Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection

Lanrui Wang1,2, Jiangnan Li1,2, Zheng Lin1,2, Fandong Meng3, Chenxu Yang1,2, Weiping Wang1, Jie Zhou3

1Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
2School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
3Pattern Recognition Center, WeChat AI, Tencent Inc, China

{wanglanrui,lijiangnan,linzheng,yangchenxu,wangweiping}@iie.ac.cn
{fandongmeng,withtomzhou}@tencent.com

Abstract

Empathy, which is widely used in psychological counselling, is a key trait of everyday human conversations. Equipped with commonsense knowledge, current approaches to empathetic response generation focus on capturing implicit emotion within dialogue context, where the emotions are treated as a static variable throughout the conversations. However, emotions change dynamically between utterances, which makes previous works difficult to perceive the emotion flow and predict the correct emotion of the target response, leading to inappropriate response. Furthermore, simply importing commonsense knowledge without harmonization may trigger the conflicts between knowledge and emotion, which confuse the model to choose incorrect information to guide the generation process. To address the above problems, we propose a Serial Encoding and Emotion-Knowledge interaction (SEEK) method for empathetic dialogue generation. We use a fine-grained encoding strategy which is more sensitive to the emotion dynamics (emotion flow) in the conversations to predict the emotion-intent characteristic of response. Besides, we design a novel framework to model the interaction between knowledge and emotion to generate more sensible response. Extensive experiments on EMPATHETICIALOGUES demonstrate that SEEK outperforms the strong baselines in both automatic and manual evaluations.1

1 Introduction

Enriching dialogue systems with human characteristics and capabilities is a hotspot in the humanlike dialogue system research area. Empathy, which is used extensively in psychological counselling (Sharma et al., 2021; Liu et al., 2021; Sharma et al., 2020), is a key trait of everyday human conversations. In contrast to generating responses with controlled emotions (Zhou et al., 2018; Zheng et al., 2021), the key to the empathetic dialogue system is to understand the user’s emotions and generate appropriate responses. Several works concentrate on improving the empathetic models’ ability to capture contextual emotions by emotion mimicry (Majumder et al., 2020), feedback-based adversarial generating (Li et al., 2019), or the mixture of experts (Lin et al., 2019). On the other hand, Sabour et al. (2021); Li et al. (2020) introduce commonsense knowledge into empathetic models so as to better perceive implicit semantic information and generate more informative and empathetic response.

However, the existing works are all about the dialogue-level emotional perception (Lin et al., 2019; Majumder et al., 2020; Li et al., 2019; Sabour...
et al., 2021; Li et al., 2020). Since emotions change dynamically throughout conversations, the coarse modeling method at the dialogue level (recognizing the emotion of the whole conversation context) cannot capture the process of emotional dynamics and makes it difficult to predict response emotions. Welivita and Pu (2020) have studied the shifting pattern of the utterances and drawn two graphs to show the most common emotion-intent flow patterns (with a frequency \( \geq 5 \)) throughout the first four dialogue turns and the global exchanging trends of emotion-intent between speakers and listeners in the EMPATHETICDIALOGUES dataset. For instance, in the first case illustrated in Fig. 1, the speaker’s emotion shifts from afraid at the beginning of the conversation to an embarrassed self-deprecation about previous experience of fearing heights (sharing such a funny story). Accordingly, it is much better that the dialogue agent should express the same self-deprecating sentiment like the gold response. Nevertheless, the baseline models have difficulty capturing subtle changes in the speaker’s emotions and can only provide response according to the fear detected. Moreover, merely introducing knowledge without making emotionally logical choices may lead to logical conflicts between knowledge and emotion in the generated responses. As illustrated in the second case illustrated in Fig. 1, the CEM (Sabour et al., 2021) model chooses the wrong knowledge and is unable to correctly give empathetic responses with nostalgic overtones, which makes knowledge and emotion come into conflict.

