Improving Empathetic Dialogue Generation with Semantics Decoupling

Rui Pan
College of Intelligence and Computing
Tianjin University
Tianjin, China
ruipan@tju.edu.cn

Minghui Zou
College of Intelligence and Computing
Tianjin University
Tianjin, China
zzzxzmh@tju.edu.cn

Sai Zhang
College of Intelligence and Computing
Tianjin University
Tianjin, China
zhang_sai@tju.edu.cn

Yongxin Yu*
College of Intelligence and Computing
Tianjin University
Tianjin, China
yyx@tju.edu.cn

Zhiyong Feng
College of Intelligence and Computing
Tianjin University
Tianjin, China
zyfeng@tju.edu.cn

ABSTRACT

Empathetic dialogue generation is dedicated to generating responses to empathize with users by perceiving and understanding context emotions and dialogue situations. Existing works typically emphasize that an empathetic response needs to express suitable emotion through perceiving context emotion but ignore the equal need to express informative content in response by understanding the dialogue situation. To this end, we propose a novel empathetic dialogue generation model abbreviated as EmpDGM, which is extended based on the Transformer by a semantics decoupler and empathetic generator. Specifically, the semantics decoupler can effectively decouple emotion semantics and content semantics in the input sequence using adversarial training and multi-task learning meanwhile ensuring the obtained content semantics is complete. And the empathetic generator introduces a gated fusion mechanism to fuse content semantics and context emotion embedding in a balanced manner throughout the whole generation process, which overcomes generally incorporating context emotion embedding as part of initial embedding in the generation module leading the insufficient emotion expression. We conduct automatic evaluation and manual evaluation on the benchmark dataset EmpatheticDialogues of empathetic dialogue generation. Experimental results reveal that our EmpDGM outperforms advanced baselines in both emotion perceptivity and content quality and generates more informative and affective responses.

CCS CONCEPTS

• **Computing methodologies** → *Discourse, dialogue and pragmatics.*

KEYWORDS

Empathetic Dialogue Generation; Semantics Decoupling; Adversarial Training

ACM Reference Format:
Rui Pan, Minghui Zou, Sai Zhang, Yongxin Yu, and Zhiyong Feng. 2022. Improving Empathetic Dialogue Generation with Semantics Decoupling. In *Proceedings of the 11th International Joint Conference On Knowledge Graphs (IJCKG '22).* ACM, New York, NY, USA, 9 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Empathetic dialogue generation has become a prominent research area, thanks to the emergence of high-quality dialogue corpora and better learning capabilities of deep neural networks such as the Transformer [19]. Empathetic dialogue generation aims to enable the dialogue model to empathize with users by perceiving and understanding context emotions and dialogue situations to generate appropriate responses. Several pieces of research have demonstrated that the dialogue model, which generates empathetic responses, can make human-computer interaction more natural [2], improve user experience, and increase user satisfaction [5, 12, 21].

There has been a lot of attention to the research of empathetic dialogue generation over recent years and proposed various approaches to make dialogue models empathetic to a certain extent [6, 9, 10, 13, 18]. Lin et al. [10] set different decoders for different context emotions and used a shared decoder to combine hidden states of each decoder based on the predicted emotion distribution to generate an empathetic response. Li et al. [9] used coarse and fine-grained emotions simultaneously to adequately capture context emotion's nuances. Majumder et al. [13] attempted to generate an empathetic response by grouping emotions and mimicking context emotion. Shen et al. [18] argued that empathetic dialogue is a bidirectional process, and empathy occurs when the emotions of

*Corresponding author.

IJCKG '22, October 27–29, 2022, Hangzhou, China

ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00
https://doi.org/XXXXXXX.XXXXXXX

both parties to the dialogue are in consensus. Gao et al. [6] emphasized the importance of the emotion cause and generated a more empathetic response by identifying the context emotion and the cause behind it.

However, we consider the existing works have **two shortcomings** that constrain the potential of dialogue models for empathetic expression: (1) Existing works typically highlight the important contribution of context emotion for generating an empathetic response and improve the quality of a generated response by modeling context emotion to enhance emotion understanding. But research in social psychology [3, 4] argues that empathetic expression has a multidimensional meaning consisting of emotion aspect and cognition aspect, and states that the cognition aspect reflected by content expression of an empathetic response requires understanding dialogue situation. Existing works overemphasize the emotion aspect and ignore the cognition aspect, which is inadequate for empathizing. (2) Existing works generally include context emotion embedding as part of initial embedding in the generation module to understand and express emotion. However, as the number of the generation module's layers deepens, context emotion has less and less influence on generating an empathetic response.

