Fine-Grained Knowledge Enhancement for Empathetic Dialogue Generation

Ai Chen, Jiang Zhong(✉), Qizhu Dai, Chen Wang, and Rongzhen Li

Chongqing University, Shapingba, Chongqing, China
ai.chen@stu.cqu.edu.cn,
{zhongjiang,daiqizhu,chenwang,lirongzhen}@cqu.edu.cn

Abstract. An engaging dialogue system is supposed to generate empathetic responses, which requires a cognitive understanding of users' situations and an affective perception of their emotions. Most of the existing work only focuses on modeling the latter, while neglecting the importance of the former. Despite some efforts to enhance chatbots' empathy in both cognition and affection, limited cognition conditions and inaccessible fine-grained information still impair the effectiveness of empathy modeling. To address this issue, we propose a novel fine-grained knowledge-enhanced empathetic dialogue generation model KEEM. We first explore strategies to filter fine-grained commonsense and emotional knowledge and leverage knowledge to construct cognitive and affective context graphs. And we learn corresponding context representations from the two knowledge-enhanced context graphs. Then we encode the raw dialogue context to learn the original cognitive and affective representations and fuse them with the knowledge-enhanced representations in cognition and affection. Finally, we feed the two fused representations into a decoder to produce empathetic replies. Extensive experiments conducted on the benchmark dataset EMPATHETICDIALOGUES verify the effectiveness of our model in comparison with several competitive models.

Keywords: Empathetic dialogue generation · Commonsense and emotional knowledge · Cognition and affection

1 Introduction

Empathy, an important aspect of engaging human conversations [14], usually refers to the ability to understand others' situations, perceive their emotions, and respond to them appropriately [5]. Research in social psychology has also shown that empathy is a vital step towards a more humanized dialogue system [20]. Consequently, we concentrate on the task of empathetic dialogue generation, which aims to empathize with users and realize a more human-like chatbot.

It is demonstrated that empathy is a complex construct involving cognition and affection [15], where cognitive empathy attends to users' situations [2] while affective one focuses on users' emotions instead [3]. But most of the existing methods [7–10,14] in empathetic dialogue generation only rely on detecting

X. Yang et al. (Eds.): ADMA 2023, LNAI 14179, pp. 78–93, 2023.
https://doi.org/10.1007/978-3-031-46674-8_6

users' emotions and modeling emotional dependencies, while ignoring the importance of realizing cognitive empathy. To achieve empathy in both two aspects, Sabour et al. [15] make an attempt to use commonsense to enhance the modeling of cognitive and affective empathy and several cognition conditions would be inferred in this method. However, limited cognitive conditions and inaccessible fine-grained information still result in inaccurate recognization of users' circumstances and feelings, thereby impairing the empathetic effect of the generated responses.

Providing external knowledge to dialogue systems has been proven to operate in favor of modeling empathy in cognition and affection [15]. Intuitively, fine-grained knowledge would be beneficial for a more comprehensive understanding of user situations and a more accurate perception of user feelings, which is shown in Fig. 1. In this case, with the related cognitive concept of "walk", the chatbot understands the user's situation that he or she encountered a snake while walking so it asks what the user did. Also, the affective concepts "dazed", "demon" and "poisonous" help the robot perceive the terrified feeling of the user.

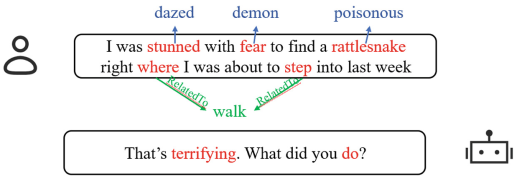
dazed
demon
poisonous
I was stunned with fear to find a rattlesnake right where I was about to step into last week
RelatedTo
walk
RelatedTo
That's terrifying. What did you do?

Fig. 1. An example from EMPATHETICDIALOGUES. Words related to cognition and affection are highlighted in red color, cognitive concepts and relations in green, and affective concepts in blue. (Color figure online)

In this paper, we propose a fine-grained **K**nowledge-**E**nhanced **EM**pathetic dialogue generation model (**KEEM**). We first explore novel knowledge selection strategies to filter commonsense, i.e. structural and semantic knowledge, and emotional knowledge, and use the selected fine-grained knowledge to construct cognitive and affective context graphs. We also learn the corresponding context representations from the above two knowledge-enhanced context graphs. Then we encode the original dialogue context to acquire the original information about cognition and affection, and fuse them with the two knowledge-enhanced representations. Finally, we feed the two fused cognitive and affective representations to a decoder to generate empathetic responses with coherent content and appropriate emotion. Extensive experiments are conducted on the benchmark dataset EMPATHETICDIALOGUES [14] for empathetic dialogue generation and the empirical results have verified that our model can produce more empathetic responses in comparison with several competitive models.

Our contributions can be summarized as follows:

- We propose KEEM, a novel approach that models cognitive and affective empathy by constructing and encoding corresponding context graphs.

