotional changes to optimize the model. However, the method in [35] requires more turns of interaction and the reward function of empathy is difficult to accurately define. Instead of setting the specific reward function, we introduce GPS [22] to utilize real responses' emotions to guide the policy network search and predict responses' emotional representations.

3 Methology

3.1 Model Overview

There are two parties (the speaker and the listener) in the dialogue. Given the context of multi-turn dialogue $C = \{P_1, R_1, P_2, R_2, \ldots, P_t\}$, where P_i and R_i denote the speaker's post

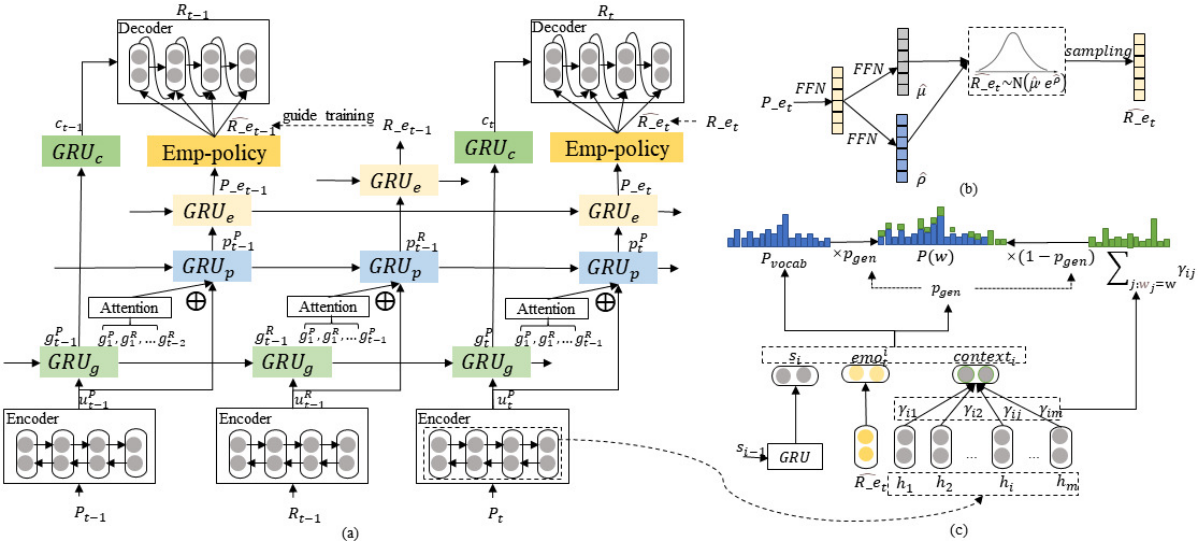

Fig. 2 The related diagrams of ETHREED, including (a) the overall architecture, (b) the policy network and (c) the pointer generation network .

and the listener's empathetic response respectively. $P_i, R_i = \{w_1, w_2, \ldots, w_m\}$ contains m words. The emotion labels of both parties in the multi-turn dialogue are represent by $Emo = \{e_1^P, e_1^R, e_2^P, e_2^R, \ldots, e_t^P\}, \forall e_i^P, e_i^R \in \{1, \ldots, n\}$, $n = 32$ in our used dataset. Our task is to understand the dialogue content and track the emotions involved in C, and then generate an empathetic response R_t. The overall structure of ETHREED is shown in Figure 2 (a).

ETHREED is a multi-turn dialogue-oriented hierarchical structure, which transfers the dialogue information and tracks the parties' emotions by multiple GRUs. The model distinguishes the speaker's posts and the listener's responses, i.e., the emotion states of both are extracted and tracked respectively. Moreover, the speaker's emotion is used to predict the response's emotion, and the listener's emotion is used to guide the policy network search. Specifically, ETHREED can be divided into three parts: emotional representation extraction, empathetic policy learning, and decoding, the detailed methods are described below.

3.2 Emotional Representations Extraction

In order to encode utterances P_i, R_i, we make use of the GloVe word embeddings [36] to get $Ew(P_i)$, $Ew(R_i)$ at first. Next, we employ Bi-GRU for textual feature extraction, then obtain the hidden state $h_i = [\overrightarrow{h_i}, \overleftarrow{h_i}]$ of each word w_i, $\overrightarrow{h_i}$ and $\overleftarrow{h_i}$ mean the forward and backward GRU hidden state. The last hidden state contains the information about the entire utterance, so we treat $u_i = [\overrightarrow{h_m}, \overleftarrow{h_1}]$ as the representation of the current utterance, where $u_t \in \mathbb{R}^{2D_h}$, D_h is the size of the unidirectional hidden layer vector. u_t^P and u_t^R denote the preliminary representations of the utterances P_t and R_t respectively.