For the shortcoming (1), we believe expressing informative (i.e., relevant and diverse) content in the generated response through better dialogue situation understanding is also important. We consider that the semantics of dialogue utterance consists of coupled emotion semantics and content semantics, and dialogue situation can usually be determined by content semantics. Thus extracting content semantics can help reduce the negative impact of potential confounding factors on understanding the dialogue situation. We attempt to decouple emotion semantics and content semantics in the input sequence by semantics decoupling and enhance the content expression of generated response by reusing content semantics. However, the following challenges exist: First, decoupling needs a reliable method to ensure that emotion semantics and content semantics in the input sequence can indeed be separated. Second, decoupling requires the establishment of a constraint to ensure that obtained content semantics is complete (i.e., preventing content semantics from being meaningless), which is essential for dialogue situation understanding. **For the shortcoming (2)**, the intuitive idea is to introduce context emotion embedding directly at each layer of the generation module. But context emotion embedding is essentially a strongly supervised signal, and frequent whole introduction tends to make the dialogue model generate a safe response for the specific emotion category. Therefore, we attempt to use a gated fusion mechanism to introduce context emotion in a balanced manner throughout generating a response.

In this paper, we propose a novel empathetic dialogue generation model, abbreviated as EmpDGM, to enable the generated empathetic response to express suitable emotion along with informative content. EmpDGM is extended based on the Transformer by a semantics decoupler and empathetic generator. Specifically, the semantics decoupler can effectively decouple emotion semantics and content semantics in the input sequence using adversarial training and multi-task learning meanwhile ensuring content semantics is complete. And the empathetic generator introduces a gated fusion mechanism that can adaptively learn a control strategy to fuse content semantics and context emotion embedding in a balanced manner during the whole generation process. Experimental results on the large-scale and commonly used benchmark dataset EmpatheticDialogues reveal that our EmpDGM outperforms advanced baselines in both automatic evaluation and manual evaluation, and the generated responses demonstrate better empathy in terms of emotion and content.

The main contributions of our work are summarized as follows:

- We use the idea of adversarial training and multi-task learning to decouple semantics of input sequence to obtain emotion semantics and content semantics meanwhile establishing a constraint to ensure content semantics is complete.
- We introduce the gated fusion mechanism, which adaptively learns a control strategy to fuse content semantics and context emotion embedding to achieve balanced participation in the generation process.
- Experimental results demonstrate that our EmpDGM benefiting from semantics decoupling and gated fusion has superior performance compared with advanced baselines, resulting in more affective and informative responses.

2 RELATED WORK

Incorporating empathy into a dialogue system can make the generated response more human and facilitate human-computer interaction. Rashkin et al. [16] contributed the benchmark dataset EmpatheticDialogues to the empathetic dialogue generation community stimulating extensive research. Lin et al. [10] set up multiple decoders corresponding to different context emotions for response generation, giving some interpretability to the generation process. Majumder et al. [13] argued that empathetic responses tend to mimic context emotion to varying degrees and proposed a new model to generate an empathetic response relying on emotion grouping and emotion mimicry. Li et al. [9] jointly considered coarse and fine-grain emotions in generating a response to fully acquire the nuances of context emotion and introduced an adversarial learning framework to use feedback to determine the degree of emotion perceptivity of the generated response in dialogue. Zheng et al. [23] modeled three factors (i.e., communication mechanism, dialog act, and emotion) influencing empathetic expression in a hierarchical manner to generate a response with better empathy. Gao et al. [6] believed that uncovering the emotion cause helps to understand the emotion better, so they proposed a framework to empower the empathetic dialogue model to identify the context emotion and the cause behind it to make more appropriate empathy. Shen et al. [18] considered empathetic dialogue a bidirectional process and integrated two-direction dialogue models with a discrete latent variable representing emotion consensus in a unified architecture to generate a better empathetic response. Sabour et al. [17] argued that the introduction of commonsense in empathetic response generation plays an important role in understanding context emotion and dialogue situation and can improve the empathetic ability of the generated response.

In fact, empathy encompasses emotion and cognition, and appropriate empathy requires perceiving, understanding, and responding to the context emotion and dialogue situation. However, existing work usually focuses on only one of these aspects.

