Using Extracted Emotion Cause to Improve Content-Relevance for Empathetic Conversation Generation

Minghui Zou[*], **Rui Pan**[*], **Sai Zhang**[†], **Xiaowang Zhang**
College of Intelligence and Computing, Tianjin University, Tianjin, China
zhang_sai@tju.edu.cn

Abstract

Empathetic conversation generation intends to endow the open-domain conversation model with the capability for understanding, interpreting, and expressing emotion. Humans express not only their emotional state but also the stimulus that caused the emotion, i.e., emotion cause, during a conversation. Most existing approaches focus on emotion modeling, emotion recognition and prediction, and emotion fusion generation, ignoring the critical aspect of the emotion cause, which results in generating responses with irrelevant content. Emotion cause can help the model understand the user's emotion and make the generated responses more content-relevant. However, using the emotion cause to enhance empathetic conversation generation is challenging. Firstly, the model needs to accurately identify the emotion cause without large-scale labeled data. Second, the model needs to effectively integrate the emotion cause into the generation process. To this end, we present an emotion cause extractor using a semi-supervised training method and an empathetic conversation generator using a biased self-attention mechanism to overcome these two issues. Experimental results indicate that our proposed emotion cause extractor improves recall scores markedly compared to the baselines, and the proposed empathetic conversation generator has superior performance and improves the content-relevance of generated responses.

1 Introduction

Open-domain conversation generation has made remarkable progress over recent years, relying on deep learning and neural networks (Serban et al., 2016; Wolf et al., 2019; Zhou et al., 2020; Huang et al., 2020). However, previous works primarily centre around improving the linguistic quality of the generated responses, such as grammatical correctness, content variety, and topic relevance, neglecting the important factor of emotion (Zhou et al., 2018). The information conveyed by humans during communication contains not only syntactic and semantic information but also emotional information. Emotion is one of the essential cognitive behaviors in humans, and artificial intelligence has the objective of enabling machines to mimic human intelligent behaviors. As an important research branch of AI, one of the long-term goals of dialogue systems is to enable machines to perceive, comprehend and express emotions. In addition, studies (Martinovski and Traum, 2003; Prendinger and Ishizuka, 2005; Prendinger et al., 2005) have shown that introducing emotional information into conversation systems can improve user engagement and satisfaction, make human-computer conversation more natural, and reduce the number of conversation terminations.

As a new research hotspot for the NLP community, most existing approaches on empathetic conversation generation focus on identifying the emotion category of the input sequence and generating a response based on the predicted emotion label. Song et al. (2019) introduce an external emotion lexicon into the generation module to achieve explicit and implicit emotion expression. Li et al. (2020) create a coarse-grained and fine-grained emotion representation for the input sequence by using an emotion classifier to identify the input sequence's emotion category and an external emotion lexicon to identify

[*]These authors contributed equally to this work and should be considered co-first authors.
[†]Corresponding author.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.
(c) Technical Committee on Computational Linguistics, Chinese Information Processing Society of China

the emotion terms. Majumder et al. (2020) improve the empathetic expression in the generated response by mimicking the input sequence's emotion while taking into account their emotional polarity. Firdaus et al. (2021) incorporate multiple emotions to varying degrees in the generation of responses to make the model more anthropomorphic.

Existing works mainly focus on emotion-related issues, focusing less on content-relevance. However, a case study of the responses generated by the existing models shows that the existing models do not guarantee the content-relevance of the generated responses very well. As shown in Table 1, in case 1, EmpDG (Li et al., 2020) generates an emotionally irrational and irrelevant response, whereas GPT2 (Radford et al., 2019) can express empathy for the user's emotion, but the generated response deviates from the conversation's topic (from diet to age); in case 2, EmpDG and GPT2 both focus too much on the user's lonely emotion but fail to develop the conversation around the user's specific situation, resulting in the generation of a safe response, which is also irrelevant to the situation.

Input 1	My son really needs to cut down on the amount of sugary snacks he has!
Emotion	Disappointed
EmpDG	That is so disgusting! I bet he is a very disgusting situation.
GPT2	I know how you feel. I have a son who is 6 and a daughter who is 5.
Input 2	I feel so lonely sometimes because all my friends live in a different country .
Emotion	Lonely
EmpDG	I am sorry you had to go through that.
GPT2	Oh no! I hope you are ok!

Table 1: Examples of empathetic conversation.

We suggest there are two main reasons: Firstly, as Gao et al. (2021) pointed out, existing works deal with emotions on a surface level and do not consider the underlying causes of the emotion, making it difficult to comprehend user's complicated emotions and badly affecting the subsequent links of emotion prediction and empathetic conversation generation. Secondly, emotion category is a strong supervisory signal, and overemphasizing its importance in the process of generating responses can easily lead to the generation of safe responses for specific emotion categories. For examples in Table 1, if the model can accurately capture the emotion cause in the input sequence (as highlighted in yellow) and incorporate them into the process of generating responses, the model will have the ability to understand the user's emotion better and generate responses with more relevant content by developing topics around the facts conveyed by the user during the generation process.