To this end, we propose a Serial Encoding and Emotion-Knowledge interaction (SEEK) method for empathetic dialogue generation. To achieve a more fine-grained perception of emotional dynamics, we use an utterance-level encoding strategy which is more sensitive to the emotion flow in the conversations and able to predict the emotion characteristic of the response. We further introduce two new emotion-intent identification tasks to understand contextual emotion and predict the emotional and intentional trait of responses. For the problem of conflicts between knowledge and emotions, we also design a framework modeling the process of bi-directional interaction between them. Extensive experimental results on the utterance-level annotated EMPATHETICDIALOGUES (ED) dataset (Welivita and Pu, 2020) demonstrate that SEEK outperforms the strong baseline with both automatic and manual evaluation metrics. Our contributions are summarized as follows:

- To the best of our knowledge, our work is the first to model the emotion flow that involves the process of emotional dynamics in the task of empathetic dialogue generation. In addition to the coarse emotion at the dialogue level, we introduce fine-grained emotions at the utterance level.
- By modelling the bi-directional interactive selection process between commonsense knowledge and emotions, we have improved not only the ability to recognize contextual emotions, but also the ability to filter out unreasonable external knowledge, allowing the model to generate more sensible empathetic responses.
- The automatic and manual evaluation on annotated-ED dataset shows that our proposed model is superior to the strong baselines and capable of generating more diverse and sensible empathetic responses.

2 Related Work

In order to control the emotion of the generated response, which is one of the fundamental characteristics of daily conversation, plenty of approaches (Zhou et al., 2018; Zheng et al., 2021; Zhong et al., 2019; Shen and Feng, 2020; Liang et al., 2021) view the target emotion as a guiding information of the models’ generator.

Contrary to controlling the emotion of the target response, the task of empathetic dialogue generation requires that the models learn a proper emotion to express empathy. Numerous researchers have attempted to improve the dialogue models’ ability to respond empathetically. Rashkin et al. (2019) proposed a benchmark and dataset to build and evaluate empathetic dialogue generation models. Lin et al. (2019) learned a precise emotion distribution of the response based on mixture of experts. Majumder et al. (2020) split the emotions into two classes and designed a framework to mimic the target emotion in a certain class. Li et al. (2019) utilized user feedback to build a multi-resolution adversarial training framework. In addition, Kim et al. (2021) and Kim et al. (2022) focused on the keywords and emotion cause of dialogue history to better understand the context-level emotion and
recognize feature transitions between utterances. As well, several datasets (Liu et al., 2021; Welivita et al., 2021) of empathetic dialogue generation have been published for further research. However, most of the current approaches do not pay enough attention to the emotion flow of the conversations.

Commonsense knowledge is widely used to build dialogue systems. Zhong et al. (2021a) utilize Commonsense knowledge graph to gain candidate words for generation. Sabour et al. (2021) adopt COMET (Bosselut et al., 2019), a pre-trained language model to generate commonsense inference for retrieving implicit information of dialogue context. In addition, Li et al. (2020) construct a graph-based framework to encode the context-knowledge graph retrieved on commonsense knowledge base. The knowledge introduced into these models might become a trigger of logical conflicts due to the absence of harmony selection.

3 Methodology

3.1 Task Formulation

The task of empathetic dialogue generation is to generate empathetic responses based on the historical context. Given a dialogue $D$, where the context and the target response are denoted as $C = [C_1, ..., C_{N−1}]$ and $Y$ respectively, with a emotion label of the whole context $e_c$. Additionally, a given sequence of emotion-intent labels $EI = [e_{i1}, ..., e_{iN−1}, e_{iY}]$ of the corresponding utterances in $D$, which includes the 32 emotion categories, and 9 common intent classes. Our goal is to generate the next utterance $Y$, which is fluent and coherent to the context, and express empathy to the speaker’s situation and feelings.