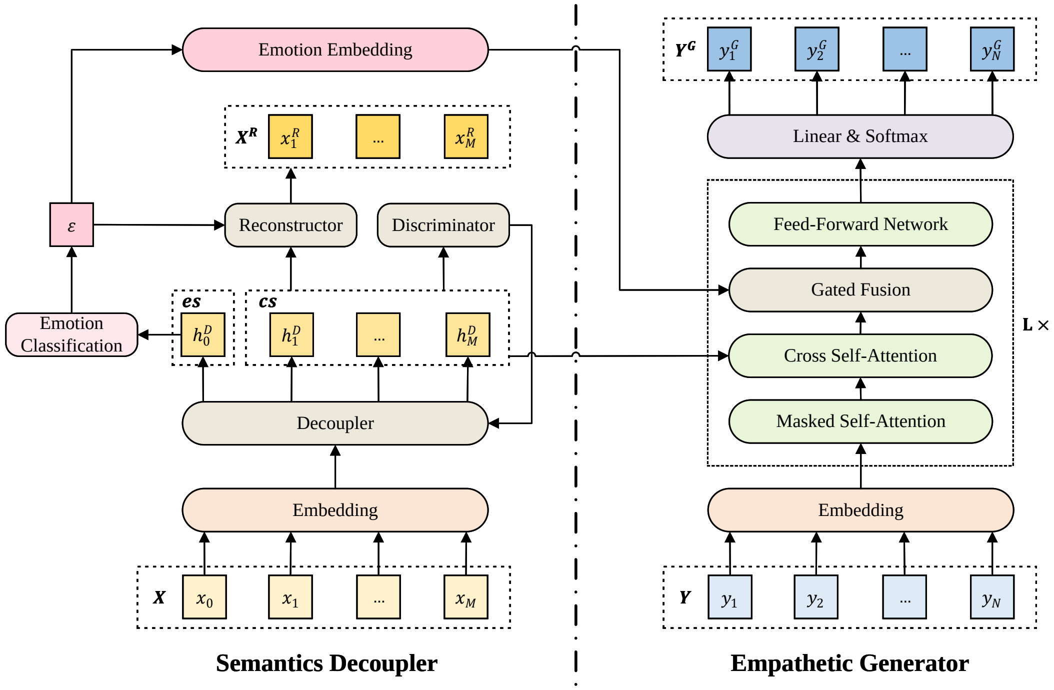

Figure 1: The overview of our EmpDGM, which contains two modules: (1) Semantics decoupler separates content semantics cs and emotion semantics es in the input sequence X using adversarial training and multi-task learning. (2) Empathetic generator introduces the gated fusion mechanism that fuses content semantics cs and context emotion ε's embedding E_ε to generate a more empathetic response Y^G.

3 TASK FORMULATION

We give the task formulate of empathetic dialogue generation as follows in the first place before going into more detail about our method.

In the multi-turn setting, the dialogue context $U = \{u_1, u_2, ..., u_K\}$ contains K utterances from the speaker and listener, and each utterance $u_i = \left\{t_1^i, t_2^i, ..., t_{T_i}^i\right\}$ contains T_i tokens. Following the previous works [6, 9, 10], we concatenate the K utterances together and flat into a token sequence. Furthermore, we separate utterances using special tokens [SEP] and insert a special token [CLS] at the start of the token sequence. Finally, we get the input sequence $X = \{x_0, x_1, ..., x_M\}$, where x_0 corresponds to the special token [CLS].

Empathetic dialogue generation requires the model to act as a listener. Formally, given an input sequence X, our model aims to generate a N-length empathetic response $Y = \{y_1, y_2, ..., y_N\}$ suitable for emotion and informative in content by maximizing probability $P(Y|X) = \prod_{n=1}^{N} P(y_n|y_{<n}, X)$.

4 METHODS

In this section, we go into depth about our EmpDGM that introduces semantics decoupling and gated fusion for empathetic response generation. Figure 1 illustrates the overview of EmpDGM, which contains two modules: semantics decoupler and empathetic generator. The semantics decoupler decouples emotion semantics and content semantics in the input sequence using adversarial training and multi-task learning. The empathetic generator introduces the gated fusion mechanism that fuses content semantics and context emotion embedding to generate a more empathetic response. We give more details about the two modules in the following subsections.

4.1 Semantics Decoupler

We propose a semantics decoupler to decouple emotion semantics and content semantics in the input sequence while ensuring that content semantics does not lose completeness. We use the idea of adversarial training and multi-task learning to implement this module. Specifically, adversarial training comprises a decoupler and a discriminator. Multi-task learning means that we not only decouple semantics (including context emotion classification and adversarial training) but also implement semantics reconstruction.

The decoupler (denoted as D) is designed to obtain emotion semantics and content semantics according to the input sequence and trick the discriminator into misclassifying the emotion of content semantics. Specifically, given an input sequence X, each token x_i's embedding E_{x_i} is firstly represented as the sum of its corresponding word embedding $E_{x_i}^W$, position embedding $E_{x_i}^P$, and dialogue state embedding $E_{x_i}^S$ (for distinguishing utterances from speaker or listener in multi-turn dialogue):

$$E_{x_i} = E_{x_i}^W + E_{x_i}^P + E_{x_i}^S \quad (1)$$

Then, we feed the embedding sequence into the Transformer encoder to obtain context representation for each token. Note that the Transformer encoder stacks L identical layers so that each position i of the sequence in the l-th ($l \in \{1, 2, ..., L\}$) layer is calculated as follows:

$$a_i^l = \begin{cases} \text{LayerNorm}\left(E_{x_i} + \text{MSAttn}\left(E_{x_i}\right)\right), & l = 1 \\ \text{LayerNorm}\left(h_i^{l-1} + \text{MSAttn}\left(h_i^{l-1}\right)\right), & 1 < l \le L \end{cases} \tag{2}$$

$$h_i^l = \text{LayerNorm}\left(a_i^l + \text{FFN}\left(a_i^l\right)\right) \tag{3}$$

where MSAttn $(\cdot)$ and FFN $(\cdot)$ denote the multi-head self-attention sub-layer and feed-forward network sub-layer proposed in [19], LayerNorm $(\cdot)$ denotes the layer normalization proposed in [1].