To this end, we propose an empathetic conversation generation model enhanced by emotion cause to improve the content-relevance of generated responses. Specifically, our model involves two components, an emotion cause extractor and an empathetic conversation generator. In order to accurately identify emotion cause in the absence of large-scale labeled data, we present a semi-supervised training method to optimize the emotion cause extractor. To integrate the extracted emotion cause into the empathetic conversation generator and minimize the damage to the general language knowledge already learned by the pre-trained language model, we introduce a biased self-attention mechanism to enhance the model's attention to the emotion cause when generating responses.

The contributions of our work are summarized as follows:

- To compensate for the scarcity of large-scale word-level emotion-cause labeled datasets, a semi-supervised training method using labeled and unlabeled data for joint training is proposed.
- To integrate the extracted emotion cause into the generation process, a biased self-attention mechanism that does not introduce new additional parameters is proposed.
- Experimental results indicate that our proposed model performs superior to the baselines and improves the content-relevance of the generated responses.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

2 Related Work

Empathetic conversation generation has made great progress in recent years. Several works (Song et al., 2019; Shen and Feng, 2020; Welivita and Pu, 2020; Zheng et al., 2021; Sabour et al., 2022; Shen et al., 2021) attempt to make dialogue models more empathetic and have achieved promising results. Song et al. (2019) introduce an external emotion lexicon into the generation module to achieve explicit and implicit emotion expression. Shen et al. (2020) present a novel framework that extends the emotional conversation generation through a dual task and alternatively generates the responses and queries. Welivita et al. (2020) combine dialogue intent modeling and neural response generation to obtain more controllable and empathetic responses. Zheng et al. (2021) propose a multi-factor hierarchical framework to model communication mechanism, dialog act and emotion in a hierarchical way. Sabour et al. (2022) introduce external commonsense information to absorb additional information about the situation and help the model better understand the user's emotion.

Emotion cause extraction is intended to discover the stimulus reasons behind the user's emotion (Lee et al., 2010; Chen et al., 2010). Although there has been a lot of excellent works in this research direction (Xia and Ding, 2019; Bao et al., 2022; Turcan et al., 2021), most of the existing datasets are at the sentence/sub-sentence level (Kim et al., 2021). There is still a lack of a large-scale word-level emotion-cause labeled dataset up till now.

Most existing approaches on empathetic conversation generation only consider superficial emotional information in the dialogue context but ignore deeper emotional causes. Recently, some researches (Gao et al., 2021; Kim et al., 2021) have attempt to investigate emotion cause in empathetic conversation generation, resulting in more relevant and empathetic responses. Since there is no large-scale word-level emotion-cause labeled dataset, Gao et al. (2021) train an emotion cause extractor using a sentence-level labeled dataset and then automatically construct a word-level labeled dataset. Kim et al. (2021) use a Bayesian conditional probability formula based on the emotion category of the dialogue context to train an emotion cause extractor in a weakly supervised way. In order to incorporate emotion cause into the process of generating responses, Gao et al. (2021) introduce a soft gating mechanism and a hard gating mechanism to make model boost the attention on emotion cause; while Kim et al. (2021) introduce the RSA framework, which is essentially a Bayesian conditional probability-based response rewriting module based on the original decoder.

3 Task Formulation

Emotion cause extraction. Given an input sequence $X_e = (x_1, x_2, ..., x_k)$, the goal is to predict the emotion cause probability $C = (c_1, c_2, ..., c_k)$ that indicates whether the token is an emotion cause. Specifically, we add special tokens [CLS] and [SEP] at the beginning and end of the sequence, respectively (as shown in Figure 1).

Empathetic conversation generation. Given an input sequence $X_g = (x_1, x_2, ..., x_n)$, the goal is to generate a response $Y = (y_1, y_2, ..., y_m)$ that is empathetic and relevant to the conversation. Specifically, follow the previous works (Lin et al., 2019; Shin et al., 2020; Gao et al., 2021), we concatenate all utterances in the dialogue context together as input and separate utterances by [SEP] tokens (as shown in Figure 1).