To extract the dialogue content and the parties' emotion more accurately, we introduce four GRUs, namely global GRU (GRU_g), party GRU (GRU_p), emotion GRU (GRU_e) and content GRU (GRU_c). Since the extraction methods of two parties' emotional representation are the same, in this section, we mainly describe how to extract the speaker's emotional representation P_e_t of the t-th turn. GRU_g is equivalent to the ContextRNN in the HRED model, which tracks the utterances of all parties and captures the context information.

$$g_t^P = GRU_g\left(g_{t-1}^R, u_t^P\right) \tag{1}$$

where $g_t^P \in \mathbb{R}^{D_g}$, D_g denotes the global state vector size. g_t^P means the current global state of t-th turn containing dialogue history information. Then, we get the sequence of global states $\{g_1^P, g_1^R, \ldots, g_{t-1}^P, g_{t-1}^R, g_t^P\}$.

Next, in order to be aware of the parties' emotion comprehensively and model the connection of parties during dialogue, we introduce GRU_p to get the party states. We follow the party states in dialogue and consider the effects between the party states. In the t-th turn, we update the speaker's state p_t^P by incoming the listener's state p_{t-1}^R and all relevant information about P_t to GRU_p.

$$p_t^P = GRU_p\left(p_{t-1}^R, \left[u_t^P; a_t^P\right]\right) \tag{2}$$

$$a_t^P = \beta\left[g_1^P, g_1^R, \ldots, g_{t-1}^R\right]^T \tag{3}$$

$$\beta = \text{softmax}\left(\left(u_t^P\right)^T W_\beta \left[g_1^P, g_1^R, \ldots, g_{t-1}^R\right]\right) \tag{4}$$

where $W_\beta \in \mathbb{R}^{2D_h \times D_g}, \beta^T \in \mathbb{R}^{2(t-1)}$, $p_t^P \in \mathbb{R}^{D_p}$. D_p is the size of the party state vector. We calculate the attention weight β to obtain the feature $a_t^P \in \mathbb{R}^{D_g}$ containing the information related to u_t^P in the dialogue history. It catenates u_t^P to get all the related information which would help the emotion classification. Then, we get the sequence of party states $\{p_1^P, p_1^R, \ldots, p_{t-1}^P, p_{t-1}^R, p_t^P\}$ by GRU_p.

After getting the party state, we use GRU_e to track emotion of the speaker and the listener respectively. According to the speaker's emotion P_e_{t-1} of last turn and the current speaker's state p_t^P, we extract the speaker's emotional representation P_e_t by GRU_e.

$$P\_e_t = GRU_e\left(P\_e_{t-1}, p_t^P\right) \tag{5}$$

where $P\_e_t \in \mathbb{R}^{D_e}$, D_e is the vector size of the emotional representation. The tracking of the two parties' emotional representation makes it more sensitive to perceive the change of emotion and helps the model to extract the emotional representation more accurately. So far, we obtain the sequence of emotional representations of two parties in multi-turn dialogue $\{P\_e_1, R\_e_1, P\_e_2, R\_e_2, \ldots, P\_e_t\}$. We employ a feed forward neural network including the softmax activation function to predict the probability distribution of the speaker's and listener's emotions. The loss function is L_1.

$$\begin{aligned} L_1 = &-\frac{1}{N}\sum_{t=1}^{N} \log P\left(\hat{e}_t^P = e_t^P\right) \\ &-\frac{1}{N-1}\sum_{t=1}^{N-1} \log P\left(\hat{e}_t^R = e_t^R\right) \end{aligned} \tag{6}$$

N means N turns dialogues, e_t^P and e_t^R are true emotion labels, $\hat{e}_t^P$ and $\hat{e}_t^R$ are predicted emotion categories.

Although the global state contains the context information, in order to reduce the impact of emotion classification errors on the response generation task and capture the main content of current dialogue turn, GRU_c is introduced to extract the dialogue content representation c_t.

$$c_t = GRU_c\left(c_{t-1}, g_t\right) \tag{7}$$

where $c_t \in \mathbb{R}^{D_c}$, D_c is the size of dialogue content representation. We get the dialogue content sequence $\{c_1, c_2, \ldots, c_t\}$ by GRU_c.

