Empathetic Dialogue Generation via Sensitive Emotion Recognition and Sensible Knowledge Selection

Lanrui Wang[1,2], **Jiangnan Li**[1,2], **Zheng Lin**[1,2]*, **Fandong Meng**[3], **Chenxu Yang**[1,2], **Weiping Wang**[1], **Jie Zhou**[3]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[3]Pattern 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 EMPATHETICDIALOGUES 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.,

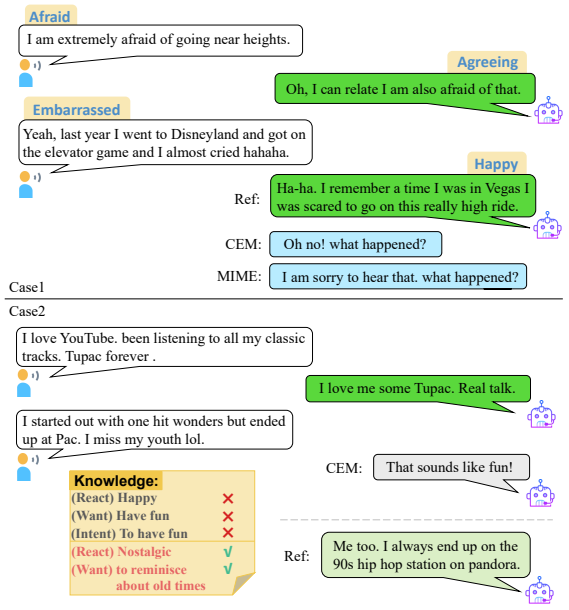

Figure 1: Two cases of multi-turn Empathetic Dialogues. The first case shows the speaker's emotion went from fear at the beginning of the conversation to an embarrassed self-deprecation, ending with a happy mood. And the second case shows that CEM chooses the wrong knowledge leading to inappropriate response.

* Zheng Lin is the corresponding author.
[1]The code is available at https://github.com/wlr737/EMNLP2022-SEEK

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 ≥ 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 $\boldsymbol{EI} = [\boldsymbol{ei}_1, ..., \boldsymbol{ei}_{N-1}, \boldsymbol{ei}_Y]$ 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:

$$\boldsymbol{H}_{U_i} = \mathbf{TRS}_{Enc}(EMB_{C_i}), \tag{1}$$

where $\boldsymbol{H}_{U_i} \in \mathbb{R}^{L_n \times d}$, L_n 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:

$$\boldsymbol{U}_i = \boldsymbol{H}_{U_i}[0]. \tag{2}$$

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 $\boldsymbol{\mathcal{K}}_i$. Similarly, we encode the knowledge text using the same Transformer Encoder, and average the encoded hidden state via mean pooling (Zhong et al., 2021b):

$$\begin{aligned} \boldsymbol{H}_{K_i} &= \mathbf{TRS}_{Enc}(\boldsymbol{\mathcal{K}}_i) \\ \boldsymbol{K}_i &= \mathbf{Mean}(\boldsymbol{H}_{K_i}) \end{aligned} \tag{3}$$

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:

$$\begin{aligned} \boldsymbol{a}_i &= [\boldsymbol{U}_i; \boldsymbol{K}_i], \\ \hat{\boldsymbol{U}}_i &= \mathbf{BiLSTM}(\boldsymbol{W}_a \boldsymbol{a}_i), \end{aligned} \tag{4}$$

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

3.3.1 Fine-grained Emotion Recognition

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

$$P_{tag}(\boldsymbol{ei}_i) = \text{Softmax}(\boldsymbol{W}_{\boldsymbol{e}} \hat{\boldsymbol{U}}_i) \tag{5}$$