4 Approach

Our proposed emotion-cause-enhanced empathetic conversation generation model consists of two main modules: Emotion Cause Extractor and Empathetic Conversation Generator. The overview is shown in Figure 1. Since there is no large-scale word-level emotion cause dataset available, we present a semi-supervised training method to obtain the emotion cause extractor using small-scale labeled data jointly trained with large-scale unlabeled data. To involve the emotion cause in the generation process, we introduce multiplicative signals to implement the biased self-attention mechanism. The multiplicative signal enhances the model's attention to the emotion cause in the generation process and improves the content-relevance of the generated responses.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

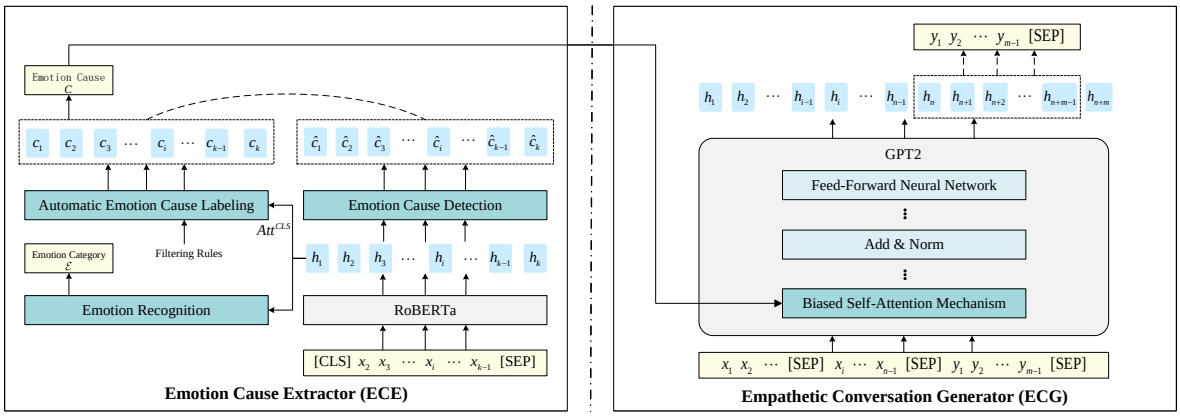

Figure 1: The overview of our proposed ECE and ECG.

4.1 Emotion Cause Extractor

The RoBERTa model (Liu et al., 2019) created by stacking the Transformer encoder (Vaswani et al., 2017) can better model contextual information in both directions. We construct the Emotion Cause Extractor (ECE for short) based on the RoBERTa to identify the emotion categories of the input sequence and its emotion causes. Thus the tasks of the ECE can be divided into emotion recognition and emotion cause detection.

4.1.1 Emotion Recognition

Emotion recognition is a classification problem aiming to predict the emotion category of the input sequence. Given a input sequence X_e, the forward propagation process of the model can be defined as:

$$H_h^E = \text{RoBERTa}(X_e) \tag{1}$$

$$P = \text{softmax}\big(W_e H_{h,1}^E + b_e\big) \tag{2}$$

where H_h^E denotes the output of the last hidden layer, and $H_{h,1}^E$ denotes the output of the first token (i.e., [CLS]) in the last hidden layer. W_e and b_e denote the parameters of the feed-forward neural network.

After obtaining the probability distribution P of emotion category, the emotion category of the X_e can be defined as $\mathcal{E} = \text{argmax}(P)$.

We employ the following loss function to optimize the parameters:

$$\mathcal{L}_{emo}(P) = -\sum_{i \in labels} t(i) \log p_i \tag{3}$$

where $labels \in \{1, 2, ..., s\}$ denotes emotion categories, and $t(i)$ denotes the ground truth distribution corresponding to the input sequence.

It is noted that the input representation of the RoBERTa contains both word embedding and positional embedding:

$$H_0^E = X_e W_e^W + X_e^P W_e^P \tag{4}$$

where W_e^W denotes the word embedding matrix, X_e^P denotes the absolute position of tokens in X_e, and W_e^P denotes the positional embedding matrix.

4.1.2 Emotion Cause Detection

Emotion cause detection is a sequence labeling problem that aims to predict whether each token in the input sequence is the emotion cause, i.e., a word-level $\{0, 1\}$ labeling problem. Since no large-scale word-level emotion cause dataset is available, this section proposes a semi-supervised training method using small-scale labeled data jointly with large-scale unlabeled data.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

For the labeled data, given an input sequence X_e, the context-aware word representation is obtained by encoding using the RoBERTa. Then, a layer of the feed-forward neural network is used for $\{0, 1\}$ sequence labeling:

$$H_h^E = \mathrm{RoBERTa}(X_e) \tag{5}$$

$$\widehat{C} = \mathrm{softmax}\big(W_c H_h^E + b_c\big) \tag{6}$$

where $\widehat{C}$ represents the emotion cause probability of each token, W_c and b_c denote the parameters of the feed-forward neural network.

The loss function applied for parameter learning is as follows:

$$\mathcal{L}_{cau}\left(\widehat{C}\right) = -\sum_{i=1}^{k} \log \mathrm{P}\left(\widehat{C}_i\right) \tag{7}$$

where k indicates the length of the input sequence, and $\mathrm{P}(\cdot)$ denotes obtaining the probability corresponding to the ground truth label of each token.