- We explore knowledge selection strategies for cognitive and emotional knowledge to obtain more accurate and fine-grained knowledge.
- We conduct automatic and human evaluations and analyses to demonstrate the effectiveness of KEEM.

2 Preliminaries

2.1 Commonsense and Emotional Knowledge

In this work, we leverage the commonsense knowledge graph ConceptNet [17] and the emotional lexicon NRC-VAD [12] to infer the speakers' situations and their emotions, which enhances the cognition and affective empathy and leads to more empathetic responses.

ConceptNet is a large-scale knowledge graph that connects words and phrases of natural language with labeled edges. It represents the general knowledge, allowing models to better understand the meanings behind the words [11]. It contains 34 relations, over $21M$ edges, and over $8M$ nodes. The edges stored in ConceptNet can be concisely represented as the quadruples of their start node, relation label, end node, and confidence score: (h, r, t, s).

NRC-VAD is a lexicon of more than $20k$ English words and their vectors of three independent dimensions, i.e. valence (positiveness-negativeness/pleasure-displeasure), arousal (active-passive), and dominance (dominant-submissive), abbreviated as VAD. The values of VAD vectors are fine-grained real numbers in the interval from 0 (lowest) to 1 (highest).

2.2 Task Formulation

Dialogue context C is a sequence of M utterances: $C = [U_1, U_2, \cdots, U_M]$, where the i-th utterance $U_i = [w_i^1, w_i^2, \cdots, w_i^{m_i}]$ consists of m_i words. Following [9], we flat C as a token sequence, and prepend a CLS token to it, thus obtaining a new context sequence: $C = [CLS, w_1^1, \cdots, w_1^{m_1}, \cdots, w_M^1, \cdots, w_M^{m_M}]$. Given C, the task of empathetic dialogue generation is to generate an empathetic response $Y = [y_1, y_2, \cdots, y_n]$ with coherent content and appropriate emotion.

3 Methodology

3.1 Overview

Figure 2 shows an overview of our proposed model. Our model (KEEM) is composed of four stages: 1)cognitive context graph constructing and encoding; 2)affective context graph constructing and encoding; 3)cognition and affection fusion; and 4)empathetic response generation, where the first two stages can be performed simultaneously.

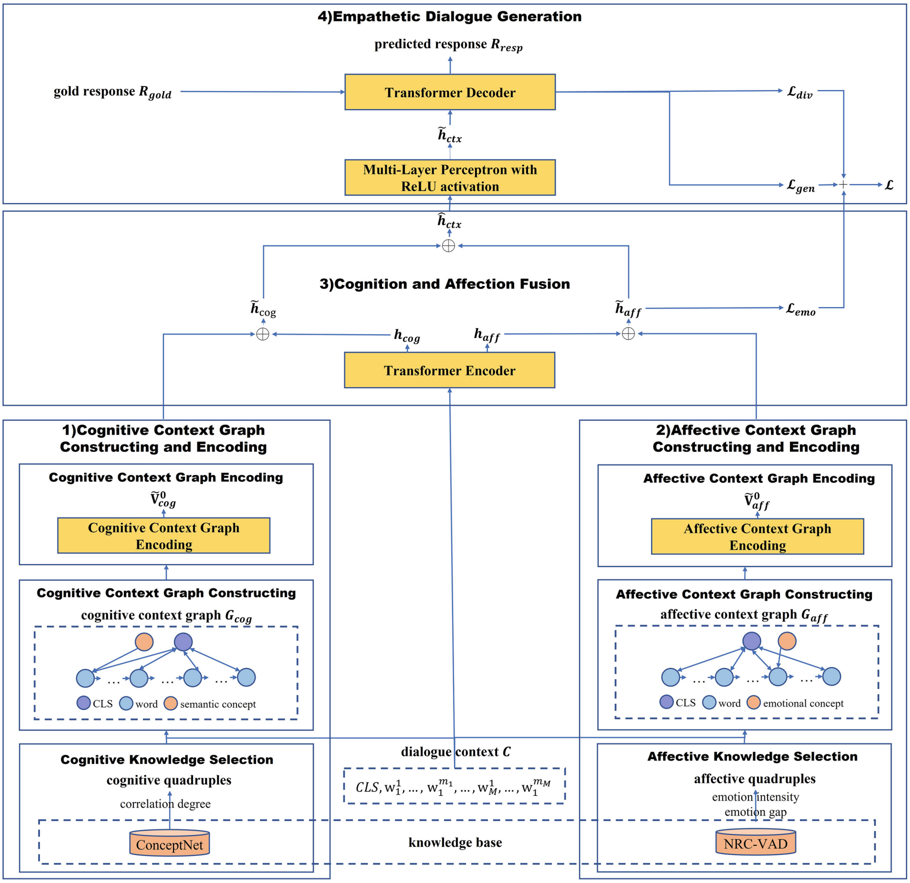
4)Empathetic Dialogue Generation
predicted response R_{resp}
gold response R_{gold}
Transformer Decoder
$\mathcal{L}_{div}$
$\widetilde{h}_{ctx}$
Multi-Layer Perceptron with ReLU activation
$\mathcal{L}_{gen}$
$\mathcal{L}$
$\widetilde{h}_{ctx}$
3)Cognition and Affection Fusion
$\widetilde{h}_{cog}$
$\widetilde{h}_{aff}$
$\mathcal{L}_{emo}$
h_{cog}
h_{aff}
Transformer Encoder
1)Cognitive Context Graph Constructing and Encoding
Cognitive Context Graph Encoding
$\widetilde{V}^{0}_{cog}$
Cognitive Context Graph Encoding
Cognitive Context Graph Constructing
cognitive context graph G_{cog}
CLS
word
semantic concept
Cognitive Knowledge Selection
cognitive quadruples
correlation degree
ConceptNet
2)Affective Context Graph Constructing and Encoding
Affective Context Graph Encoding
$\widetilde{V}^{0}_{aff}$
Affective Context Graph Encoding
Affective Context Graph Constructing
affective context graph G_{aff}
CLS
word
emotional concept
Affective Knowledge Selection
affective quadruples
emotion intensity
emotion gap
NRC-VAD
dialogue context C
$CLS, w_1^1, \ldots, w_1^{m_1}, \ldots, w_M^1, \ldots, w_1^{m_M}$
knowledge base