Finally, we collect all outputs of the last layer to obtain the context representation sequence $h^D = \left\{h_0^D, h_1^D, ..., h_m^D\right\}$.

The emotion semantics es represented as h_0^D can be learned via an emotion classification task aiming to predict context emotion ε based on h_0^D as follows:

$$\varphi = \text{Softmax}\left(\boldsymbol{W_D} h_0^D + b_D\right) \tag{4}$$

$$\varepsilon = \arg\max\left(\varphi\right) \tag{5}$$

where $\boldsymbol{W_D}$ and b_D are trainable parameters. During training, the emotion semantics es is optimized by minimizing cross-entropy between ground truth distribution $t\left(\cdot\right)$ and predicted emotion distribution with the following loss function:

$$\mathcal{L}_{emo}\left(\Theta_D\right) = -\sum_{i \in labels} t\left(i\right) \log\left(\varphi_i\right) \tag{6}$$

where $labels$ denotes the index of emotion categories, Θ_D denotes parameters of the decoupler.

Furthermore, we treat the sequence consisting of context representations starting from h_1^D as the content semantics cs, i.e., $cs = \left\{h_1^D, h_2^D, ..., h_m^D\right\}$. The content semantics cs aims to make the discriminator unable to identify its emotion correctly by maximizing the predicted emotion distribution's entropy (i.e., minimizing the negative entropy) with the following loss function:

$$\mathcal{L}_{adv}\left(\Theta_D\right) = -\mathcal{H}\left(e^C | cs, \Theta_C\right) \tag{7}$$

where $\mathcal{H}\left(p\right) = -\Sigma_{i \in labels} p_i \log p_i$, e^C denotes emotion distribution of the content semantics cs predicted by the discriminator, Θ_C denotes parameters of the discriminator.

The discriminator (denoted as C) is essentially an emotion classifier designed to correctly classify the emotion of the content semantics cs. Specifically, we add the special token [CLS]'s embedding E_{CLS} in front of the content semantics cs and then feed it into another Transformer encoder with a different parameter set. We use the context representation h_0^C corresponding to the [CLS] token to predict the emotion distribution of the content semantics cs:

$$h^C = \text{Transformer}_{\text{encoder}}^{\text{C}}\left(\left[E_{\text{CLS}}; cs\right]\right) \tag{8}$$

$$e^C = \text{Softmax}\left(\boldsymbol{W_C} h_0^C + b_C\right) \tag{9}$$

where $[\cdot;\cdot]$ denotes concatenation, $\boldsymbol{W_C}$ and b_C are trainable parameters. Note that the calculation procedure of $\text{Transformer}_{\text{encoder}}^{\text{C}}\left(\cdot\right)$ is the same as that described in Equation (2)-(3).

The discriminator is trained with the following loss function to minimize cross-entropy between ground truth distribution $t\left(\cdot\right)$ and predicted emotion distribution:

$$\mathcal{L}_{dis}\left(\Theta_C\right) = -\sum_{i \in labels} t\left(i\right) \log\left(e_i^C\right) \tag{10}$$

To prevent the decoupler from generating content semantics that does not contain emotion but loses completeness, we introduce a reconstructor (denoted as R) into the semantics decoupler. Specifically, the reconstructor is implemented by a Transformer encoder, which reconstructs the input sequence based on the content semantics cs and the context emotion embedding E_ε:

$$h^R = \text{Transformer}_{\text{encoder}}^{\text{R}}\left(cs \oplus E_\varepsilon\right) \tag{11}$$

$$x_i^R = \text{Softmax}\left(\boldsymbol{W_R} h_i^R + b_R\right) \tag{12}$$

where $\oplus$ denotes element-wise addition, $\boldsymbol{W_R}$ and b_R are trainable parameters.

The reconstructor requires minimizing cross-entropy loss among ground truth distribution x_i and reconstructed token distribution x_i^R using the following loss function:

$$\mathcal{L}_{rec}\left(\Theta_D, \Theta_R\right) = -\frac{1}{M}\sum_{i=1}^{M} x_i \log\left(x_i^R\right) \tag{13}$$

where M denotes the reconstructed input sequence's length, Θ_R denotes parameters of the reconstructor.

4.2 Empathetic Generator

After obtaining the content semantics cs from the semantics decoupler, we design an empathetic generator (denoted as G) to generate an empathetic response $Y = \left\{y_1^G, y_2^G, ..., y_N^G\right\}$. The empathetic generator is based on a Transformer decoder, which also stacks L identical layers much like the Transformer encoder. But each layer contains a masked multi-head self-attention sub-layer, a cross multi-head self-attention sub-layer, and a feed-forward network sub-layer. Next, we detail how to incorporate the content semantics cs and context emotion embedding E_ε into the empathetic generator to participate in response generation.