For the unlabeled data, we observe that the model needs to pay attention to the emotion cause when predicting the emotion category of the input sequence. Thus the attention weight distribution of the model in predicting emotion categories can be used to predict whether each token is an emotion cause or not. Given an input sequence X_e, emotion recognition is performed using the RoBERTa to obtain the attention weight distribution Att^{CLS} of the first [CLS] token in the last hidden layer. Then, simple filtering based on the rules (including removing punctuation, special words, stop words, etc.) is applied, and the tokens with top-k weights are selected as the emotion cause of the input sequence. In this way, emotion cause labels can be automatically constructed for unlabeled data, and the rest of the processing is similar to labeled data.

However, the above method of automatic emotion cause labeling requires converting each token from vector to text at the realization and then performing rule-based filtering. This leads to the fact that the computational graph of automatic emotion cause labeling module is not fully linked with that of emotion cause detection module, i.e., the loss function $\mathcal{L}_{cau}$ of emotion cause detection is not derivable for Att^{CLS}, and cannot be directly involved in the optimization of Att^{CLS}. Thus we propose an additional auxiliary loss function to link the computational graph and introduce the regularization constraint by computing the vector inner product of Att^{CLS} and $\widehat{C^1}$:

$$\mathcal{L}_{aux}\left(Att^{CLS}, \widehat{C}\right) = Att^{CLS} \cdot \widehat{C^1} \tag{8}$$

where $\widehat{C^1} = \widehat{C}\,[1, :]$ denotes the probability that each token is the emotion cause.

In summary, we employ the following loss function to optimize the emotion cause extractor:

$$\mathcal{L}^{ECE} = \lambda_1 \mathcal{L}_{emo} + \lambda_2 \mathcal{L}_{cau} + \lambda_3 \mathcal{L}_{aux} \tag{9}$$

where λ_i indicates the weight of each loss function (we set $\lambda_1 = 1/3$, $\lambda_2 = \lambda_3 = 1$).

4.2 Empathetic Conversation Generator

4.2.1 Conversation Generation

Given a input sequence X_g and the corresponding probability of emotion cause C, the goal of the Empathetic Conversation Generator (ECG for short) is to maximize the probability $P\left(Y|X_g, C\right)$. The empathetic conversation generator proposed in this section is implemented based on the GPT2 (Radford et al., 2019). Forward propagation process of the GPT2 in conversation generation task can be defined as:

$$H_h^G = \mathrm{GPT2}(X_g) \tag{10}$$

$$\widehat{Y} = \mathrm{softmax}\big(W_g H_h^G + b_g\big) \tag{11}$$

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

where W_g and b_g denote the parameters of the feed-forward neural network.

The loss function is as follows:

$$\mathcal{L}^{ECG}\left(\widehat{Y}\right) = -\sum_{i=1}^{m} \log \mathrm{P}\left(\widehat{Y}_i\right) \tag{12}$$

where m denotes the length of the sequence, and $\mathrm{P}(\cdot)$ denotes obtaining the probability corresponding to the ground truth.

It is noted that the input representation of the GPT2 contains three parts: word embedding, positional embedding and role embedding:

$$H_0^G = X_g W_g^W + X_g^P W_g^P + X_g^R W_g^R \tag{13}$$

where X_g^R denotes the role identifier of each token in the input sequence X_g (used to distinguish different speakers), and W_g^R denotes the role embedding matrix.

4.2.2 Biased Self-Attention Mechanism

In order to integrate the emotion cause into the generation progress of the GPT2, it is typical to introduce a new attention mechanism layer. However, considering that the GPT2 has large-scale, trained parameters, if a new attention mechanism layer is introduced in the fine-tuning phase, it may greatly impact the original parameters and destroy the general knowledge already learned by the GPT2. Therefore we chose to introduce multiplicative signals based on emotion cause on top of the original self-attention mechanism of the GPT2 to enhance the model's attention to emotion cause during generation. Meanwhile, the above possible problems are avoided since no additional parameters are introduced.

Moreover, considering that deep neural networks are biased toward modelling syntactic information at the bottom level and semantic information at the top level, the first few layers of the GPT2 network do not require special attention for the emotion cause. We use the layer number information to scale the above multiplicative signals. As the number of layers increases, the multiplicative signals based on the emotion cause gradually strengthen.

The original self-attention mechanism of the GPT2 is defined as:

$$\mathrm{MaskedAttention}(Q, K, V) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M - \lambda\left(I - M\right)\right) V \tag{14}$$

where $\odot$ denotes the multiplication of the corresponding elements of the matrix, λ denotes an infinite scalar (generally taken as $\lambda = 10000$). M denotes the lower triangular matrix with all non-zero elements being 1, I denotes the matrix where all elements are 1.