3.2 Utterance and Knowledge Encoder

Utterance Encoding: To get a precise representation of each utterance, we firstly encode the context at the utterance level to extract the contextual information. We employ Transformer (Vaswani et al., 2017) to encode the utterance. The embedding of the input is the sum of the word embedding, positional embedding, and dialogue state embedding. Following previous work, we prepend the utterance $u_i$ with [CLS] token to obtain the utterance input $C_i = [w_{CLS}, w_1, w_2, ..., w_{L_i}]$. The embedding is then fed into the Transformer, and we obtain the representation:

$$H_{U_i} = T_{RS_{Enc}}(EMB_{C_i}),$$  (1)

where $H_{U_i} \in \mathbb{R}^{L_i \times d}$, $L_i$ is the length of the utterance, and $d$ is the hidden size of the encoder. We take the representation of [CLS] to represent the utterance:

$$U_i = H_{U_i}[0].$$  (2)

Knowledge Encoding: In order to generate high-quality commonsense inferences for the corresponding context, we utilize COMET (Bosselut et al., 2019), which is a pre-trained GPT (Radford et al., 2018) language model and fine-tuned on ATOMIC (Sap et al., 2019), to generate five types of commonsense knowledge: the effect of the person (xEffect), the reaction of the person speaking the corresponding sentence (xReact), the intent before the person speaking (xIntent), what the person needs (xNeed), and what the person wants after speaking the sentence (xWant). Appending these five special relation tokens after the utterance and feeding them into COMET, we get 5 commonsense inferences texts for each relation of input utterance and then concatenate them to $K_i$. Similarly, we encode the knowledge text using the same Transformer Encoder, and average the encoded hidden state via mean pooling (Zhong et al., 2021b):

$$H_{K_i} = TRS_{Enc}(K_i)$$

$$K_i = \text{Mean}(H_{K_i})$$  (3)

3.3 Emotion Flow Perceiver

Regarding the task of emotional understanding of each utterance as a tagging task, we use a Bi-LSTM to model the emotion dynamics and the interactions between different utterances for the contextual understanding process.

The input of Bi-LSTM is the concatenation of the encoded utterances and knowledge:

$$a_i = [U_i; K_i],$$

$$\hat{U}_i = \text{BiLSTM}(W_a a_i),$$  (4)

where $W_a \in \mathbb{R}^{2d \times d}$ is a trainable weight, and $\hat{U}_i \in \mathbb{R}^{2d}$ represents the processed utterance representation.

3.3.1 Fine-grained Emotion Recognition

For better understanding of the conversation, we pass $\hat{U}_i$ through a tagging classifier to produce a fine-grained emotion-intent tagging distribution $P_{tag} \in \mathbb{R}^{t}$:

$$P_{tag}(e_{i}) = \text{Softmax}(W_a \hat{U}_i)$$  (5)
Figure 2: An overall architecture of our proposed model.

where \( t \) is the number of emotion-intent categories.

We train the tagging module with the cross-entropy loss between the predicted distribution and the ground truth label for a conversation context:

\[
L_{\text{emo}} = - \sum_{i=1}^{N-1} \log(P_{\text{log}}(e_i)).
\]  

(6)

3.3.2 Response Emotion-Intent Prediction

The shift in emotion and intent in empathetic dialogue conforms to an intuitive pattern. We use the attention mechanism to learn the shift pattern of emotion and intent between utterances.

\[
\hat{h}_{\text{pre}} = \text{attention}(\hat{U}_1, \hat{U}_2, ..., \hat{U}_{N-1}),
\]

\[
P_{\text{pre}} = \text{Softmax}(W_p\hat{h}_{\text{pre}}),
\]  

(7)

where \( \hat{h}_{\text{pre}} \in \mathbb{R}^{2d} \) is the representation of the predicted emotion-intent characteristic of response, and \( W_p \in \mathbb{R}^{2d \times t} \) is the weight vector for the linear layer. \( P_{\text{pre}} \) denotes the predicted distribution of the emotion-intent of the target response, \( t \) is the number of emotion and intent categories.

During training, we then minimize the cross-entropy loss between the emotion-intent distribution of the predicted response \( P_{\text{pre}} \) and the ground truth label \( e_i^N \) of the target response:

\[
L_{\text{pre}} = -\log(P_{\text{pre}}(e_i^N)).
\]  

(8)

3.3.3 Dialogue Emotion Recognition

The sequence of utterances representation not only has the contextual information of utterances themselves but also indicates the emotional trait of the whole dialogue. Similarly, we employ the attention mechanism to summarize the holistic emotion label, based on the sequence \( [\hat{U}_1, \hat{U}_2, ..., \hat{U}_{N-1}] \):

\[
\hat{h}_{\text{dia}} = \text{attention}(\hat{U}_1, \hat{U}_2, ..., \hat{U}_{N-1}),
\]

\[
P_{\text{dia}} = \text{Softmax}(W_d\hat{h}_{\text{dia}}),
\]  

(9)

where \( h_{\text{dia}} \in \mathbb{R}^{2d} \), and \( W_d \in \mathbb{R}^{2d \times q} \) is the weight vector for the linear layer. The \( P_{\text{dia}} \) is the distribution of the dialogue emotion, \( q \) is the number of available emotion categories.