Fig. 2. Overview of our model (KEEM).

3.2 Cognitive Context Graph Constructing and Encoding

Cognitive Knowledge Selection. The structural and semantic information of ConceptNet, i.e. the relations and concepts therein, can help enhance cognition. Hence, for each non-stop word w_i^j of C, we first retrieve all of its quadruples from ConceptNet as candidates. And then we filter the cognitive knowledge by the following heuristic steps:1) We remove the quadruples that have low confidence scores (i.e. scores lower than 1.0) or inappropriate relations (i.e. relations unrelated to cognition). 2) We define and calculate the correlation degrees of the retrieved concepts. The correlation degree of a concept is the number of C's words that link to it in ConceptNet. If the concept itself appears in the dialogue context, the correlation degree is increased by one. 3) We rank the correlation degrees of the quadruples and select the top K_1 ones as the knowledge needed in cognitive graph construction.

Cognitive Context Graph Constructing. To build the cognitive graph, we first take all the tokens of C, including the CLS token, as the initial nodes of the graph. We also add the semantic concepts as the new nodes to the graph, which are selected from Cognitive Knowledge Selection. Then we connect vertex pairs of three types: 1) every two consecutive words in dialogue context; 2) every word in dialogue context and all of its semantic-related concepts; 3) CLS token and all words in dialogue context. Note that the edges connecting vertices are directed.

Thus, the dialogue context is enhanced by external structural and semantic knowledge and represented as the cognitive context graph G_{cog}, and the links of the graph are stored in the adjacency matrix A_{cog}.

Cognitive Context Graph Encoding. Similar to [8], we first initialize the cognitive vector presentation of every vertex v_{sem}^i by summing up its word embedding $\boldsymbol{E}_w\left(v_{cog}^i\right) \in \mathbb{R}^{d_{model}}$, positional embedding $\boldsymbol{E}_p\left(v_{cog}^i\right) \in \mathbb{R}^{d_{model}}$, and dialogue state embedding $\boldsymbol{E}_s\left(v_{cog}^i\right)$:

$$\mathbf{v}_{cog}^i = \boldsymbol{E}_w\left(v_{cog}^i\right) + \boldsymbol{E}_p\left(v_{cog}^i\right) + \boldsymbol{E}_s\left(v_{cog}^i\right) \tag{1}$$

where d_{model} is the dimension of the embeddings.

Then, we adopt a multi-head graph-attention mechanism, followed by a residual connection and layer normalization, thus v_{cog}^i attending to all its immediate neighbors $\left\{v_{cog}^j\right\}_{j \in A_{cog}^i}$ to update its cognitive presentation with structural and semantic knowledge:

$$\hat{\mathbf{v}}_{cog}^i = \text{LayerNorm}\left(v_{cog}^i + \|_{n=1}^H \sum_{j \in A_{cog}^i} \text{att}_{cog}^n\left(\mathbf{v}_{cog}^i, \mathbf{v}_{cog}^j\right) \mathbf{W}_{cog}^{nv} \mathbf{v}_{cog}^j\right) \tag{2}$$

where LayerNorm is layer normalization, $\|$ is the concatenation of H attention heads, A_{cog}^i are the immediate neighbors of v_{cog}^i presented in the adjacency matrix A_{cog}, $\mathbf{W}_{cog}^{nv} \in \mathbb{R}^{d_{model} \times d_h}$ is the linear transformation, $d_h = \frac{d}{H}$ is the dimension of each head, and $\text{att}_{cog}^n\left(\mathbf{v}_{cog}^i, \mathbf{v}_{cog}^j\right)$ is the self-attention mechanism for the n-th attention head:

$$\text{att}_{cog}^n\left(\mathbf{v}_{cog}^i, \mathbf{v}_{cog}^j\right) = \frac{\exp\left(\left(\mathbf{W}_{cog}^{nq} \mathbf{v}_{cog}^i\right)^\top \mathbf{W}_{cog}^{nk} \mathbf{v}_{cog}^j\right)}{\sum_{k \in A_{cog}^i} \exp\left(\left(\mathbf{W}_{cog}^{nq} \mathbf{v}_{cog}^i\right)^\top \mathbf{W}_{cog}^{nk} \mathbf{v}_{cog}^k\right)} \tag{3}$$

where $\mathbf{W}_{cog}^{nq} \in \mathbb{R}^{d_{model} \times d_h}$, $\mathbf{W}_{cog}^{nk} \in \mathbb{R}^{d_{model} \times d_h}$ are the linear transformations. And notably, when $\mathbf{v}_{cog}^i$ is the vector of a word in dialogue context and $\mathbf{v}_{cog}^j$ is the counterpart of a concept semantic-related, following [1], we update $\mathbf{v}_{cog}^j$ with the subtraction between the concept embedding and the corresponding relation embedding:

$$\mathbf{v}_{cog}^j = \mathbf{v}_{cog}^j - \boldsymbol{E}_r\left(\mathbf{r}_{cog}^{ij}\right) \tag{4}$$

where $\boldsymbol{E}_r\left(\boldsymbol{r}_{cog}^{ij}\right) \in \mathbb{R}^{d_{model}}$ is the relation embedding between $\mathbf{v}_{cog}^{i}$ and $\mathbf{v}_{cog}^{j}$.

After that, we apply Transformer layers [18] to update the vector representations of vertices, incorporating global cognitive information into all vertices in G_{cog}:

$$\tilde{\mathbf{v}}_{cog}^{i} = \mathrm{TRSEnc}\left(\hat{\mathbf{v}}_{cog}^{i}\right) \tag{5}$$

where $\tilde{\mathbf{v}}_{cog}^{i} \in \mathbb{R}^{d_{model}}$ is the semantic vector representation of v_{cog}^{i}, and TRSEnc represents Transformer encoder layers.

Finally, we use the cognitive vertex presentation of CLS token, i.e. $\tilde{\mathbf{v}}_{cog}^{0}$, as the global cognitive representation of G_{cog}.

3.3 Affective Context Graph Constructing and Encoding

The procedures of Subsects. 3.2 and 3.3 are similar, with the strategies of knowledge selection and the approaches of graph encoding being different.

Affective Knowledge Selection. To inject appropriate affective knowledge into the dialogue context to detect users' emotions, we filter emotional concepts with high emotional intensity value and low emotion gap value between them and the dialogue context.

Be analogous to Cognitive Knowledge Selection., for non-stop word w_i^j of C, we first acquire its quadruples from ConceptNet and filter the affective knowledge by removing the quadruples that have low confidence scores or affection-unrelated relations.

Then we retrieve the VAD vectors (mentioned in Sect. 2) of the emotional concepts and calculate their emotion intensity values [8,21]. The formula for calculating the emotional intensity value of concept c is as follows:

$$EI(c) = \min - \max\left(\left\|V(c) - \frac{1}{2}, \quad \frac{A(c)}{2}\right\|\right) \tag{6}$$

where $\min - \max$ is the min-max normalization, $\|\|$ is L_2 norm, $V(c)$ and $A(c)$ are concept c's values of the valence and arousal dimensions in the VAD vector, respectively. If c is not in NRC-VAD, $EI(c)$ will be set to 0.

Besides, we compute the values of the emotion gap between each nonstop word and its emotional concepts. The formula for computing the emotion gap value between word w and concept c is as follows:

$$EG(w, c) = \frac{abs(V(w) - V(c)) + abs(A(w) - A(c))}{2} \tag{7}$$

where abs is the operation to get the absolute value.

Eventually, we filter the quadruples whose emotion gap values between head concept and tail concept (i.e. word in dialogue context and its emotional concept) are lower than 0.5, rank the emotion intensity values of the quadruples, and select the top K_2 ones.

Affective Context Graph Constructing. We construct the affective context graph in a way similar to Cognitive Context Graph Constructing and represent it as G_{aff}, and the emotional links in the knowledge-enhanced graph are stored in the affective adjacency matrix A_{aff}.

Affective Context Graph Encoding. Taking the same steps as Cognitive Context Graph Encoding, we initialize the affective vector presentation of every vertex v^i_{aff}, and then update v^i_{aff} with affective knowledge. But notice that, when encoding the affective context graph, the relations between vertices would not be considered. We also employ Transformer layers [18] to update the vector representations of vertices, incorporating global emotional information into all vertices in G_{aff}.