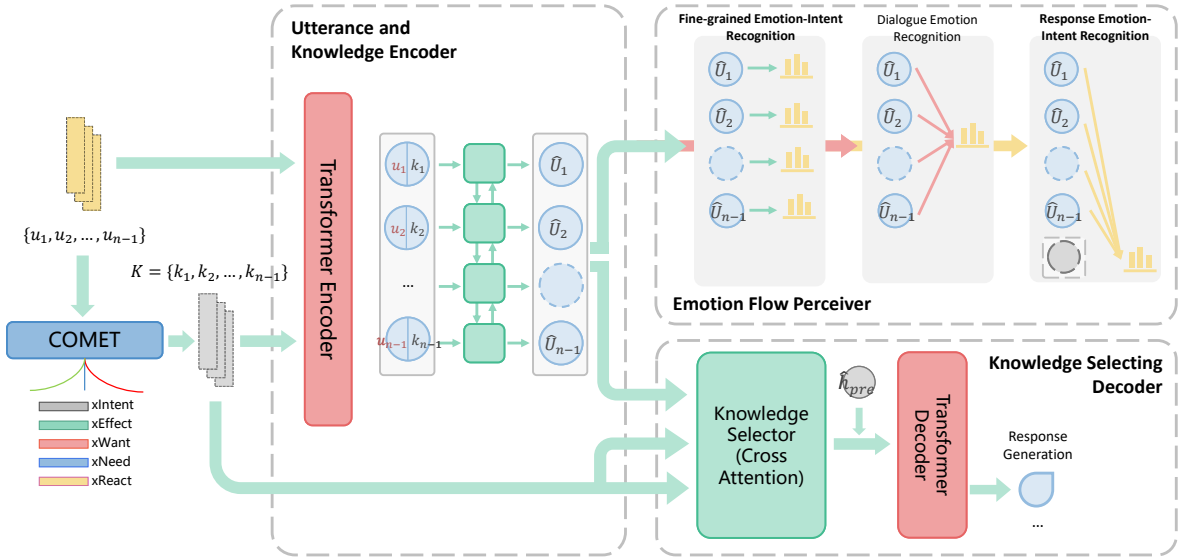

Figure 2: An overall architecture of our proposed model.

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:

$$\mathcal{L}_{emo} = -\sum_{i=1}^{N-1} log(P_{tag}(\boldsymbol{ei}_i)). \quad (6)$$

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.

$$\begin{aligned} \hat{\mathbf{h}}_{pre} &= \text{attention}([\hat{\boldsymbol{U}}_1, \hat{\boldsymbol{U}}_2, ..., \hat{\boldsymbol{U}}_{N-1}]), \\ P_{pre} &= \text{Softmax}(\boldsymbol{W_p}\hat{\mathbf{h}}_{pre}), \end{aligned} \quad (7)$$

where $\hat{\mathbf{h}}_{pre} \in \mathbb{R}^{2d}$ is the representation of the predicted emotion-intent characteristic of response, and $\boldsymbol{W}_p \in \mathbb{R}^{2d \times t}$ is the weight vector for the linear layer. P_{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_{pre} and the ground truth label $\boldsymbol{ei}_N$ of the target response :

$$\mathcal{L}_{pre} = -log(P_{pre}(\boldsymbol{ei}_N)). \quad (8)$$

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 $[\hat{\boldsymbol{U}}_1, \hat{\boldsymbol{U}}_2, ..., \hat{\boldsymbol{U}}_{N-1}]$:

$$\begin{aligned} \hat{\mathbf{h}}_{dia} &= \text{attention}([\hat{\boldsymbol{U}}_1, \hat{\boldsymbol{U}}_2, ..., \hat{\boldsymbol{U}}_{N-1}]), \\ P_{dia} &= \text{Softmax}(\boldsymbol{W_d}\hat{\mathbf{h}}_{dia}), \end{aligned} \quad (9)$$

where $\mathbf{h}_{dia} \in \mathbb{R}^{2d}$, and $\boldsymbol{W}_d \in \mathbb{R}^{2d \times q}$ is the weight vector for the linear layer. The P_{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:

$$\mathcal{L}_{dia} = -log(P_{dia}(\boldsymbol{e}^*)). \quad (10)$$

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 $[\hat{\boldsymbol{U}}_1, \hat{\boldsymbol{U}}_2, ..., \hat{\boldsymbol{U}}_{N-1}]$ 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 $\mathcal{K} = [\mathcal{K}_1, ...\mathcal{K}_{N-1}]$. The hidden representation of selected knowledge is as follows:

$$\boldsymbol{\mathcal{S}} = \text{Cross-Attention}(\hat{\boldsymbol{U}}, \mathcal{K}, \mathcal{K}), \quad (11)$$

where $\boldsymbol{\mathcal{S}} \in \mathbb{R}^{L_s \times d}$, 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):

$$\boldsymbol{S} = \text{pooling}(\boldsymbol{\mathcal{S}}). \quad (12)$$

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

$$[\text{SOS}] = \boldsymbol{W}_k([\boldsymbol{S}; \hat{\mathbf{h}}_{pre}]) \quad (13)$$

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

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

$$\mathcal{L}_{nll} = -\sum_{t=1}^{T} log(P(y_t|C, y_{<t}). \quad (14)$$

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:

$$\mathcal{L}_{cls} = \mathcal{L}_{tag} + \mathcal{L}_{pre} + \mathcal{L}_{dia}. \quad (15)$$

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):

$$\mathcal{L}_{div} = -\sum_{t=1}^{T}\sum_{i=1}^{V} w_i \delta_t(c_i) log(P(y_t|C, y_{<t}), \quad (16)$$

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 $\delta_t(c_i)$ 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:

$$\mathcal{L} = \alpha\mathcal{L}_{nll} + \beta\mathcal{L}_{cls} + \gamma\mathcal{L}_{div}, \quad (17)$$

where α, β, and γ are hyper-parameters used to balance three losses. In our experiments, we set α= 1, β= 1, and γ= 1.5.