The ground truth label of the dialogue emotion is denoted as \( e^* \). The cross-entropy loss utilized to optimize the process of summarizing the conversational emotion is calculated by:

\[
L_{\text{dia}} = -\log(P_{\text{dia}}(e^*)).
\]  

(10)

3.4 Knowledge Selecting Decoder

Merely introducing commonsense knowledge into empathetic models without making an emotionally logical selection to is not ideal. Sabour et al. (2021) select commonsense inferences with an implicit procedure. On the contrary, our method models the process of bi-directional interactions between emotion and knowledge of the corresponding utterance in the conversations.

We adopt \( s \) layers of Cross-Attention Transformer to perform the harmony of emotion and knowledge. Since the utterance representation sequence \( [\hat{U}_1, \hat{U}_2, ..., \hat{U}_{N-1}] \) passed through the three tasks of emotion, it contains emotional characteristics of the corresponding utterances. The
inputs of Cross-Attention Knowledge Selector are composed of the utterance representation sequence acting as the query vector, the key and value vector which are both the knowledge text generated from the COMET model $\mathcal{K} = [\mathcal{K}_1, \ldots, \mathcal{K}_{N-1}]$. The hidden representation of selected knowledge is as follows:

$$S = \text{Cross-Attention}(\hat{U}, \mathcal{K}, \mathcal{K}),$$

(11)

where $S \in \mathbb{R}^{L_s \times d}$, $L_s$ is the maximum length of the knowledge text, and $d$ is the hidden size of the model.

Afterward, we average the harmonized knowledge via mean pooling (Zhong et al., 2021b):

$$\bar{S} = \text{pooling}(S).$$

(12)

We take the Transformer Decoder as the backbone of the Decoder. We perform a concatenation operation between the averaged harmonized knowledge $S$ and the prediction of response representation $\hat{h}_{pre}$ to get a mixture of these two types of information to represent the [SOS] token:

$$[\text{SOS}] = W_k([S; \hat{h}_{pre}])$$

(13)

where $W_k \in \mathbb{R}^{2d \times d}$ is the weight vector for the linear layer.

At the training stage, we prepend the target response $u_N = [y_1, \ldots, y_T]$ with the [SOS] token and get the final input of the Decoder $Y = [[\text{SOS}], y_1, \ldots, y_T]$. The training loss is the standard negative log-likelihood (NLL) loss on the target response $u_N$:

$$L_{nll} = - \sum_{t=1}^{T} \log(P(y_t|C, y_{<t})).$$

(14)

### 3.5 Training Objectives

During the training process, we need to minimize three classification losses and a response generation loss. The classification losses are weighted equally:

$$L_{cls} = L_{tag} + L_{pre} + L_{dia}.$$  

(15)

In order to improve the diversity of the generated response, we adopt Frequency-Aware Cross-Entropy (FACE) (Jiang et al., 2019) as an additional loss to penalize high-frequency tokens, similar to Sabour et al. (2021):

$$L_{div} = - \sum_{t=1}^{T} \sum_{i=1}^{V} w_i \delta_i(c_i) \log(P(y_t|C, y_{<t})), $$

(16)

where $w_i$ is a frequency weight value of the $i$-th token in the vocabulary $V$, $c_i$ represents a candidate token in the vocabulary and $\delta_i(c_i)$ is a function indicate whether $c_i$ equals to the ground truth token $y_t$.