The affective vertex presentation of the CLS token, that is $\tilde{\mathbf{v}}^0_{aff}$, is also used as the global affective representation of G_{aff}.

3.4 Cognition and Affection Fusion

To full exploit the original information of the dialogue context, we encode the raw dialogue context. We use the same initialization method as Cognitive Context Graph Encoding to acquire the embeddings of context sequence, i.e. E_C, and then feed it into new Transformer encoder layers to get the hidden representations of C:

$$H = TRSEnc(C) \tag{8}$$

Then we use the hidden representation of the CLS token to represent the context sequence:

$$h = H[0] \tag{9}$$

And then we perform the cognitive and affective linear transformation on the hidden representation h to obtain the corresponding representations of the dialogue context, respectively:

$$h_{cog} = \mathbf{W}_{coc} h \tag{10}$$

$$h_{aff} = \mathbf{W}_{aoc} h \tag{11}$$

where $\mathbf{W}_{coc} \in \mathbb{R}^{d_{model} \times d_{model}}$, $\mathbf{W}_{aoc} \in \mathbb{R}^{d_{model} \times d_{model}}$ are the cognitive and affective linear transformations.

Emotion Classification. Our proposed model learns to predict the users' emotional state to guide the empathetic response generation. We concatenate $\widetilde{\mathbf{v}}^0_{aff}$ with h_{aff} to obtain a fused affective representation $\widetilde{h}^0_{aff}$:

$$\tilde{h}_{aff} = \tilde{\mathbf{v}}^0_{aff} \oplus h_{aff} \tag{12}$$

where $\oplus$ denotes concatenation and $\tilde{h}_{aff} \in \mathbb{R}^{d_{model}}$.

Hence, we pass $\tilde{h}_{aff}$ through a linear layer followed by a Softmax operation to produce the emotion category distribution $P_{emo} \in \mathbb{R}^q$, where q is the number of emotion categories:

$$P_{emo} = \text{Softmax}\left(W_{emo}\tilde{h}_{aff}\right) \tag{13}$$

where $W_{emo} \in \mathbb{R}^{2d_{model} \times q}$ is the emotional linear transformation. During training, we conduct the parameter learning by minimizing the Cross-Entropy (CE) loss between the ground truth label e^* and the predicted label e:

$$\mathcal{L}_{emo} = -\log\left(P_{emo}\left(e = e^*\right)\right) \tag{14}$$

Information Integration. To generate empathetic responses, we integrate cognitive and affective information into the dialogue context. First concatenate the global cognitive representation of the cognitive context graph, i.e. $\tilde{\mathbf{v}}^0_{cog}$, and the cognitive representation of the raw dialogue context, i.e. h_{cog}, to obtain a fused cognitive representation:

$$\tilde{h}_{cog} = \tilde{\mathbf{v}}^0_{cog} \oplus h_{cog} \tag{15}$$

where $\tilde{h}_{cog} \in \mathbb{R}^{2d_{model}}$.

Then $\tilde{h}_{cog}$ and $\tilde{h}_{aff}$ are concatenated and the combination of them is passed through a Multi-Layer Perceptron with ReLU activation, which aims to learn a contextualized representation with adequate cognition and affective information:

$$\hat{h}_{ctx} = \tilde{h}_{cog} \oplus \tilde{h}_{aff} \tag{16}$$

$$\tilde{h}_{ctx} = \text{MLP}\left(\sigma\left(\hat{h}_{ctx}\right) \odot \hat{h}_{ctx}\right) \tag{17}$$

where $\hat{h}_{ctx} \in \mathbb{R}^{d_{model}}$, MLP denotes Multi-Layer Perceptron, and $\odot$ denotes element-wise multiplication.

3.5 Empathetic Response Generation

The contextualized representation of cognition and affection, i.e. $\tilde{h}_{ctx}$, and the word embeddings of the target response R_{gold}, i.e. $E_w\left(R_{gold}\right)$, are fed as the inputs into the Transformer decoder layers to generate a response:

$$O = \text{TRSDec}\left(E_w\left(R_{gold}\right), \tilde{h}_{ctx}\right) \tag{18}$$

$$P_{resp} = \text{softmax}\left(W_o O\right) \tag{19}$$

$$p\left(R_t \mid R_{<t}, G_{\text{cog}}, G_{\text{aff}}\right) = P_{\text{resp}}[t] \tag{20}$$

where $O \in \mathbb{R}^{l_R \times d_{model}}$, l_R is the length of the predicted response, TRSDec represents the Transformer decoder layers, $P_{resp} \in \mathbb{R}^{l_R \times |V|}$, $|V|$ is the vocabulary size, and $p\left(R_t \mid R_{<t}, G_{\text{cog}}\ , G_{\text{aff}}\ \right)$ is the distribution over the vocabulary V for the t-th word R_t.