During training, the embedding E_{y_i} of each token y_i of gold response is obtained by summing its corresponding word embedding $E_{y_i}^W$, position embedding $E_{y_i}^P$ and dialogue state embedding $E_{y_i}^S$:

$$E_{y_i} = E_{y_i}^W + E_{y_i}^P + E_{y_i}^S \tag{14}$$

Then, the embedding sequence is fed into the masked multi-head self-attention sub-layer (denoted as MMSAttn $(\cdot)$), which calculates the self-attention of the current position by masking part after the current position to satisfy the autoregressive property. Specifically, each position i in the masked multi-head self-attention sub-layer of the l-th ($l \in \{1, 2, ..., L\}$) layer is calculated as follows:

$$m_i^l = \begin{cases} \text{LayerNorm}\left(E_{y_i} + \text{MMSAttn}\left(E_{y_i}\right)\right), & l = 1 \\ \text{LayerNorm}\left(f_i^{l-1} + \text{MMSAttn}\left(f_i^{l-1}\right)\right), & 1 < l \le L \end{cases} \tag{15}$$

Next, the masked multi-head self-attention representation m_i^l is fed into the cross multi-head self-attention sub-layer (denoted as CMSAttn $(\cdot)$), which incorporates the content semantics cs to guide content expression of the generated response. Formally, each position i in the cross multi-head self-attention sub-layer is calculated as follows:

$$c_i^l = \mathrm{CMSAttn}\left(m_i^l, cs\right) \tag{16}$$

As mentioned earlier, existing work generally involves context emotion embedding as part of initial embedding in the generation module to help understand and express emotion. However, as the number of the generation module's layers deepens, context emotion has less and less influence on expressing emotion in the generated response.

To this end, we introduce a gated fusion sub-layer (denoted as $\mathrm{GF}(\cdot)$) that enables both informative content expression and stable emotion expression in the empathetic response by adaptively fusing the cross multi-head self-attention representation c_i^l and context emotion embedding E_ε using adjustable weights at the current position. Specifically, c_i^l and E_ε are first mapped to the same underlying space. Then, the *sigmoid* function (σ for short) calculates the adaptive adjustment weights ρ. Finally, the weighted summation is performed. Formally, each position i in the gated fusion sub-layer is calculated as follows:

$$\tilde{c}_i^l = \tanh\left(\boldsymbol{W_c} c_i^l + b_c\right) \tag{17}$$

$$\tilde{E}_\varepsilon = \tanh\left(\boldsymbol{W_\varepsilon} E_\varepsilon + b_\varepsilon\right) \tag{18}$$

$$\rho = \sigma\left(\boldsymbol{W_\rho}\left[\tilde{c}_i^l; \tilde{E}_\varepsilon\right]\right) \tag{19}$$

$$\mathrm{GF}\left(c_i^l, E_\varepsilon\right) = \rho \otimes \tilde{c}_i^l + (1-\rho) \otimes \tilde{E}_\varepsilon \tag{20}$$

where $\otimes$ denotes element-wise multiplication.

After that, we use residual connections around the gated fusion sub-layer and perform layer normalization as follows:

$$g_i^l = \mathrm{LayerNorm}\left(m_i^l + \mathrm{GF}\left(c_i^l, E_\varepsilon\right)\right) \tag{21}$$

Later, we pass the fusion representation g_i^l into the feed-forward network sub-layer (denoted as $\mathrm{FFN}(\cdot)$) as follows:

$$f_i^l = \mathrm{LayerNorm}\left(g_i^l + \mathrm{FFN}\left(g_i^l\right)\right) \tag{22}$$

After calculating L layers, we get the final representation h^G. The probability distribution used to predict token in the response sequence is calculated as follows:

$$y_i^G = \mathrm{Softmax}\left(\boldsymbol{W_G} h_i^G + b_G\right) \tag{23}$$

Following is the cross-entropy loss function used to train the empathic generator:

$$\mathcal{L}_{dec}\left(\Theta_D, \Theta_G\right) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(y_i^G\right) \tag{24}$$

where N denotes the response sequence's length, Θ_G denotes parameters of the empathetic generator.

4.3 Training Strategy

The generation loss function $\mathcal{L}_{gen}\left(\Theta_D, \Theta_R, \Theta_G\right)$ of EmpDGM includes context emotion classification loss, empathetic generator loss, decoupler loss, and reconstructor loss, defined as follows:

$$\begin{aligned}\mathcal{L}_{gen}\left(\Theta_D, \Theta_R, \Theta_G\right) =& \mathcal{L}_{emo}\left(\Theta_D\right) + \mathcal{L}_{dec}\left(\Theta_D, \Theta_G\right) \\ &+ \lambda\mathcal{L}_{adv}\left(\Theta_D\right) + \eta\mathcal{L}_{rec}\left(\Theta_D, \Theta_R\right)\end{aligned} \tag{25}$$

where λ and η are hyperparameters.

In summary, EmpDGM alternately optimizes the generation loss $\mathcal{L}_{gen}\left(\Theta_D, \Theta_R, \Theta_G\right)$ and discriminator loss $\mathcal{L}_{dis}\left(\Theta_C\right)$ at training time. The training strategy is described as shown in Algorithm 1.