Lastly, all the parameters for our proposed model are jointly trained and optimized by minimizing the weighted sum of the three mentioned losses:

$$L = \alpha L_{nll} + \beta L_{cls} + \gamma L_{div},$$

(17)

where $\alpha$, $\beta$, and $\gamma$ are hyper-parameters used to balance three losses. In our experiments, we set $\alpha = 1$, $\beta = 1$, and $\gamma = 1.5$.

### 4 Experimental Setup

#### 4.1 Dataset

Our experiments are conducted on the utterance-level annotated EMPATHETIC DIALOGUES (ED) (Rashkin et al., 2019; Welivita and Pu, 2020). ED is a large-scale multi-turn dialogue dataset that contains 25k empathetic conversations between a speaker and a listener. ED provides 32 evenly distributed emotion labels which are common in daily chats. However, the emotion labels of ED dataset are on the context level, there are no explicit signals for utterance-level emotions. Welivita and Pu (2020) annotated ED dataset with 41 new categories of utterance-level emotional and intentional labels, which provide fine-grained information about the empathetic dialogues in ED dataset.

#### 4.2 Baselines

We select several strong baseline models for comparison, including:

**MIME:** Majumder et al. (2020) proposed a Transformer-based model employing mimicry strategy to sample the emotion of target responses based on the detected user emotion. The emotions are separated into two classes (positive and negative). The model utilizes a VAE to get the representations of the mimicking and non-mimicking emotions.

**EmpDG** (Li et al., 2019): An adversarial training framework is composed of an empathetic generator and a semantic-emotional discriminator. The discriminator ensures that the responses generated by the generator are relevant to the context and also empathetic. The converged generator trained on the adversarial framework can generate empathetic responses with high diversity.

**KEMP:** Li et al. (2020) employed a graph encoder to extract the contextual and concept information
Table 1: Automatic Evaluation results of baselines and our model. The improvement of SEEK to four strong baselines is statistically significant (paired t-tests with p-values < 0.05).

<table>
<thead>
<tr>
<th>Models</th>
<th>PPL</th>
<th>Dist-1</th>
<th>Dist-2</th>
<th>DE Acc.</th>
<th>UEI Acc.</th>
<th>REI Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIME</td>
<td>37.08</td>
<td>0.31</td>
<td>1.03</td>
<td>29.38</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>EmpDG</td>
<td>37.77</td>
<td>0.59</td>
<td>2.48</td>
<td>30.03</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KEMP</td>
<td>36.89</td>
<td>0.61</td>
<td>2.65</td>
<td>37.58</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>CEM</td>
<td>37.03</td>
<td>0.66</td>
<td>2.99</td>
<td>36.44</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>SEEK</td>
<td>37.09</td>
<td>0.73</td>
<td>3.23</td>
<td>41.85</td>
<td>34.08</td>
<td>25.67</td>
</tr>
</tbody>
</table>

Table 2: Ablation study of our proposed model SEEK. The best results are marked with bold.

<table>
<thead>
<tr>
<th>Models</th>
<th>PPL</th>
<th>Dist-1</th>
<th>Dist-2</th>
<th>DE Acc.</th>
<th>UEI Acc.</th>
<th>REI Acc.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEEK</td>
<td>37.09</td>
<td>0.73</td>
<td>3.23</td>
<td>41.85</td>
<td>34.08</td>
<td>25.67</td>
</tr>
<tr>
<td>w/o Utter</td>
<td>37.37</td>
<td>0.70</td>
<td>3.13</td>
<td>38.9</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>w/o Res</td>
<td>37.97</td>
<td>0.63</td>
<td>2.74</td>
<td>40.82</td>
<td>50.48</td>
<td>-</td>
</tr>
<tr>
<td>w/o Utter &amp; Res</td>
<td>38.48</td>
<td>0.60</td>
<td>2.70</td>
<td>39.7</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>w/o Emo</td>
<td>37.67</td>
<td>0.61</td>
<td>2.66</td>
<td>41.27</td>
<td>35.88</td>
<td>23.37</td>
</tr>
<tr>
<td>w/o Know</td>
<td>37.35</td>
<td>0.31</td>
<td>1.19</td>
<td>41.07</td>
<td>33.53</td>
<td>25.58</td>
</tr>
<tr>
<td>+ Others know</td>
<td>37.50</td>
<td>0.69</td>
<td>2.88</td>
<td>38.25</td>
<td>34.43</td>
<td>24.45</td>
</tr>
<tr>
<td>+ Context Enc</td>
<td>38.68</td>
<td>0.67</td>
<td>2.60</td>
<td>41.81</td>
<td>32.86</td>
<td>24.45</td>
</tr>
</tbody>
</table>