Our proposed biased self-attention mechanism based on the emotion cause can be defined as:

$$\mathrm{MaskedScore}(Q, K) = \mathrm{softmax}\left(\frac{QK^T}{\sqrt{d_k}} \odot M - \lambda\left(I - M\right)\right) \tag{15}$$

$$\mathrm{BiasedScore}(Q, K) = \mathrm{Normalize}\left(\mathrm{MaskedScore}(Q, K) \odot \left(I + \frac{h_i}{h} C\right)\right) \tag{16}$$

$$\mathrm{Normalize}(X) = \frac{x_{i,j}}{\sum_i x_{i,j}} \tag{17}$$

$$\mathrm{BiasedAttention}(Q, K, V) = \mathrm{BiasedScore}(Q, K)\, V \tag{18}$$

where C represents the probability of each token being an emotion cause, $h_i \in \{1, 2, ..., h\}$ denotes the serial number of the self-attention layer, $\mathrm{Normalize}(\cdot)$ denotes the function for normalization by row.

4.3 Training Strategy

Our proposed model is trained using a two-stage training strategy.

In the first stage, the ECE is trained using a semi-supervised training method, as shown in Algorithm 1.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

Algorithm 1: The training process of ECE

```
Input: ECE, EmoCause-1 dataset and EmpDialog dataset
Loading the RoBERTa and randomly initializing other parameters;
for training iteration do
    for data ∈ EmoCause-1 do
        Train ECE in a supervised method;
    end
    for data ∈ EmpDialog do
        Construct emotion cause labels automatically;
        Train ECE in a supervised method based on the emotion cause labels;
    end
end
Output: ECE
```

In the second stage, the ECG is trained based on the emotion cause extracted by the ECE, and the parameters of the ECE are frozen in this stage. The training process is shown in Algorithm 2.

Algorithm 2: The training process of ECG

```
Input: ECG, ECE and EmpDialog dataset
Loading the ECE;
Loading the GPT2 and randomly initializing other parameters;
for training iteration do
    for data ∈ EmpDialog do
        Extract the emotion cause of the input sequence using ECE;
        Integrate the extracted emotion cause into ECG using biased self-attention mechanism;
        Update the parameters of the ECG;
    end
end
Output: ECG
```

5 Experiments

5.1 Datasets

We use the following two datasets to conduct experiments.

Label	Hopeful	
Situation	I have been making goals each week for earning money. I'm hoping to save enough to start renovations on my house.	
Conversation	Speaker:	I have big renovation plans for my house. I've made a money plan and have kept to it so far.
	Listener:	Well at least you have a plan. Are you planning to start the renovation soon?
	Speaker:	Yes, hopefully it will all go as planned. So far so good.
	Listener:	Awesome. I'm sure it's going to turn out great.

Table 2: An example of the EmpDialog dataset.

EmpatheticDialogues (EmpDialog for short) is a dataset for empathetic conversation generation created by Rashkin et al. (2019). The dataset, which contains 19,533 conversations in the training set, 2770 conversations in the validation set and 2547 conversations in the test set, is collected and created by the Amazon Mechanical Turk platform. EmpDialog defines 32 emotion categories, and each conversation

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

is created based on an emotional category and a situation description. An example of the EmpDialog dataset is shown in Table 2.

Label	Hopeful
Situation	I have been making goals each week for earning money. I'm hoping to save enough to start renovations on my house.
Cause	goals, earning, money

Table 3: An example of the EmoCause dataset.

EmoCause is a word-level emotion cause dataset created by Kim et al. (2021) based on the validation and test sets of EmpDialog. The dataset is also collected and created by the Amazon Mechanical Turk platform. The workers are asked to vote for each token in a given $situation$ to determine whether it is the emotion cause. EmoCause have 2770 validation data and 2547 test data. An example of the EmoCause dataset is shown in Table 3.

As described in subsection 4.3 our proposed model is trained in two stages and the experimental data used in different stages are different.

Experimental Data for ECE: The experimental data used by ECE are obtained from EmpDialog and EmoCause. First, the validation set of EmoCause is randomly divided into two equal parts (denoted as EmoCause-1 and EmoCause-2). Then, the training set (unlabeled) of EmpDialog is combined with EmoCause-1 (labeled) to form the training set used in the experiments, EmoCause-2 is used as the validation set for experiments, and the test set of EmoCause is used as the test set for experiments.

Experimental Data for ECG: The experimental data used in ECG are derived from EmpDialog, and the division of the training set, validation set and test set is the same as the original dataset.