3.3 Empathetic Policy Learning

Empathy can be seen as the transfer of emotion between the parties, so we define the process of predicting the listener's emotion state $\hat{R\_e_t}$ based on the speaker's emotion state P_e_t as the empathetic policy. The listener's emotion does not just mimic the emotion of speaker's utterance, and the responses to one emotion are also diverse. Therefore, we employ a stochastic policy network to learn the empathetic policy (empathetic policy network) from real empathetic dialogue data, which takes advantage of the exploratory characteristic of reinforcement learning. In the empathetic policy network, the state in the t-th turn is P_e_t, the action for the current state is $\hat{R\_e_t}$, the action function is the probability distribution of the action $\hat{R\_e_t}$ under the condition of the state P_e_t. For getting the predicted $\hat{R\_e_t}$, we define the probability density function of $\hat{R\_e_t}$ as the policy network π, and regard the probability density function of the normal distribution as the policy function, i.e. $\pi\left(\hat{R\_e_t} \mid P\_e_t\right) = \frac{1}{\sqrt{2\pi}\sigma} \cdot \exp\left(-\frac{\left(\hat{R\_e_t}-\mu\right)^2}{2\sigma^2}\right)$, $\hat{R\_e_t} \sim N\left(\mu, e^\rho\right)$. As shown in Figure 2 (b), in order to obtain this normal distribution, we learn the approximation of the mean and variance using two different fully connected networks.

$$sta_t = FFN_S\left(P\_e_t\right) \tag{8}$$

$$\hat{\mu} = W_\mu sta_t + b_\mu \tag{9}$$

$$\hat{\rho} = W_\rho sta_t + b_\rho \tag{10}$$

$$FFN(x) = \sigma(Wx + b) \tag{11}$$

where $\hat{\mu}$ is the approximate mean, $\hat{\rho}$ is the approximate logarithm of the variance, FFN represents the feed forward network, σ denotes the tanh activation function. In addition, the fully connected network does not change the vector dimension, i.e. $\hat{\mu}, \hat{\rho}, sta_t \in \mathbb{R}^{D_e}$. Then, we sample from this distribution to obtain response emotional representation $\hat{R\_e_t}$.

In terms of optimizing the policy network, we want the predicted emotional representation $\hat{R\_e_t}$ to be similar to the real response emotional representation R_e_t, so we apply the Constraint-based GPS to guide the optimization of the empathetic policy which utilizes the emotion category distribution of the real response as constraint. The GPS method was first proposed by Levine [22], which can be regarded as a combination of reinforcement learning and imitation learning. The GPS divides the policy search method into a control phase and a supervision phase. The policy network generates data through the interaction with the control phase, and the supervision phase learns and optimizes from the data generated by the control phase. The Constraint-based GPS utilizes KL divergence for optimization involving two probability distributions, q and p. In detailed, p is the target probability distribution, q is the approximate distribution of the target probability

distribution, and the loss function L_2 is as follow.

$$L_2 = KL(q(\tau)\|p(\tau)) \tag{12}$$

where $\tau = \{R\_e_1, R\_e_2, \ldots, R\_e_t\}$. However, as $\hat{R\_e_t}$ and R_e_t are the high-dimensional features, and we only assume that $\hat{R\_e_t} \sim N(\mu, e^{\rho})$, we can't obtain their probability distributions concurrently. Therefore, we use the emotion classifier as an auxiliary network to obtain the emotion category distributions of $\hat{R\_e_t}$ and R_e_t. At this time, the action trajectory is transformed into $\{e_1^R, e_2^R, \ldots, e_t^R\}$ and then the probability distributions are calculated as follows:

$$q = P\left(e_t^R \mid P_1, R_1, \ldots, P_t\right) = \alpha\left(W_e R\_e_t\right) \tag{13}$$

$$p = P\left(\hat{e_t^R} \mid P_1, R_1, \ldots, P_t, R_t\right) = \alpha\left(W_e \hat{R\_e_t}\right) \tag{14}$$

where $W_e \in \mathbb{R}^{32 \times D_e}$. α means softmax function. We get the emotional representation $\hat{R\_e_t}$ by optimizing the distance between p and q.

3.4 Decoding

Before decoding, we convert the content representation c_t and the response's emotional representation $\hat{R\_e_t}$ to con_t and emo_t by the feed forward network for unifying the magnitude. In order to make the responses contain more context-related words, we employ the pointer generation network [37] to decode, and introduce the emotion attention to dynamically integrate emo_t into the generation. The implementation details are shown in Figure 2 (c).

In the i-th generation step, the generated word of our model is jointly determined by the decoder hidden state s_i, the context vector $context_i$, and the emotional representation emo_t^i. Concretely, on the i-th step, the decoder unit (a single-layer unidirectional GRU) accesses the word embedding of last word out_{i-1} and the last decoder state s_{i-1}.

$$s_i = GRU_d\left(s_{i-1}, out_{i-1}\right) \tag{15}$$

Among them, $out_i = Ew(w_i)$, $out_i \in \mathbb{R}^{D_w}$, $s_{i-1}, s_i \in \mathbb{R}^{D_s}$, D_s is the size of decoder state vector. The initial decoder state s_0 is con_t.