Algorithm 1: Training Strategy

```
  Input: Given an input sequence $X = \{x_0, x_1, ..., x_M\}$.
  Output: An empathetic response $Y = \{y_1, y_2, ..., y_N\}$.
1 foreach mini-batch do
2     if step in G-steps then
3         minimize $\mathcal{L}_{gen}(\Theta_D, \Theta_R, \Theta_G)$, Optimized decoupler,
            reconstructor, and generator.
4     end
5     if step in D-steps then
6         minimize $\mathcal{L}_{dis}(\Theta_C)$, Optimized discriminator.
7     end
8 end
```

Note that the notations *step* and *G-steps* in row 2 denote the current training step and training frequency of the corresponding module, respectively. Therefore, the judgment condition in row 2 (i.e., "*step* in *G-steps*") indicates whether the current training step requires training the corresponding module. The notations in row 5 are similar.

5 EXPERIMENTS

5.1 Dataset

We conduct experiments to evaluate the effectiveness of our EmpDGM using the benchmark dataset EmpatheticDialogues proposed by Rashkin et al. [16], which is a large-scale multi-turn dataset collected on the Amazon Mechanical Turk platform.

The EmpatheticDialogues contains 24,850 multi-turn dialogues and defines 32 fine-grain emotions. Each dialogue is created based on an emotion and a situation. Specifically, two paired crowd-sourced workers are assigned different roles, i.e., Speaker and Listener. Speaker selects an emotion and describes the situation based on that emotion. After that, Speaker and Listener conduct multi-turn dialogues based on that situation description. Rashkin et al. constrained the Speaker's emotion selection during dialogue collection to make the distribution of emotion in the dataset more uniform.

Finally, we divide the EmpatheticDialogues in the ratio of 19,533: 2770: 2547 to obtain the training, validation, and test sets.

5.2 Implementation Details

We use PyTorch[1] to implement our EmpDGM and use pre-trained GloVE vectors [15] to initialize the word embeddings, which are shared for all corresponding modules. The hidden dimension for all modules is set to 300. We adopt Adam [7] to optimize the model, and

[1] https://pytorch.org/

the batch size and learning rate are set to 16 and 0.0005, respectively. Hyperparameters λ and η in $\mathcal{L}_{gen}$ $(\Theta_D, \Theta_R, \Theta_G)$ are set to 1. During training, the *G-steps* is set to 1, and *D-steps* is set to 5, which means training every step and once every five steps, respectively. Meanwhile, we use the early stop strategy to stop training when the loss of the validation set is no longer decreasing. The threshold of the early stop strategy is 5. For fairness, our EmpDGM and baselines use the greedy search strategy in the inference phase.

5.3 Baselines

We selected the following representative baselines to compare with our proposed EmpDGM to evaluate its effectiveness:

Transformer [19] is established based on an encoder-decoder architecture. The encoder and decoder are both formed by stacking several identical layers containing mainly the multi-head self-attention sub-layer and feed-forward network sub-layer.

MIME [13] is a Transformer-based model proposed in 2020 that argues that grouping emotions based on polarity and mimicking context emotion can improve the quality of the generated empathetic response to some extent. In addition, it introduces a stochastic strategy during training to make the response more diverse.

RecEC [6] is a Transformer-based model proposed in 2021 that identifies sequence-level context emotion and the word-level emotion cause behind it in empathetic dialogue generation for the first time and later incorporates the emotion cause into the response generation process.

CEM [17] is a Transformer-based model proposed in 2022 that uses commonsense to obtain additional information to understand the context emotion and dialogue situation better and incorporates this information into the generation process to enhance the empathetic expression of responses.

5.4 Automatic Evaluation

5.4.1 Metrics.

For automatic evaluation, we evaluate the effectiveness of our EmpDGM in two dimensions: content level and emotion level.

At the content level, BELU [14], originally applied to machine translation tasks, has proved unsuitable for evaluating the quality of generated responses in the dialogue generation task [11]. Therefore, we select the following metrics from the content level in the automatic evaluation. (1) **Perplexity:** Perplexity [20] is a common metric for evaluating the fluency of generated responses. A smaller Perplexity indicates a higher probability of response generation and better fluency. (2) **Distinct-n:** Distinct-n [8] is often used for evaluating the diversity of generated responses. A higher Distinct-n indicates a better diversity of responses. Here we use Distinct-1 and Distinct-2 in our experiments. (3) **BERTScore:** We follow the work [6] and use the BERTscore [22] as one metric to evaluate the generation quality. BERTscore uses the pre-trained language model to encode the generated and gold responses separately and then calculates the sum of weighted cosine similarities between embeddings of their tokens. BERTscore includes three more specific metrics: precision score ($\mathrm{P_{BERT}}$), recall score ($\mathrm{R_{BERT}}$), and F1 score ($\mathrm{F_{BERT}}$). At the emotion level, following the previous works [6, 17, 18], we adopt **Emotion Accuracy** to evaluate the agreement between predicted context emotion and ground truth.