4.3 Implementation Details

We implement our model using Pytorch (Paszke et al., 2019), and utilize Adam (Kingma and Ba, 2015) optimizer to optimize the model. We use 300-dimensional pre-trained GloVE vectors (Pennington et al., 2014) to initialize the word embeddings, which are shared between the encoder and the decoder. During the training stage, the learning rate is initialized as 0.0001 and we vary the learning rate following Vaswani et al. (2017). Our model is trained on one NVIDIA GeForce RTX 3090 GPU using a batch size of 32 and the early stopping strategy. For other settings, such as dropout rate, maximum decoding steps, and so forth, we keep the same as Sabour et al. (2021). The training time of SEEK is about 3 hours for around 27000 iterations.

4.4 Automatic Evaluation

Since Liu et al. (2016) had proved that some automatic metrics based on word overlapping might be improper to evaluate the dialogue systems, such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), we adopt Perplexity (PPL) and Distinct-n (Dist-n) (Li et al., 2016) as the main automatic metrics of generation quality. For the conversational emotion recognition and our newly introduced two tasks including fine-grained emotion-intent tagging and response emotion-intent prediction, we employ dialogue emotion accuracy (DE Acc.), utterance emotion-intent accuracy (UEI Acc.) and response emotion-intent accuracy (REI Acc.).

To examine whether SEEK can generate more sensible response with fine-grained emotion recognition, we compare the performance of our model with the strong baselines. As shown in Table 1, the diversity scores (Dist-1 and Dist-2) of SEEK outperform all of the baselines, which indicates our models can generate more informative response based on the external knowledge. We attribute this improvement to the knowledge selector and the predicted emotion of the target responses, with which the cross-attention mechanism helps to select the related knowledge based on the contextual information of utterances, and the predicted vector provides...
Table 3: Human evaluation results. We apply Fleiss’s Kappa, denoted as $\kappa$, to measure inter-annotator agreement, where $0.4 < \kappa < 0.6$ indicates moderate agreement.

<table>
<thead>
<tr>
<th>Models</th>
<th>Coh.</th>
<th>Emp.</th>
<th>Flu.</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIME</td>
<td>2.84</td>
<td>2.97</td>
<td>2.87</td>
</tr>
<tr>
<td>EmpDG</td>
<td>2.85</td>
<td>2.78</td>
<td>2.76</td>
</tr>
<tr>
<td>KEMP</td>
<td>2.73</td>
<td>2.80</td>
<td>2.80</td>
</tr>
<tr>
<td>CEM</td>
<td>2.82</td>
<td>2.99</td>
<td>2.75</td>
</tr>
<tr>
<td>SEEK</td>
<td><strong>2.91</strong></td>
<td><strong>3.02</strong></td>
<td><strong>3.07</strong></td>
</tr>
</tbody>
</table>

Table 4: Human A/B test (%) on the three aspects: coherence, empathy, and fluency. The comparison results directly show SEEK outperforms the four baselines models.