We calculate attention scores γ_{ij} over encoder hidden state h_j to get the context vector $context_i$ that represents the context information related to s_{i-1}. In addition, we follow the coverage mechanism [38] to constrain the distribution of attention weights.

$$context_i = \sum_{j=1}^{m} \gamma_{ij} h_j \tag{16}$$

$$\gamma_{ij} = \frac{\exp\left(con\_h_{ij}\right)}{\sum_{k=1}^{T} \exp\left(con\_h_{ik}\right)} \tag{17}$$

$$con\_h_{ij} = \sigma\left(s_{i-1} W_{con} h_j + W_r cover_{ij} + b_i\right) \tag{18}$$

$$cover_i = \sum_{i'=0}^{i-1} \gamma_{i'} \tag{19}$$

where W_{con} and W_r are learnable parameters, σ means tanh activation function. We choose the bilinear function to calculate attention weight incorporating the coverage vector. Thereinto, $cover_i$ is able to control the attention weight and avoid repeating attention problem.

We argue that each word in response contains a different degree of emotion, so we dynamically integrate emo_t during generating on each step. Therefore, we use emotion attention to extract the empathetic response emotion emo_t^i.

$$emo_t^i = \sigma(s_i W_{emo} emo_t + b_i) * emo_t \tag{20}$$

where $emo_t^i \in \mathbb{R}^{D_e}$, $W_{emo} \in \mathbb{R}^{D_e}$. After capturing s_i, $context_i$ and emo_t^i, we make use of them to get the generated word probability distribution $P(w)$. $P(w)$ consists of two parts: the probability distribution of the words in vocabulary and the probability distribution of words in the speaker's utterance. And we utilize p_{gen} to balance the two.

$$P(w) = p_{gen} P_{vocab}(w) + \left(1 - p_{gen}\right) \sum_{j:w_j=w} \gamma_{ij} \tag{21}$$

$$P_{vocab} = \alpha\left(V\left[s_i; context_i; emo_t^i\right] + b\right) \tag{22}$$

$$P_{gen} = \sigma\left(W_c^T context_i + W_s^T s_i + W_e^T emo_t^i + b_{ptr}\right) \tag{23}$$

where $b, V \in \mathbb{R}^{D_s + 2D_h + D_e}$. α and σ denote softmax and sigmoid function respectively. The attention weight γ_{ij} serves as a pointer to copy a word from the speaker's utterance.

The loss function is described as follow. The cross-entropy loss function L_3 is the goal of text generation. Moreover, the coverage loss L_4 controls attention to uniformly obtain information from the encoder hidden states.

$$L_3 = -\frac{1}{\sum_{s=1}^{N} m(s)} \sum_{t=1}^{N} \sum_{i=1}^{m(t)} y(w_i) \log P(w_i) \tag{24}$$

$$L_4 = \frac{1}{\sum_{s=1}^{N} m(s)} \sum_{t=1}^{N} \sum_{i=1}^{m(t)} \min\left(cover_{ij}, \gamma_{ij}\right) \tag{25}$$

Among them, N means N turns dialogues, m means m words in utterances, $y(w_i)$ denotes the true distribution of vocabulary on the i-th step. The complete loss function is L.

$$L = L_1 + L_2 + L_3 + L_4 \tag{26}$$

4 Experimental Settings

4.1 Data

We conduct experiments on the Empathetic Dialogues dataset proposed by Hannah Rashkin et al. [3] to evaluate our method, which is a common dataset in EDG. The train / validation / test data contains 19533 / 2770 / 2547 sets of multi-turn dialogues, and the specific example is shown in Fig. 1. Each set of dialogue contains dialogue situation, emotion label, and dialogue content. There are 32 emotion categories, and the distribution is relatively even.

The emotion label of the Empathetic Dialogues dataset is just for the speaker in the current situation, while our model tracks the emotions of both parties in the dialogue and needs the emotion label of real responses to optimize the empathetic policy. Therefore, we train an emotion classifier to label the emotion categories of the responses in the dataset. We collect the situation discourses in the dataset as training data and fine-tune the Roberta-Large[1)] pre-trained model [39]. The accuracy of emotion classifier achieves 62.8% in the test set, and we adopt this classifier to label the emotion categories of the listener's utterances in the dataset.