Then a standard Negative Log-Likelihood (NLL) is used to optimize generated responses:

$$\mathcal{L}_{gen}=-\sum_{t=1}^{l_R} \log p\left(R_t \mid R_{<t}, G_{cog}, G_{aff}\right) \tag{21}$$

To avoid generating generic empathetic responses, following [15], we adopt Frequency-Aware Cross-Entropy (FACE) [4] as an additional loss to penalize high-frequency tokens. Therefore, during the training process, we first compute the relative frequency of each token $word_i$ in the training corpus:

$$RF_i=\frac{\operatorname{freq}\left(word_i\right)}{\sum_{j=1}^V \text { freq }\ \left(word_i\right)} \tag{22}$$

where V is the vocabulary size of the training corpus. Accordingly, the frequency-based weight w_i can be calculated as follows:

$$w_i=a \times RF_i+1 \tag{23}$$

where $a=-\left(\max_{1 \leq j \leq V}\left(RF_j\right)\right)^{-1}$ is the frequency slope, 1 is added as the bias so that w_i falls into [0, 1]. As done by [15], we normalize w_i to have a mean of 1. The diversity loss is finally computed as below:

$$\mathcal{L}_{div}=-\sum_{t=1}^T \sum_{i=1}^V w_i \delta_t\left(c_i\right) \log \mathrm{P}\left(c_i \mid y_{<t}, C\right) \tag{24}$$

where c_i is a candidate token in the vocabulary and $\delta_t\left(c_i\right)$ is the indicator function, which equals to 1 only if $c_i=y_t$ and 0 otherwise.

Eventually, all the parameters of our proposed model are trained and optimized by jointly minimizing the emotional loss (Eq. 14), the generation loss (Eq. 21) and the diversity loss (Eq. 24) as follows:

$$\mathcal{L}=\gamma_1 \mathcal{L}_{emo}+\gamma_2 \mathcal{L}_{gen}+\gamma_3 \mathcal{L}_{div} \tag{25}$$

where γ_1, γ_2, γ_3 are hyper-parameters to balance the above three losses.

4 Experiments

4.1 Baselines

We compare our proposed model with the following baselines:

- **MoEL** [9]: A variation of Transformer consisting of one encoder and several decoders that focus on each emotion accordingly.

- **EmpDG** [7]: A adversarial model that encodes semantic context and multi-resolution emotional context respectively and interacts with user feedback.
- **MIME** [10]: Another variation of Transformer. It does emotion grouping and applies emotion stochastic sampling and emotion mimicry.
- **KEMP** [8]: A knowledge-enriched model that uses commonsense and emotional lexical knowledge to explicitly understand and express emotions.
- **CEM** [15]: Another knowledge-enriched Transformer-based model that uses commonsense to obtain more information about users' situations.

4.2 Implementation Details

We conduct our experiments on EMPATHETICDIALOGUES [14], a large-scale dataset of 25k conversations, grounded in emotional situations. the dataset considers 32 emotion labels, of which the distribution is close to evenly distributed. For our experiments, we use the original 8:1:1 train/validation/test split of this dataset.

We use Pytorch[1] to implement the proposed model. The word embeddings are initialized with pre-trained Glove vectors[2] [13], and the relation embeddings are randomly initialized and fixed during training. For the positional embeddings, we follow the original paper [18]. The dimension of embeddings is set to 300 empirically. The maximum introducing numbers of external concepts per dialogue and per token are set as 10 and 1, respectively. And the loss weights γ_1, γ_2, γ_3 are all set to 1. We set the same hyper-parameters of Transformer as [8], including the hidden size, the number of attention heads, etc. When training our proposed model, we use Adam and early stopping with a batch size of 16 and an initial learning rate of 1e5. We varied the learning rate during training following [18] and use a batch size of 1 and a maximum of 30 decoding steps during testing and inference.

4.3 Evaluations

We evaluate our models from two aspects, i.e. automatic and human evaluations.

Automatic Evaluation. To evaluate the performance of KEEM, we first adopt Emotion Accuracy, i.e. the accuracy of emotion detection. The **Perplexity** [16] is also utilized to measure the high-level general quality of the generation model. A response with higher confidence will result in a lower perplexity. Furthermore, **Distinct-1** and **Distinct-2** [6] are used to measure the proportion of the distinct unigrams and bigrams in all the generated results, which indicate the diversity of the produced responses.

[1] https://pytorch.org/.

[2] https://github.com/stanfordnlp/GloVe.

Table 1. Results of automatic evaluation.

Models	Accuracy (%)	Perplexity	Distinct-1	Distinct-2
MoEL	32.00	38.04	0.44	2.10
EmpDG	34.31	37.29	0.46	2.02
MIME	34.24	37.09	0.47	1.90
KEMP	39.31	36.89	0.55	2.29
CEM	39.11	**36.11**	0.66	2.99
KEEM (ours)	**40.54**	36.28	**0.72**	**3.10**
w/o Cog	40.01	36.21	0.56	2.32
w/o Aff	39.50	36.25	0.67	3.02

The results of the automatic and manual evaluations are shown in Table 1. In Table 1, we observe that KEEM achieves the highest emotion accuracy, which suggests the new strategy of selecting affective knowledge is beneficial for users' emotion detection. Although CEM [15] gets a slightly lower perplexity score than ours, our proposed model also considerably outperforms the baselines in terms of Distinct-1 and Distinct-2, which highlights the importance of the novel approaches to incorporating commonsense knowledge and constructing the cognitive context graph.