<table>
<thead>
<tr>
<th>Comparisons</th>
<th>Aspects</th>
<th>Win</th>
<th>Lose</th>
<th>Tie</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEEK vs. MIME</td>
<td>Coh.</td>
<td><strong>24.3</strong></td>
<td>17.1</td>
<td>58.6</td>
</tr>
<tr>
<td>SEEK vs. MIME</td>
<td>Emp.</td>
<td><strong>31.4</strong></td>
<td>22.2</td>
<td>46.4</td>
</tr>
<tr>
<td>SEEK vs. MIME</td>
<td>Flu.</td>
<td><strong>28.6</strong></td>
<td>25.9</td>
<td>45.5</td>
</tr>
<tr>
<td>SEEK vs. EmpDG</td>
<td>Coh.</td>
<td><strong>32.1</strong></td>
<td>26.3</td>
<td>41.6</td>
</tr>
<tr>
<td>SEEK vs. EmpDG</td>
<td>Emp.</td>
<td><strong>35.5</strong></td>
<td>27.4</td>
<td>37.1</td>
</tr>
<tr>
<td>SEEK vs. EmpDG</td>
<td>Flu.</td>
<td><strong>26.9</strong></td>
<td>22.3</td>
<td>50.8</td>
</tr>
<tr>
<td>SEEK vs. KEMP</td>
<td>Coh.</td>
<td><strong>29.2</strong></td>
<td>25.2</td>
<td>45.6</td>
</tr>
<tr>
<td>SEEK vs. KEMP</td>
<td>Emp.</td>
<td><strong>28.8</strong></td>
<td>19.9</td>
<td>51.3</td>
</tr>
<tr>
<td>SEEK vs. KEMP</td>
<td>Flu.</td>
<td><strong>38.7</strong></td>
<td>15.6</td>
<td>45.7</td>
</tr>
<tr>
<td>SEEK vs. CEM</td>
<td>Coh.</td>
<td><strong>27.3</strong></td>
<td>24.8</td>
<td>47.9</td>
</tr>
<tr>
<td>SEEK vs. CEM</td>
<td>Emp.</td>
<td><strong>33.4</strong></td>
<td>27.5</td>
<td>39.1</td>
</tr>
<tr>
<td>SEEK vs. CEM</td>
<td>Flu.</td>
<td><strong>35.7</strong></td>
<td>21.6</td>
<td>42.7</td>
</tr>
</tbody>
</table>

4.6 Ablation Studies

To study the effect of tasks and modules employed in our model, we remove the newly introduced tasks and the interaction process between emotion and knowledge. Additionally, we replace the knowledge type and encoding strategy respectively. The results are demonstrated in Table 2.

Removing the task of fine-grained Utterance Emotion-Intent tagging and Response Emotion-Intent prediction (w/o Utter, w/o Res, and w/o Utter & Res) causes the drop of accuracy of dialogue emotion recognition and generative quality, as these variants lose the fine-grained understanding of the dialogue and the ability to predict the emotion-intent characteristics of the target response.

The margin between the variant (w/o Emo) without emotional harmonization of the knowledge and SEEK proves the importance of the interaction between knowledge and emotion-intent from the Knowledge Selection module of our model. The variant without knowledge (w/o Know) indicates the importance of external knowledge for the diversity of responses the model generated.

Moreover, the decreased performance by replacing the type of knowledge + Others Know and the encoding strategy + Context Enc shows the superiority of our method. Using Others type of knowledge in our model rather than PersonX results in a considerable decrease in all performance, which indicates that the PersonX type of commonsense helps the model to understand the utterances more effectively. The encoding strategy employed...
by baselines (as the variant + Context Enc used) emphasizes on overall understanding of the whole conversation, ignoring an accurate grasp of utterances, which leads to a decline of performance.

Remarkably, the UEI Accuracy of w/o Utter and REI Accuracy of w/o Res are higher than SEEK. This is possibly due to the noise of the utterance label of annotated ED dataset and the subtle differences between intent categories (e.g. agreeing and acknowledging, counselling and questioning), which means the classification supervision signal of utterances or the response will make the input vector of attention module harder and lose some information of other classes. The loss of information about the hidden states may confuse another classifier and leads to a decrease in accuracy. In any case, although there exists a trade-off between these two tasks, they can simultaneously improve the ability of the model to generate more sensible empathetic responses by modeling the emotion flow.

4.7 Case Study
The first case of figure 1 illustrates how emotion shifts during a multi-turn conversation. For better compares generated responses of our model and the baselines, we show two of the generated result of our model and baselines in Table 5. In the first case, the baselines failed to give responses with nostalgic overtones, similar to the commonsense knowledge demonstrated in figure 1, where CEM choose the wrong knowledge to generate response with a happy emotion and the intent to have fun. On the contrary, SEEK successfully gives a response with more sensitive and accurate emotional perception. Similarly, in the second case, all of the baselines generate responses based on the explicit emotion guilty, without fine-grained understanding which is more accurate. Unlike the baselines, SEEK respond sensitively with sympathizing intent.