4.2 Model Comparisons

We choose the following baselines and state-of-art models for comparison:

MIME [11] encodes imitation and non-imitation representation for positive and negative emotion, then combines both for the decoding part.

EMPDG [6] extracts the emotional representations from discourse-level and token-level, meanwhile, it utilizes the adversarial generative network to learn the empathetic dialogue policy.

KEMP [15] constructs emotional context graph using multi-type knowledge and incorporates emotions into the decoder by emotional attention mechanism.

EmpHi [2] uses the latent variable to learn the intention distribution of empathetic dialogues and combines implicit and explicit intention distributions to generate empathetic responses.

Additionally, we also conduct ablation experiments to analyze the influence of different components in our model:

w/o emp-policy removes the stochastic policy network and the speaker's emotional representation is directly used in the decoder.

w/o GPS eliminates the KL loss, then $R\hat{\_}e_t$ is only a hidden variable in dialogue generation and loses the meaning of response's emotional representation.

w/o GRU_e regards the party state as emotional representation for classification and dialogue generation.

w/o GRU_p inputs the global state to GRU_e obtain the emotional representation directly without the party state.

w/o emo-track connects the inputs and outputs of the GRU_e at all times, which does not distinguish between speaker's posts and listener's empathetic responses.

4.3 Evaluation Metrics

Automatic evaluation: We adopt **Emotion Accuracy** (Acc for short) to evaluate the model's ability to recognize speakers' emotions. **Perplexity** (PPL for short) measures the uncertainty of the generated responses. **Distinct-1** and **Distinct-2** [40] (D-1 and D-2 for short) are the ratio of different unigrams / bigrams in the generated responses which are used to evaluate the response diversity. We utilize the BERT score ($\mathbf{F_{BERT}}$) [41] to calculate the semantic similarity between generated responses and real responses. We also provide the number of models' parameters.

Human evaluation: In the experiments of this paper, human evaluation includes **human rating** and **A / B test** in this paper. Specifically, we use the common rating method in EDG following [11]. We randomly sample 100 dialogue contexts and generated responses of each model from the test set, then three annotators rates the generated responses from 1 to 5 using the Likert scale in terms of empathy, relevance, and fluency, where 1, 3, and 5 indicate unacceptable, moderate, and excellent performance. In the A / B test, we resample 100 dialogues, and three annotators select the better response among the utterance generated by the two comparison models. If the quality of the two models' responses is considered to be of the same degree, then the annotators can choose a tie. We evaluate different 100 dialogues for every A / B test to ensure the fairness of the results.

4.4 Implementation Details

We implement the full model through the Pytorch framework and perform the model optimization using the Adam method. To ensure the fairness of the experimental results, we reproduce the comparison models based on the hyperparameters mentioned in their paper. We repeat the experiment five times

[1)] https://huggingface.co/roberta-large

Table 1 The results of automatic evaluation and human rating. The best performing result for each metric is highlighted. We repeat 5 runs with different seeds and average the results for each automatic metrics. Emp., Rel. and Flu. are short for Empathy, Relevance and Fluency. The Flessia-Kappa κ for three rating aspects are 0.47,0.51 and 0.41, which indicate three annotators reach a moderate agreement.

Method	Params.	Automatic evaluation					Human rating		
		Acc	PPL	Dis-1	Dis-2	F_{BERT}	Emp.	Rel.	Flu.
MIME	17.80M	0.3239	37.28	0.39	1.56	0.189	3.14	3.19	**4.85**
EmpDG	29.29M	0.3278	36.26	0.45	2.13	0.195	3.21	3.19	4.72
KEMP	32.54M	0.3852	34.56	0.61	2.69	0.223	3.20	3.16	4.79
EmpHi	10.92M	0.4046	33.49	0.98	4.64	0.187	3.07	3.05	4.67
ETHREED	11.98M	**0.4294**	**27.74**	**1.14**	**4.76**	**0.226**	**3.25**	**3.23**	**4.85**

for each model and take the average value as the final results. We set the batch size to 8 and the maximum length of decoding to 40. The hidden state dimension is 300 which is same as other baselines. We use the validation set to select the appropriate learning rate and decay ratio. We first experimented with a learning rate from 0.0005 to 0.0001. We find the lower learning rate takes longer time to train, with lower emotion accuracy and lower PPL, and the higher learning rate is easy to overfit but achieve the high emotion accuracy quickly. So, we introduce the decay trick, which sets the learning rate to 0.0005 and decayed every three epochs with a decaying ratio of 0.3 to give a better balance between emotion accuracy and PPL. In addition, we apply the teacher forcing strategy in the process of training the decoder.