5.2 Comparison Methods

For ECE, we chose the following three models as baselines: (1) **EmpDG** (Li et al., 2020): a Transformer-based model that creates the coarse and fine-grained emotion representation by emotion classification and external emotion lexicon. In addition, it uses two discriminators to interact with user feedback. Here, we select the coarse-grained tokens as the emotion cause. (2) **RoBERTa_Att:** a RoBERTa-based (Liu et al., 2019) model that is trained on the emotion recognition task, we obtain emotion cause by the attention weight distribution of the first special token [CLS]. (3) **GEE** (Kim et al., 2021): a BART-based (Lewis et al., 2020) model that uses a Bayesian conditional probability formula based on the emotion category labels of context to predict emotion cause.

For ECG, we chose the following three models as baselines: (1) **EmpDG** (Li et al., 2020): the same as mentioned above. (2) **RecEC** (Gao et al., 2021): a Transformer-based model that incorporates emotion cause into response generation with gating mechanisms. It constructs emotion cause labels using a pre-trained sentence-level emotion cause extractor. (3) **GPT2** (Radford et al., 2019): a GPT2-based model that is fine-tuned on the conversation generation task.

5.3 Evaluation Metrics

For ECE, we conducted the automatic evaluation to evaluate with the following metrics: emotion classification accuracy (Accuracy for short) and emotion cause recall rate (Recall for short).

For ECG, we used automatic evaluation and manual evaluation to verify the effectiveness. The metrics used for the automatic evaluation included Perplexity, Distinct-1, Distinct-2, and emotion classification accuracy (Accuracy for short), well-known metrics commonly used to evaluate conversation generation. Additionally, we introduced BERTscore (Zhang et al., 2020) to measure the cosine similarity between the generated response and the gold response. BERTscore contains three more specific metrics, namely recall rate ($\mathrm{R_{BERT}}$), precision rate ($\mathrm{P_{BERT}}$) and F1 score ($\mathrm{F_{BERT}}$).

The manual evaluation included both quantitative and qualitative components. The quantitative component required scorers to score on three dimensions of Empathy, Relevance, and Fluency, with each dimension being scored in an increasing value domain from 1 to 5. The qualitative component required

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

scorers to rank the response generated by different models in order of preference. The manual evaluation randomly selected 100 test data and disrupted the responses generated by different models. Afterwards, these responses are distributed to 3 scorers for scoring, and the final results are averaged. The above approach fully ensures the fairness of the manual evaluation.

5.4 Parameter Settings

ECE is constructed based on RoBERTa-base, and ECG is constructed based on GPT2-base. Table 4 is drawn to show the parameter settings in detail.

	ECE	ECG
Initial learning rate	0.00002	0.00002
Gradient reduce strategy	ReduceLROnPlateau	ReduceLROnPlateau
Gradient clip threshold	1	1
Gradient accumulation threshold	1	2
Batch size	64	8
Early stopping strategy	Top-5 Recall	Perplexity
Early stopping threshold	5	5

Table 4: Parameter setting of ECE and ECG.

5.5 Experimental Results and Analysis

Model	Accuracy	Top-1 Recall	Top-3 Recall	Top-5 Recall
EmpDG	0.31	0.134	0.362	0.493
Roberta_Att	0.58	0.148	0.399	0.596
GEE	0.40	0.173	0.481	0.684
ECE (Ours)	**0.58**	**0.227**	**0.565**	**0.727**

Table 5: Results on comparative experiments of the different Emotion Cause Extractors.

Table 5 shows the experimental results of different emotion cause extractors. Our ECE performs optimally in all metrics compared to the comparison methods. Compared with the Roberta_Att, ECE maintains its original strong competitiveness in emotion classification accuracy while achieving remarkable improvement in emotion cause recall rate. These achievements demonstrate that our proposed semi-supervised training method can effectively narrow the gap between emotion recognition and emotion cause detection and significantly improve the emotion cause detection ability of the model.

Training Dataset	Accuracy	Top-1 Recall	Top-3 Recall	Top-5 Recall	Training Method
train	0.56	0.147	0.410	0.607	unsupervised
valid	0.56	0.246	0.514	0.556	supervised
merge (ours)	**0.58**	**0.227**	**0.565**	**0.727**	semi-supervised
merge w/o $\mathcal{L}_{aux}$	0.58	0.208	0.523	0.709	semi-supervised

Table 6: Results on ablation study of the ECE.