4 Experimental Setup

4.1 Dataset

Our experiments are conducted on the utterance-level annotated EMPATHETICDIALOGUES (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

Models	PPL	Dist-1	Dist-2	DE Acc.	UEI Acc.	REI Acc.
MIME	37.08	0.31	1.03	29.38	-	-
EmpDG	37.77	0.59	2.48	30.03	-	-
KEMP	**36.89**	0.61	2.65	37.58	-	-
CEM	37.03	0.66	2.99	36.44	-	-
SEEK	37.09	**0.73**	**3.23**	**41.85**	**34.08**	**25.67**

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).

Models	PPL	Dist-1	Dist-2	DE Acc.	UEI Acc.	REI Acc.
SEEK	**37.09**	**0.73**	**3.23**	**41.85**	34.08	25.67
w/o Utter	37.37	0.70	3.13	38.9	-	**30.41**
w/o Res	37.97	0.63	2.74	40.82	**50.48**	-
w/o Utter & Res	38.48	0.60	2.70	39.7	-	-
w/o Emo	37.67	0.61	2.66	41.27	35.88	23.37
w/o Know	37.35	0.31	1.19	41.07	33.53	25.58
+ Others know	37.50	6.90	2.88	38.25	34.43	24.32
+ Context Enc	38.68	0.67	2.60	41.81	32.86	24.45

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

of the context graph constructed on external knowledge. The knowledge-enriched context graph contains emotional dependencies which helps to understand the emotion characteristic of conversations.

CEM: Sabour et al. (2021) use COMET to generate commonsense knowledge based on the last utterance said by the speaker in dialogue. The authors use five specific prefixes (xIntent, xEffect, xWant, xNeed, xReact) to obtain five types of knowledge corresponding to the last utterance. The model can generate more informative empathetic responses.

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 initialed 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

Models	Coh.	Emp.	Flu.
MIME	2.84	2.97	2.87
EmpDG	2.85	2.78	2.76
KEMP	2.73	2.80	2.80
CEM	2.82	2.99	2.75
SEEK	**2.91**	**3.02**	**3.07**

Table 3: Human evaluation results. We apply Fleiss's Kappa, denoted as κ, to measure inter-annotator agreement, where $0.4 < \kappa < 0.6$ indicates moderate agreement.

additional information of the generating process.

To prove if SEEK has better understanding of the dialogue emotion, we list the accuracy of the baselines and our proposed model. Remarkably, SEEK surpasses all of the baselines by a large margin, we attribute the increase of performance to the two fine-grained tasks we introduced. The better comprehension of the utterances in dialogue, the more accuracy it takes. In terms of the two new accuracy scores, UEI Accuracy and REI Accuracy, SEEK reaches satisfying performances, as the number of the categories of these two tasks are 41.

4.5 Human Evaluation

Following previous works, we conduct a human evaluation based on three aspects: *coherence* (**Coh.**): How much does the response relevant to the context? *empathy* (**Emp.**): How much does the model know about the speaker's situation and emotion characteristic? Does the model respond empathetically enough or give suggestions? *fluency* (**Flu.**): How much the generated response obey the grammar? We randomly choose 100 dialogues and assign the responses generated by the models to three crowd-sourced workers for the evaluation. Each aspect is on a scale of 1 to 5. Moreover, considering the variation between different individuals, we conduct another human A/B test to directly compare our method with other baselines. Three professional annotators score the questionnaire of the response pairs to choose one of the responses in random order or select "Tie" when the quality of provided sentence is difficult to distinguish. As the results of the human rating and A/B test are shown in Table 3 and table 4, SEEK outperforms the baselines in all the three aspects.