Human Evaluation. For qualitative evaluation, we take human A/B tests to compare KEEM and five baselines, following [15]. For a given dialogue context, our model's response is paired with a response from the baselines, and annotators are asked to choose the better one from the following three aspects: 1) **Empathy**: which one shows more understanding of the user's situation and feelings; 2) **Coherence**: which one is more on-topic and relevant to the context; 3) **Fluency**: which one is more fluent and natural. We randomly sample 100 dialogues and their corresponding results from our model as well as the baselines, and then assign three crowdsourcing workers to annotate each pair.

As displayed in Table 2, responses generated by KEEM are more often preferred by human judges in empathy and coherence compared to the baselines. This also demonstrates that, with the enhancement of commonsense and emotional knowledge, our model is able to produce more empathetic and relevant responses. We also notice that KEEM does not significantly outperform the baselines not enriched by external knowledge in fluency, which might imply that the knowledge incorporated has a negative influence on fluency. There is one reasonable explanation that selected knowledge contains not only useful information but also noises, which decrease the fluency of the responses.

Table 2. Results of human evaluation (%).

Comparisons	Aspects	Win	Lose	Tie
KEEM vs. MoEL	Empathy	32	9	59
	Coherence	51	8	41
	Fluency	36	33	31
KEEM vs. EmpDG	Empathy	36	10	54
	Coherence	53	10	37
	Fluency	34	35	31
KEEM vs. MIME	Empathy	50	9	41
	Coherence	39	9	52
	Fluency	24	27	49
KEEM vs. KEMP	Empathy	40	26	34
	Coherence	44	23	33
	Fluency	34	36	30
KEEM vs. CEM	Empathy	47	35	29
	Coherence	45	31	24
	Fluency	24	18	58

4.4 Ablation Study

We conduct the ablation study to verify the effect and contribution of each component of our KEEM. More specifically, we consider the following three variants of KEEM:

- **w/o Cog**: We remove the procedures in Sect. 3.2) and delete the concatenation of the global and original cognitive representation of dialogue context (Eq. 15). The fused cognitive representation is replaced with the latter in subsequent calculations.
- **w/o Aff**: We ablate the new approach to filtering affective knowledge (Eq. 7). The effects of introducing affective knowledge and building and encoding affective context graphs have been proven in [8].

The results of the above two variants are in Table 1, which show that each component contributes to KEEM from different aspects. Specifically, removing the cognitive knowledge decreases most of the performance, suggesting that incorporating extra cognitive information helps to recognize the users' situations and identify their emotions, to varying degrees. And ablating the novel strategy of selecting affective knowledge influences the emotion classification considerably and impairs the quality of the generated results slightly.

4.5 Case Study

Cases from KEEM and five baseline models are listed in Table 3. In the first case, KEEM shows the best recognition of the user's action and feeling, by incorporating the concepts of “destroy”, “joy”, and “great”, which correspond to the word “fell”, “cheer”, “nice” in dialogue context. In contrast, the baseline MoEL only attends to “baby” and ignores what the user did for the baby, and the other four baselines are only able to identify the positivity of the dialogue context, producing some context-unrelated content. In the second case, KEEM generates the most context-consistent and emotion-appropriate response, which expresses a good wish for the user about his future promotion, while MoEL and EmpDG do not even detect the user’s hopeful mood. Both two cases demonstrate that KEEM can generate responses of empathy.

Table 3. Responses generated by different models.

Emotion	**caring**
Context	**Speaker**:So my friend’s baby fell down and scraped his knee yesterday and i was trying so hard to cheer him up. **Listener**:That is nice of you. What did you do? **Speaker**:I was singing nursery songs :)
Gold	**Which one?**
MoEL	Oh no, I bet he was a good kid.
EmpDG	That is a great attitude to have.
MIME	That is a good thing to do. I am glad you have a good friend!
KEMP	That is a great feeling. I am sure you will be fine.
CEM	That is awesome! I bet you were very proud of him.
Sem. Know.	**(fell, RelatedTo, destroy)**, **(scraped, RelatedTo, destroy)**
Emo. Know.	**(nice, great)**, **(cheer, joy)**
KEEM	**That is a great thing to do. I am sure he’ll be fine.**
Emotion	**hopeful**
Context	**Speaker**:I guess I will be the next team leader. **Lister**:Really? Is that a promotion? **Speaker**:Yes. I was expecting that very much.
Gold	Good for you.
MoEL	I am not sure what you are going to do.
EmpDG	What kind of leader you?
MIME	That is awesome! I hope you get it!
KEMP	I am sure you will get a new one.
CEM	That is great!
Sem. Know.	**(promotion, RelatedTo, leader)**, **(leader, RelatedTo, promotion)**
Emo. Know.	**(next, new)**, **(leader, hero)**
KEEM	**I am sure you will do well.**