Table 2 The results of A / B test. The Flessia-Kappa $\kappa \in [0.41, 0.6]$ denotes the moderate agreement, the Flessia-Kappa $\kappa \in [0.21, 0.4]$ denotes the fair agreement.

ETHREED vs	Win(%)	Loss(%)	Tie(%)	κ
MIME	46.67	27.33	26.00	0.45
EmpDG	41.33	35.33	23.33	0.44
KEMP	43.33	31.67	25.00	0.46
EmpHi	40.67	37.67	21.67	0.36

Table 3 The results of ablation study. The best performing result for each metric is highlighted.

Model	Acc	PPL	D-1	D-2	F_{BERT}
ETHREED	0.4294	**27.74**	**1.14**	**4.76**	0.227
w/o emp-policy	0.4271	27.98	1.09	4.48	0.222
w/o GPS	0.4279	27.95	1.07	4.47	0.227
w/o GRU_p	0.4283	27.77	1.09	4.53	0.226
w/o GRU_e	**0.4349**	27.84	1.12	4.69	0.223
w/o emo-track	0.4152	27.81	1.09	4.61	**0.228**

5 Result and Discussion

5.1 Generation Performance

Automatic Evaluation: The results of the automatic evaluation are shown in Table 1. Firstly, the ETHREED model has the highest accuracy in speakers' emotion recognition, which verifies that the model has a good ability to perceive emotions. Secondly, the PPL of our model significantly reduces, which indicates the high-level general quality of our model. In addition, the promotion on D-1 and D-2 reveals that the responses generated by ETHREED are more diverse. We believe that the flexibility of the empathetic policy plays a great role in responses' diversity. At last, the result on F_{BERT} metric shows that ETHREED has a higher semantic similarity to real responses, although the responses are more diverse. We also provide the number of parameters for the model, which have a huge difference from LLMs, but our model still performs well with a small number of parameters.

Human Evaluation: Table 1 presents the results of human rating. Our model is optimal in empathy, relevance, and fluency (the fluency score is consistent with MIME). Instead of rating responses from one particular perspective, the A / B test evaluates the responses from the whole, that is to choose the more appropriate empathetic response for the current dialogue context. The results in Table 2 show that the winning percentage of our model is larger than the losing percentage, and compared with the baseline MIME model, our model are better significantly. It confirms our model's effectiveness.

5.2 Ablation Study

To better verify the effect of each part of ETHREED, we conduct ablation experiments, and the results are shown in Table 3. Firstly, in the w/o emp-policy and w/o GPS experiments, the D-1 and D-2 decrease significantly, and the PPL increases. The results indicate that the stochastic policy network can make the empathetic responses more flexible and diverse, and using GPS can better learn the empathetic interaction pattern from the data. Secondly, to prove the effect of hierarchical structure, we conduct experiments without GRU_e and GRU_p respectively. The experimental results show that the quality of the generated responses would decrease af-

Table 4 Examples of comparison model responses

Label	Joyful
Context	I am so excited because I am finally going to visit my parents next month! I did not see them for 3 years
Golden	3 years is a long time. How come?
MIME	That is great! I am happy for you!
EmpDG	That is awesome! Where are you going?
KEMP	That is great! I hope you have a great time!
EmpHi	I bet you are excited about it.
ETHREED	**That's awesome! What are you planning on doing?**
Label	Afraid
Context	I was afraid when i went into the movies alone
Golden	Really why is that?
MIME	Oh no! I am so sorry to hear that. What happened?
EmpDG	I am glad you were okay. I am scared to go back to the movies.
KEMP	That is scary! I am glad you are okay.
EmpHi	I know what you mean. I have a movie movie.
ETHREED	**Why were you scared**
Label	Faithful
Context	Speaker: I've been married almost 23 years. My husband travel so much, but we both have trust in each other to be true and loyal. Listener: That is great, why does he travel so much? Speaker: For his job. So, it can be hard, but we trust each other.
Golden	What does he do for his job?
MIME	I am happy for you
EmpDG	I am sure he will be fine
KEMP	I am sure you will do great!
EmpHi	I agree, I'm glad you're going to be a good friend.
ETHREED	**That's great. I hope you have a good relationship with him.**
Label	Sad
Context	Speaker: One of my best friends is moving far away and i do not know when i will be able to see him again. Listener: That's so sad! can you email them? Speaker: Yeah i can call him and email him, it is just that i do not know when we will see each other again
Golden	can you see him before he goes?
MIME	I am sorry to hear that. I hope you have a good time.
EmpDG	I am sorry to hear that. I hope you find a new friend.
KEMP	I am sure you will do great!
EmpHi	I know, I'm sure you'll get to get a good friend.
ETHREED	**I hope you are able to find a good time to see him.**