We further draw a heat map to illustrate the cross-attention weights of commonsense knowledge in a certain case. The detailed information of that case and analysis will be shown in Appendix A.

5 Conclusion
In this paper, we study the task of empathetic dialogue generation. The strong baselines ignore emotion flow of the conversations. We therefore proposed a Serial Encoding and Emotion-Knowledge interaction (SEEK) method for empathetic dialogue
generation, to predict the correct emotion of the target response by perceiving the emotion flow of the context and harmonizing commonsense knowledge with fine-grained emotions to avoid conflicts. Experiments on the utterance-level annotated EMPATHETICDIALOGUES show that our model outperforms the baselines, and the ablation studies indicate that all the components of our model, the encoding strategy, and the commonsense knowledge work.

In the future, we will focus on further usage (e.g. providing online-emotion aid) of empathetic systems and try to improve normalization capabilities of our model on other datasets.

Limitations

The limitation of our work mainly comes from the shortage of datasets in the task of empathetic dialogue generation. Although there are several newly released large-scale datasets (Liu et al., 2021; Welivita et al., 2021), most of the research can only be carried out on the English corpus EMPATHETICDIALOGUES. Another limitation is the problem of evaluation metrics. As mentioned in Liu et al. (2016), the scores of standard automatic evaluation metrics are not consistent with human evaluation results. The lack of task-specifically automatic metrics makes it troublesome for evaluating empathetic dialogue generation.

Ethical Considerations

The data (Rashkin et al., 2019; Welivita and Pu, 2020) used in our work is all drawn from open-source datasets. The conversations of the dataset are around given emotions and carried out by employed crowd-sourced workers, with no personal privacy issues involved.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61976207, No. 61906187).

References


A More Cases

To show the process of knowledge selection of our proposed model, we clearly show the attention weights on the commonsense knowledge in Table 6. We firstly get the weights matrix from Cross-Attention outputs and search the words in the knowledge text by the index of high-value elements. To directly show the selecting process, we mark the knowledge words based on the color in the heatmap we drew: the higher weight the knowledge words have the darker blue marks them in the table.

In this case, the context of the case is mainly about a couple of parents asking for the gender of the baby in a hospital and the COMET totally model generates 25 commonsense inferences based on it. The speaker reacts excitedly to knowing the gender of their baby which infers something to celebrate, and SEEK chooses the correct knowledge and expresses congratulation.


We asked the doc to put the ultrasound in an envelope so we could record our reaction to the gender reveal. I was very happy when I finally saw it! (Excited)

<table>
<thead>
<tr>
<th>Type</th>
<th>x_intent</th>
<th>x_need</th>
<th>x_want</th>
<th>x_effect</th>
<th>x_react</th>
</tr>
</thead>
<tbody>
<tr>
<td>Knowledge</td>
<td>to see the baby</td>
<td>to have an ultrasound</td>
<td>to see what the baby is</td>
<td>to see the baby</td>
<td>happy</td>
</tr>
<tr>
<td></td>
<td>to know the gender</td>
<td>to see the ultrasound</td>
<td>to show it to their friends</td>
<td>to see the gender</td>
<td>excited</td>
</tr>
<tr>
<td></td>
<td>to know the sex</td>
<td>to have the ultrasound</td>
<td>to show it to everyone</td>
<td>to see the ultrasound</td>
<td>surprised</td>
</tr>
<tr>
<td></td>
<td>to be informed</td>
<td>to have a baby</td>
<td>to show it to others</td>
<td>to be happy</td>
<td>joyful</td>
</tr>
<tr>
<td>none</td>
<td>to get the ultrasound</td>
<td></td>
<td>to see the baby</td>
<td>we get excited</td>
<td>relieved</td>
</tr>
</tbody>
</table>

| Context | We asked the doc to put the ultrasound in an envelope so we could record our reaction to the gender reveal. I was very happy when I finally saw it! (Excited) |
| SEEK | Congratulations! |
| Gold | Congrats! what gender did your child end up being? |

Table 6: The visualization of the cross-attention weights of selecting knowledge in SEEK.