We design the ablation study to further analyze the effectiveness of our proposed semi-supervised training method. In Table 6, the "train" (or "valid") in Training Dataset represents that ECE uses only the training (or validation) set of EmoCause for unsupervised (or supervised) training. Similarly, "merge" represents that ECE uses the training set of EmpDialog with EmoCause-1 for semi-supervised training. Note that in the "valid" set of experiment, the test set of EmoCause is used as the validation set, which is actually not a regular practice and is only required here to meet the need of the ablation experiments because we do not have more labeled data.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

The experimental results in Tabel 6 show that the supervised training method is outstanding on Top-1 Recall and Top-3 Recall compared with the unsupervised training method. Still, the supervised training method is significantly weaker than the unsupervised training method on Top-5 Recall. This phenomenon declares that the supervised training method is superior to the unsupervised training method in performance, but it can easily cause overfitting and lead to instability. In contrast, the semi-supervised training method has the advantage of combining the two. On the one hand, supervised training can be used to provide a clear, task-appropriate optimization goal for emotion cause detection. On the other hand, the labeled data can guide the processing of automatic emotion cause labeling and the unlabeled data can avoid overfitting that may result from using only labeled data. In addition, an ablation study on $\mathcal{L}_{aux}$ under the semi-supervised training method also validates the effectiveness of our proposed auxiliary loss function.

Model	Perplexity	Distinct-1	Distinct-2	P_{BERT}	R_{BERT}	F_{BERT}	Accuracy
EmpDG	34.311	0.018	0.069	0.252	0.213	0.232	0.314
RecEC	177.825	0.019	0.090	0.225	0.177	0.201	0.412
GPT2	14.132	**0.027**	**0.112**	0.304	0.238	0.271	/
ECG (Ours)	**14.063**	0.025	0.109	**0.307**	**0.240**	**0.273**	**0.598**

Table 7: Results on Automatic Evaluation of the ECG. It should be noted that the particularly large Perplexity of RecEC is because the model is trained with F_{BERT} as the optimization target for the early stop strategy.

Table 7 demonstrates the automatic evaluation results of different empathetic conversation generation models. Our ECG achieves remarkable improvements in all metrics compared with EmpDG and RecEC, which are Transformer-based models. ECG also makes a small improvement in all metrics except Distinct compared with the pre-trained language model GPT2. The above phenomenon suggests that our ECG can improve the quality of the generated responses by introducing attention to emotion cause on the basis of pre-trained language models. Regarding the poor performance of ECG on Distinct, it may be due to the limitations caused by the emotion cause in the generation process.

Model	Empathy	Relevance	Fluency
EmpDG	2.927	2.763	4.497
RecEC	2.893	2.790	4.677
GPT2	3.213	3.257	4.753
ECG (Ours)	**3.383**	**3.553**	**4.763**

Table 8: Results on Manual Evaluation of the ECG.

Pref. (%)	EmpDG	RecEC	GPT2	**ECG (Ours)**
EmpDG	/	47.1	26.7	29.9
RecEC	52.9	/	38.2	31.0
GPT2	73.3	61.8	/	42.5
ECG (Ours)	**70.1**	**69.0**	**57.5**	/

Table 9: Preference test (%) between any two method.

Table 8 shows the manual evaluation results of different empathetic conversation generation models. The improvement in Empathy and Relevance of the responses generated by ECG is remarkable, which indicates that introducing attention to the emotion cause in the generation process can promote the model's understanding of user emotion and generate more content-relevant emotional responses. Table 9 presents the preferences of scorers for different models. The scorers' preference for our ECG is greater than the other models, which verifies the validity of the ECG.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.

5.6 Case Study

Input	I recently went to eat a restaurant that was not very clean. The whole place had an odd odor and made me uncomfortable.
Emotion	Disgusted
EmpDG	I would have cried right now.
RecEC	Oh no! I hate when people do that.
GPT2	Oh no! Did you clean it up?
ECG (Ours)	That's not good. Did you complain to the manager?
Input	My son was just born, My first boy after 6 girls !
Emotion	Joyful
EmpDG	Wow, That is amazing. How old is your son?
RecEC	Oh that's great. Is he a big boy?
GPT2	Wow! That's amazing! I bet you are proud of him!
ECG (Ours)	Congratulations ! That is a very exciting time for you and your family .

Table 10: Two cases of responses generated by different models.

To further illustrate that focusing on the emotion cause helps improve the content-relevance of the generated responses, we show two cases in Table 10. In the first case, ECE identifies the emotion cause in user input (as highlighted in yellow) and understands the stimulus behind the user's disgusted emotion is the poor environment of the restaurant, which prompts ECG to generate an empathetic response expressing sympathy and concerning for subsequent development (as highlighted in pink). In the second case, ECE recognizes the emotion cause in user input (as highlighted in yellow) and understands the stimulus behind the user's joyful emotion is the long-awaited birth of a son, prompting ECG to generate an empathetic response that congratulates to the user and fits the user's family situation (as highlighted in pink).

Comparing the responses generated by different models in the above two cases, it can be seen that our proposed model can accurately capture the emotion cause in user input and effectively incorporate it into the generation process, showing stronger content-relevance compared to other baselines, which further illustrates the important role of the emotion cause in the content-relevance of generated responses.