5 Related Work

With the support of the newly proposed datasets [14,19], research in empathetic dialogue generation has developed rapidly. Rashkin et al. [14] exploit adaptions of dialogue models for empathetic responding. Lin et al. [9] design listeners (i.e. decoders) in responding to different emotions and softly combining different listeners' outputs. Li et al. [7] propose a multi-resolution adversarial model to capture the nuances of user emotion and consider the potential of user feedback. Majumder et al. [10] believe that empathic responses often mimic users' emotions to varying degrees. Li et al. [8] construct an emotional context graph to perceive implicit emotions and learn emotional interactions. All of them are about how to perceive and express emotions.

Recently, some work [15] has attempted to boost both the cognitive and affection empathy of dialogue models. Sabour et al. [15] argue that the cognitive empathy of the user's situation should be considered, and they introduce commonsense to further enhance it. Nonetheless, due to some weaknesses of the knowledge base used (e.g. the limited commonsense relations, the unsatisfactory inference accuracy, and the inability to acquire fine-grained information), the effectiveness of this proposed approach would be impacted.

6 Conclusion

In this paper, we propose a novel Knowledge-Enhanced EMpathetic dialogue generation model (KEEM) to demonstrate how leveraging commonsense and emotional knowledge is beneficial to the cognition of users' situations and the detection of users' feelings, which helps produce more empathetic responses. We conduct experiments on EMPATHETICDIALOGUES dataset, and our automatic and manual evaluations have empirically proven the significance of our approach in empathetic response generation. Nevertheless, as the results demonstrate, the model still has shortcomings in terms of fluency, which is one of the future directions for us to challenge.

Acknowledgements. The authors would like to thank the Associate Editor and anonymous reviewers for their valuable comments and suggestions. This work is funded in part by the National Natural Science Foundation of China under Grants No.62176029. This work also is supported in part by the Chongqing Technology Innovation and Application Development Special under Grants CSTB2022TIAD-KPX0206. Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

1. Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., Yakhnenko, O.: Translating embeddings for modeling multi-relational data. In: Advances in Neural Information Processing Systems, vol. 26 (2013)
2. Cuff, B., Brown, S.J., Taylor, L., Howat, D.J.: Empathy: a review of the concept. Emot. Rev. **8**, 144–153 (2014)
3. Elliott, R., Bohart, A.C., Watson, J.C., Murphy, D.: Therapist empathy and client outcome: an updated meta-analysis. Psychotherapy **55**(4), 399 (2018)
4. Jiang, S., Ren, P., Monz, C., de Rijke, M.: Improving neural response diversity with frequency-aware cross-entropy loss. In: The World Wide Web Conference, pp. 2879–2885 (2019)
5. Keskin, S.C.: From what isn't empathy to empathic learning process. Procedia. Soc. Behav. Sci. **116**, 4932–4938 (2014)
6. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.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, pp. 110–119 (2016)
7. 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, pp. 4454–4466 (2020)
8. Li, Q., Li, P., Ren, Z., Ren, P., Chen, Z.: Knowledge bridging for empathetic dialogue generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 10993–11001 (2022)
9. 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 (EMNLP-IJCNLP), pp. 121–132 (2019)
10. Majumder, N., et al.: Mime: mimicking emotions for empathetic response generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8968–8979 (2020)
11. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. Computer Science (2013)
12. Mohammad, S.: Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (volume 1: Long Papers), pp. 174–184 (2018)
13. 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), pp. 1532–1543 (2014)
14. 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, pp. 5370–5381 (2019)
15. Sabour, S., Zheng, C., Huang, M.: CEM: commonsense-aware empathetic response generation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 36, pp. 11229–11237 (2022)
16. Serban, I.V., Sordoni, A., Bengio, Y., Courville, A., Pineau, J.: Hierarchical neural network generative models for movie dialogues. arXiv preprint arXiv:1507.04808 **7**(8), 434–441 (2015)

17. Speer, R., Chin, J., Havasi, C.: Conceptnet 5.5: an open multilingual graph of general knowledge. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 31 (2017)
18. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
19. Welivita, A., Pu, P.: A taxonomy of empathetic response intents in human social conversations. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 4886–4899 (2020)
20. Zech, E., Rimé, B.: Is talking about an emotional experience helpful? effects on emotional recovery and perceived benefits. Clin. Psychol. Psychoth. Int. J. Theory Pract. **12**(4), 270–287 (2005)
21. Zhong, P., Wang, D., Miao, C.: Knowledge-enriched transformer for emotion detection in textual conversations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 165–176 (2019)