ter reducing the GRU. Among them, in the w/o GRU_e experiment, we recognize emotion on the party state directly with higher accuracy, but the PPL increases and F_{BERT} decreases. We believe that the emotion recognition task can act more directly on u_t and reach the optimum faster in the w/o GRU_e experiment, but at the same time, the error of emotion recognition impacts the dialogue generation easier. At last, to further verify the effect of emotion tracking, we conduct the w/o emo-track experiment, in which the accuracy of emotion recognition decreases significantly. So, tracking two parties' emotion respectively by GRU_e and GRU_p can better capture the emotional representation for generating empathetic responses. Moreover, the accuracy of emotion recognition based on the recurrent hierarchical architecture is at a high level, which illustrates that the RNN's ability to process time sequences helps model extract parties' emotional representation more accurately.

5.3 Case Study

Some single-turn and multi-turn dialogues are chosen in Table 4 to provide a more visual display of the responses generated by different models. In the first example, our generated response contains both emotional feedback and questions about relevant content which is the specific and diverse. In the second example, our response is consistent with the gold response and is non-generic. The last two examples

are multi-turn dialogues. The responses are more contextually relevant and empathetic, besides, they contain context-related words such as "good relationship". This shows that ETHREED is effective in handling multi-turn dialogues.

5.4 Error Analysis

Data labeling error: In order to make the empathetic policy learn the responses' emotional representations from real data, we label the emotion categories of listeners' responses in the dataset. Due to the small number of utterances with labeled emotions and the similarity of some emotion categories, the accuracy of emotion recognition model is just 62.8%, which needs to be further improved. The accuracy of response emotion recognition is 48.81% in this paper. If the response emotion is further accurately labeled, the quality of the generated responses will also be improved.

6 Conclusion

In this paper, we propose a novel empathetic dialogue generation model ETHREED, which relies on hierarchical GRUs to transmit dialogue history information, track the emotional representation of both parties in the dialogue separately. Besides, we predict the responses' emotional representations by using the stochastic policy network and the guided policy search method. The experimental results show that our responses have better diversity, empathy and relevance. In the future, we will consider introducing the dialogue behavior to guide the response generation.

References

1. Decety J, Lamm C. Human empathy through the lens of social neuroscience. TheScientificWorldJOURNAL, 2006, 6: 1146–1163
2. Chen M Y, Li S, Yang Y. Emphi: Generating empathetic responses with human-like intents. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2022, 1063–1074
3. Rashkin H, Smith E M, Li M, Boureau Y L. Towards empathetic open domain conversation models: A new benchmark and dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019, 5370—5381
4. Welivita A, Xie Y, Pu P. A large-scale dataset for empathetic response generation. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021, 1251–1264
5. Zhong P, Zhang C, Wang H, Liu Y, Miao C. Towards persona-based empathetic conversational models. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 6556–6566
6. Li Q, Chen H, Ren Z, Ren P, Tu Z, Chen Z. Empdg: Multi-resolution interactive empathetic dialogue generation. In: Proceedings of the 28th International Conference on Computational Linguistics. 2020, 4454–4466
7. Idicula M A. A multi-resolution mechanism with multiple decoders for empathetic dialogue generation. In: Proceedings of International Conference on Smart Computing and Communications. 2021, 240–245
8. Shen L, Zhang J, Ou J, Zhao X, Zhou J. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In: Findings of the Association for Computational Linguistics: EMNLP. 2021, 3124–3134
9. Varshney D, Ekbal A, Bhattacharyya P. Modelling context emotions using multi-task learning for emotion controlled dialog generation. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. 2021, 2919–2931
10. Lin Z, Madotto A, Shin J, Xu P, , Fung P. Moel: Mixture of empathetic listeners. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). 2019, 121–132
11. Majumder N, Hong P, Shanshan Peng J L, Ghosal D, Gelbukh A, Mihalcea R, Poria S. Mime: Mimicking emotions for empathetic response generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020, 8968—8979
12. Gao J, Liu Y, Deng H, Wang W, Du Y C J, Xu R. Improving empathetic response generation by recognizing emotion cause in conversations. In: Findings of the Association for Computational Linguistics: EMNLP2021. 2021, 807–819
13. Wang J, Li W, Lin P, Mu F. Empathetic response generation through graph-based multi-hop reasoning on emotional causality. Knowl. Based Syst., 2021, 233: 107547
14. Pan R, Zou M, Zhang S, Yu Y, Feng Z. Improving empathetic dialogue generation with semantics decoupling. In: Proceedings of the 11th International Joint Conference on Knowledge Graphs. 2022, 55–63
15. Li Q, Li P, Ren Z, Ren P, Chen Z. Knowledge bridging for empathetic dialogue generation. In: Proceedings of the 36th AAAI Conference on Artificial Intelligence. 2022, 10993–11001
16. Zandie R, Mahoor M H. Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems. In: Proceedings of the 33th International Florida Artificial Intelligence Research Society Conference. 2020, 276–281
17. Zhou H, Huang M, Zhang T, Zhu X, Liu B. Emotional chatting machine: Emotional conversation generation with internal and external memory. In: Proceedings of the 32th AAAI Conference on Artificial Intelligence. 2018, 730–739
18. Zhou X, Wang W Y. Mojitalk: Generating emotional responses at scale. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018, 1128–1137
19. Serban I V, Sordoni A, Bengio Y, Courville A C, Pineau J. Building end-to-end dialogue systems using generative hierarchical neural network models. In: Proceedings of the 30th AAAI Conference on