6 Conclusion

In this paper, we present an empathetic conversation generation model enhanced by the emotion cause to make the generated responses more content-relevant. Our proposed model comprises an emotion cause extractor and an empathetic conversation generator. To compensate for the scarcity of large-scale word-level emotion-cause labeled datasets, we suggest a semi-supervised training method that simultaneously uses labeled and unlabeled data for training. To integrate the extracted emotion cause into the generation process, we propose a biased self-attention mechanism that does not introduce new additional parameters. Experimental results indicate that our proposed model performs superior to the baselines and the generated responses of our model are more empathetic and content-relevant.

Acknowledgements

This work was supported by the Joint Project of Tianjin University-Bohai Bank Joint Laboratory for Artificial Intelligence Technology Innovation and Bayescom.

References

Yinan Bao, Qianwen Ma, Lingwei Wei, Wei Zhou, and Songlin Hu. 2022. Multi-granularity semantic aware graph model for reducing position bias in emotion cause pair extraction. In *Findings of the Association for*

Computational Linguistics: the 60th Conference of the Association for Computational Linguistics (ACL), pages 1203–1213. Association for Computational Linguistics, Dublin, Ireland.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 179–187. Tsinghua University Press, Beijing, China.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2021. More the merrier: Towards multi-emotion and intensity controllable response generation. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence (AAAI)*, pages 12821–12829. AAAI Press, Virtual Event.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 807–819. Association for Computational Linguistics, Punta Cana, Dominican Republic (Virtual Event).

Minlie Huang, Xiaoyan Zhu, and Jianfeng Gao. 2020. Challenges in building intelligent open-domain dialog systems. *ACM Transactions on Information Systems*, 38(3):1–32.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2021. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2227–2240. Association for Computational Linguistics, Punta Cana, Dominican Republic (Virtual Event).

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53. Association for Computational Linguistics, Los Angeles, USA.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 7871–7880. Association for Computational Linguistics, Online.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 4454–4466. International Committee on Computational Linguistics, Barcelona, Spain (Online).

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. In *Proceedings of the 24th Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 121–132. Association for Computational Linguistics, Hong Kong, China.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIME: mimicking emotions for empathetic response generation. In *Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8968–8979. Association for Computational Linguistics, Online.

Bilyana Martinovski and David Traum. 2003. The error is the clue: Breakdown in human-machine interaction. In *Proceedings of the International Speech Communication Association Tutorial and Research Workshop on Error Handling in Spoken Dialogue Systems*, pages 11–16. ISCA Archive, Château-d'Oex, Vaud, Switzerland.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied artificial intelligence*, 19(3-4):267–285.

Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International journal of human-computer studies*, 62(2):231–245.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

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)*, pages 5370–5381. Association for Computational Linguistics, Florence, Italy.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. CEM: Commonsense-aware empathetic response generation. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI)*, pages 11229–11237. AAAI Press, Virtual Event.

Iulian Vlad Serban, Ryan Lowe, Laurent Charlin, and Joelle Pineau. 2016. Generative deep neural networks for dialogue: A short review. *arXiv preprint arXiv:1611.06216*.

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)*, pages 556–566. Association for Computational Linguistics, Online.

Lei Shen, Jinchao Zhang, Jiao Ou, Xiaofang Zhao, and Jie Zhou. 2021. Constructing emotional consensus and utilizing unpaired data for empathetic dialogue generation. In *Findings of the Association for Computational Linguistics: the 26th Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3124–3134. Association for Computational Linguistics, Punta Cana, Dominican Republic (Virtual Event).

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2020. Generating empathetic responses by looking ahead the user's sentiment. In *Proceedings of the 45th International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 7989–7993. IEEE, Barcelona, Spain.

Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating responses with a specific emotion in dialog. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pages 3685–3695. Association for Computational Linguistics, Florence, Italy.

Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. 2021. Multi-task learning and adapted knowledge models for emotion-cause extraction. In *Findings of the Association for Computational Linguistics: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 3975–3989. Association for Computational Linguistics, Online Event.

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 *Proceedings of the 30th Conference on Annual Conference Neural Information Processing Systems (NeurIPS)*, pages 5998–6008. MIT Press, Long Beach, USA.

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)*, pages 4886–4899. International Committee on Computational Linguistics, Barcelona, Spain (Online).

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL)*, pages 1003–1012. Association for Computational Linguistics, Florence, Italy.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*. OpenReview.net, Addis Ababa, Ethiopia.

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: The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 813–824. Association for Computational Linguistics, Online.

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 32nd AAAI Conference on Artificial Intelligence (AAAI)*, pages 730–739. AAAI Press, New Orleans, USA.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2020. The design and implementation of xiaoice, an empathetic social chatbot. *Computational Linguistics*, 46(1):53–93.

Proceedings of the 21st China National Conference on Computational Linguistics, pages 811-823, Nanchang, China, October 14 - 16, 2022.