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

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

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(**w/o Emo**). 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

Context	*Speaker:* I love YouTube. I've been listening to all my **classic** tracks. Tupac forever. (**Nostalgic**)
	Listener: I love me some Tupac. Real talk. (**Acknowledging**)
	Speaker: I started out with *One Hit Wonders* but ended up at Pac. **I miss my youth lol**. (**Nostalgic**)
Speaker Emotion	Nostalgic
MIME	I am sure you will be able to get a new one.
EmpDG	I am sure you will be fine.
KEMP	I am sure you will be fine.
CEM	That sounds like fun!
SEEK	I do not like that, but i think they will be able to get the old times . (**Nostalgic**)
Golden	Me too. I always end up on the 90s hip hop station on Pandora. (**Nostalgic**)

Context	*Speaker:* Yeah about 10 years ago I had a horrifying experience. It was 100% their fault, but they hit the water barrels and survived. They had no injuries, but they almost ran me off the road. (**Guilty**)
	Listener: Did you suffer any injuries? (**Questioning**)
	Speaker: No, I was not hit. It turned out they were drunk. I felt guilty, but realized it was his fault. (**Guilty**)
Speaker Emotion	Guilty
MIME	I would be too. I would be so scared.
EmpDG	I am sorry to hear that.
KEMP	I am sorry to hear that.
CEM	That is so sad.
SEEK	That is pretty scary! I am glad that you were able to get in out! (**Sympathizing**)
Golden	Why did you feel guilty? People really shouldn't drive drunk. (**Questioning**)

Table 5: Two cases of generated responses by SEEK and the baselines. We annotated each turn with the emotional or intentional labels at the end of the utterances. The words relevant to the predicted labels in SEEK's response are highlighted in red.

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 Table 5 compares some generated responses of our model and the baselines. In the first case, the baselines failed to give responses with nostalgic overtones. As the commonsense knowledge demonstrated in figure 1, 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

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 2879–2885. ACM.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 2227–2240. Association for Computational Linguistics.

Wongyu Kim, Youbin Ahn, Donghyun Kim, and Kyong-Ho Lee. 2022. Emp-rft: Empathetic response generation via recognizing feature transitions between utterances. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 4118–4128. Association for Computational Linguistics.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, pages 110–119. The Association for Computational Linguistics.

Qintong Li, Hongshen Chen, Zhaochun Ren, Zhumin Chen, Zhaopeng Tu, and Jun Ma. 2019. Empgan: Multi-resolution interactive empathetic dialogue generation. *CoRR*, abs/1911.08698.

Qintong Li, Piji Li, Zhumin Chen, and Zhaochun Ren. 2020. Empathetic dialogue generation via knowledge enhancing and emotion dependency modeling. *CoRR*, abs/2009.09708.

Yunlong Liang, Fandong Meng, Ying Zhang, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021. Infusing multi-source knowledge with heterogeneous graph neural network for emotional conversation generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI*

2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021, pages 13343–13352. AAAI Press.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. 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)*, pages 121–132, Hong Kong, China. Association for Computational Linguistics.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2122–2132. The Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *CoRR*, abs/2106.01144.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: MIMicking emotions for empathetic response generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979, Online. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. *CoRR*, abs/1912.01703.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543. ACL.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: commonsense-aware empathetic response generation. *CoRR*, abs/2109.05739.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press.

Ashish Sharma, Inna W. Lin, Adam S. Miner, David C. Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. *CoRR*, abs/2101.07714.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *CoRR*, abs/2009.08441.

Lei Shen and Yang Feng. 2020. CDL: curriculum dual learning for emotion-controllable response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 556–566. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Anuradha Welivita and Pearl Pu. 2020. A taxonomy of empathetic response intents in human social conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics.

Anuradha Welivita, Yubo Xie, and Pearl Pu. 2021. A large-scale dataset for empathetic response generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1251–1264. Association for Computational Linguistics.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: A multi-factor hierarchical framework for empathetic response generation. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 813–824. Association for Computational Linguistics.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021a. CARE: commonsense-aware emotional response generation with latent concepts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14577–14585. AAAI Press.

Peixiang Zhong, Di Wang, Pengfei Li, Chen Zhang, Hao Wang, and Chunyan Miao. 2021b. CARE: commonsense-aware emotional response generation with latent concepts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 14577–14585. AAAI Press.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. An affect-rich neural conversational model with biased attention and weighted cross-entropy loss. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7492–7500. AAAI Press.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 730–739. AAAI Press.

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 heat map 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.

Type	x_intent	x_need	x_want	x_effect	x_react
Knowledge	to see the baby	to have an ultrasound	to see what the baby is	to see the baby	happy
	to know the gender	to see the ultrasound	to show it to their friends	to see the gender	excited
	to know the sex	to have the ultrasound	to show it to everyone	to see the ultrasound	surprised
	to be informed	to have a baby	to show it to others	to be happy	joyful
	none	to get the ultrasound	to see the baby	we get excited	relieved
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.