5.4.2 Results and Analysis.

Table 1 demonstrates the automatic evaluation results of different models in terms of content level and emotion level. In the overall view, our EmpDGM achieves the best performance on all metrics except the metric $\mathrm{R_{BERT}}$.

Specifically, for the metric Perplexity, EmpDGM achieves the lowest score, which indicates that the responses generated by EmpDGM are more fluent than baselines. In addition, our EmpDGM also significantly outperforms the baselines on the metrics Distinct-1 and Distinct-2, implying that EmpDGM can better understand diverse dialogue situations and thus generate more informative responses. For the metric BERTScore, EmpDGM roughly achieves the best results, indicating a higher similarity between responses generated by EmpDGM and gold responses, which usually implies a higher generation quality. Regarding the metric Accuracy, our EmpDGM achieves a similar result to RecEC while slightly outperforming CEM, which is surprising because we do not make an explicit extra effort in perceiving and understanding context emotion. As discussed in the ablation analysis, we argue that this may be due to the positive connection between the semantics decoupling and context emotion classification.

Note that Transformer has no result on the metric Accuracy because Transformer generates responses based on context representations of the input sequence without utilizing context emotion. In addition, RecEC has a huge Perplexity because we train the model using the metric $\mathrm{F_{BERT}}$ as the target for model selection, following the authors' settings, to reproduce the results reported by the authors as much as possible. In contrast, other baselines and our EmpDGM use the metric Perplexity as the target for model selection.

5.5 Manual Evaluation

5.5.1 Settings and Metrics.

Manual evaluation is more convincing and essential for conversation generation [11]. Therefore, we conduct the widely used manual evaluation to assess the quality of generated responses. Specifically, we randomly select 100 test data to feed into our EmpDGM and baselines separately to generate empathetic responses. After that, the responses generated by each model are disrupted and assigned to three annotators for scoring. The disrupted approach ensures the fairness of scoring.

Annotators use three metrics **Empathy**, **Relevance** and **Fluency** to score responses generated by all models. Specifically, Empathy measures the understanding and expression degree of context emotion the response demonstrates. Relevance assesses the relevance of the generated response to dialogue context. Fluency evaluates the grammatical correctness and readability of the generated response. Each metric is scored on a Likert scale from 1 to 5, with "5" representing the best. Finally, the scores of three annotators are averaged to obtain a final score for each model.

We also perform the **human A/B test** on the responses generated by our EmpDGM and baselines from the perspective of user propensity. We combine two responses generated by models A and B corresponding to the same dialogue context, where A is our EmpDGM and B is a baseline. These combinations are then randomly assigned to three annotators who are asked to choose a more satisfactory response. If the annotator has difficulty deciding between two responses, the annotator can choose "`Tie`".

Table 1: Results on automatic evaluation of our EmpDGM and baselines in terms of content level (i.e., Perplexity, Distinct-1, Distinct-2, $\mathrm{P}_{\mathrm{BERT}}$, $\mathrm{R}_{\mathrm{BERT}}$**, and** $\mathrm{F}_{\mathrm{BERT}}$**) and emotion level (i.e.,** Emotion Accuracy**). Note that the metric** Emotion Accuracy **is abbreviated as** Accuracy**, and the best result for all models on each metric is highlighted in bold.**

Models	Perplexity	Distinct-1	Distinct-2	$\mathrm{P}_{\mathrm{BERT}}$	$\mathrm{R}_{\mathrm{BERT}}$	$\mathrm{F}_{\mathrm{BERT}}$	Accuracy (%)
Transformer	37.61	0.44	1.99	0.281	0.201	0.240	-
MIME	37.26	0.44	1.76	0.274	0.204	0.238	30.58
RecEC	150.90	0.70	2.87	0.286	**0.227**	0.256	39.41
CEM	36.54	0.63	2.76	0.291	0.208	0.249	37.74
EmpDGM	**34.18**	**0.90**	**3.46**	**0.307**	0.222	**0.264**	**39.42**

Table 2: Results on manual evaluation of our EmpDGM and baselines around the three metrics Empathy, Relevance**, and** Fluency**. The best result for all models on each metric is highlighted in bold.**

Models	Empathy	Relevance	Fluency
Transformer	3.017	2.977	4.333
MIME	3.137	3.083	4.467
RecEC	3.253	3.203	4.583
CEM	3.227	3.237	4.577
EmpDGM	**3.723**	**3.533**	**4.603**

Table 3: Results of human A/B test.

Models	Win (%)	Loss (%)	Tie (%)
EmpDGM vs. Transformer	49.7	29.0	21.3
EmpDGM vs. MIME	51.3	26.0	22.7
EmpDGM vs. RecEC	44.7	31.7	23.7
EmpDGM vs. CEM	42.3	34.7	23.0

5.5.2 Results and Analysis. Table 2 shows the results of the manual evaluation around the three metrics Empathy, Relevance, and Fluency. In the overall view, Our EmpDGM achieves the highest scores in all metrics and indicates that responses generated by EmpDGM can express not only suitable emotion but also informative content.