Artificial Intelligence. 2016, 3776–3784

20. Majumder N, Poria S, Hazarika D, Mihalcea R, Gelbukh A F, Cambria E. Dialoguernn: An attentive RNN for emotion detection in conversations. In: Proceedings of the 33th AAAI Conference on Artificial Intelligence. 2019, 6818–6825
21. Cho K, Merrienboer v B, Gülçehre Ç, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 2014, 1724–1734
22. Levine S. Motor skill learning with local trajectory methods. PhD thesis, Stanford University, USA, 2014
23. Liu Y, Maier W, Minker W, Ultes S. Empathetic dialogue generation with pre-trained roberta-gpt2 and external knowledge. In: Conversational AI for Natural Human-Centric Interaction - 12th International Workshop on Spoken Dialogue System Technology. 2021, 67–81
24. Lee Y, Lim C, Choi H. Does GPT-3 generate empathetic dialogues? A novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In: Proceedings of the 29th International Conference on Computational Linguistics. 2022, 669–683
25. Liu Y, Du J, Li X, Xu R. Generating empathetic responses by injecting anticipated emotion. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2021, 7403–7407
26. Lee J Y, Lee K A, Gan W. Improving contextual coherence in variational personalized and empathetic dialogue agents. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing. 2022, 7052–7056
27. Zaremba W, Sutskever I, Vinyals O. Recurrent neural network regularization. arXiv preprint, 2014, arXiv:1409.2329
28. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I. Attention is all you need. In: Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017. 2017, 5998–6008
29. Serban I V, Sordoni A, Lowe R, Charlin L, Pineau J, Courville A C, Bengio Y. A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the 31th AAAI Conference on Artificial Intelligence. 2017, 3295–3301
30. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI blog, 2019, 1(8): 9
31. Lubis N, Sakti S, Yoshino K, Nakamura S. Positive emotion elicitation in chat-based dialogue systems. IEEE ACM Trans. Audio Speech Lang. Process., 2019, 27(4): 866–877
32. Xie Y, Pu P. Empathetic dialog generation with fine-grained intents. In: Proceedings of the 25th Conference on Computational Natural Language Learning. 2021, 133–147
33. Saha T, Ananiadou S. Emotion-aware and intent-controlled empathetic response generation using hierarchical transformer network. In: Proceedings of International Joint Conference on Neural Networks. 2022, 1–8
34. Yan X, Yang J, Sohn K, Lee H. Attribute2image: Conditional image generation from visual attributes. In: Proceedings of the Computer Vision - ECCV 2016 - 14th European Conference, Part IV. 2016, 776–791
35. Shin J, Xu P, Madotto A, Fung P. Generating empathetic responses by looking ahead the user's sentiment. In: Proceedings of 2020 IEEE International Conference on Acoustics, Speech and Signal Processing. 2020, 7989–7993
36. Pennington J, Socher R, Manning C D. Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014. 2014, 1532–1543
37. Vinyals O, Fortunato M, Jaitly N. Pointer networks. In: Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015. 2015, 2692–2700
38. See A, Liu P J, Manning C D. Get to the point: Summarization with pointer-generator networks. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017, 1073–1083
39. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V. Roberta: A robustly optimized bert pretraining approach. arXiv preprint, 2019, arXiv:1907.11692
40. Li J, Galley M, Brockett C, Gao J, Dolan B. A diversity-promoting objective function for neural conversation models. In: Proceedings of The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016, 110–119
41. Zhang T, Kishore V, Wu F, Weinberger K Q, Artzi Y. Bertscore: Evaluating text generation with BERT. In: Proceedings of the 8th International Conference on Learning Representations. 2020