Specifically, for the metric Empathy, our EmpDGM considerably outperforms the baselines, highlighting the effectiveness of the gated fusion mechanism. Indeed, the gated fusion mechanism can adaptively incorporate context emotion embedding into the whole generation process, thus allowing context emotion to function consistently and stably under a multi-layer network structure. For the metric Relevance, Our EmpDGM is also substantially ahead of all baselines suggesting that extracting the input sequence's content semantics by semantics decoupling allows the model to understand better the dialogue situation, which is beneficial in improving the content expression of generated responses. In terms of metric Fluency, it can be observed that the gap between all models is less pronounced. It can be explained by the fact that all models are based on the Transformer structure, which has excellent generative capability by learning informative context representations through the multi-head self-attention mechanism and thus generates responses with better grammaticality and readability.

Table 3 presents the results of the human A/B test. We notice that responses generated by our EmpDGM are preferred compared to other baselines, suggesting that EmpDGM has superior empathy through better perceiving the context emotion and understanding the dialogue situation.

5.6 Ablation Analysis

We perform ablation studies to better understand each module's contribution to our proposed model. Specifically, we design two variants of EmpDGM: **(1) w/o SD:** the discriminator and reconstructor are removed, and the corresponding loss functions $\mathcal{L}_{adv}$ (Θ_D), $\mathcal{L}_{dis}$ (Θ_C) and $\mathcal{L}_{rec}$ (Θ_D, Θ_R) are also removed from the training objectives (i.e., Equation (7), (10), and (13)). Thus the semantics decoupler degenerates into a normal Transformer encoder. **(2) w/o GF:** the gated fusion mechanism is removed, and the empathetic generator includes context emotion embedding as part of initial embedding in this module, i.e., the embedding E_{y_i} in Equation (14) will add context emotion embedding E_ε additionally. The results of the ablation analysis of our proposed EmpDGM are shown in Table 4.

When we remove the semantics decoupler (i.e., w/o SD), all metrics except the metric $\mathrm{R}_{\mathrm{BERT}}$ perform worse. Especially, the metrics Perplexity, Distinct, and Accuracy performance become obviously worse, indicating that the semantics decoupler decoupling emotion semantics and content semantics helps better understand context emotion and dialogue situation. In general, the dialogue situation can usually be determined by content semantics. The content semantics contains fewer potential negative factors, allowing EmpDGM to understand the dialogue situation better. As for the improvement of EmpDGM in emotion accuracy may be due to some connection between the semantics decoupling and context emotion classification that can be mutually reinforcing. Intuitively, to better decouple emotion semantics and content semantics, EmpDGM naturally needs to understand context emotion better.

When we remove the gating fusion mechanism (i.e., w/o GF), the performance of all metrics deteriorates, suggesting that the gated fusion mechanism can better integrate context emotion into the generation process than as part of initial embedding in the empathetic generator, thus allowing the model to generate a more empathetic response.

5.7 Case Study

We present two cases from the generated responses of our EmpDGM and the baselines in Table 5. In the first case, although all

Table 4: Results on ablation analysis of our proposed EmpDGM. Note that w/o SD denotes EmpDGM without the discriminator and reconstructor, and w/o GF denotes EmpDGM without the gated fusion mechanism. The best result is highlighted in bold.

Models	Perplexity	Distinct-1	Distinct-2	P_{BERT}	R_{BERT}	F_{BERT}	Accuracy (%)
EmpDGM	**34.18**	**0.90**	**3.46**	**0.307**	0.222	**0.264**	**39.42**
w/o SD	35.90	0.71	2.68	0.297	**0.228**	0.262	37.48
w/o GF	34.38	0.79	3.18	0.285	0.219	0.252	38.62

Table 5: Two cases of the generated responses from EmpDGM and baselines.

Emotion	Proud
Context	My wife called me this morning and told me she scored a new job with Microsoft!
Transformer	That is a good thing to do. I hope you have a good time.
MIME	That is a good thing to do. I am sure she will be able to get it.
RecEC	That's awesome. I hope she continues to get that.
CEM	That is great! Congratulations!
EmpDGM	That's awesome! What kind of job is it?
Gold	That's amazing! She must be excited.
Emotion	Caring
Context	Speaker: A year ago, I found an injured kitten in the road while driving. Listener: Oh no! Did you take the kitten to an animal hospital or care for it yourself? Speaker: I cared for it myself. Now she is my beloved cat, Muffin.
Transformer	I am glad you are not a good person.
MIME	I am sure you will be fine.
RecEC	That's great! I hope she isn't too close.
CEM	That is great, I love cats.
EmpDGM	That's so sweet of you. I'm sure she is a great friend.
Gold	Awe, That's a sweet story! The kitten got a happy ending!

responses express a suitable emotion, EmpDGM and the baselines in situation understanding have significant differences. Specifically, Transformer focuses on the wrong core person of the dialogue (i.e., not *user* but *user's wife*). MIME misunderstands the subtle difference between *having gotten a job* and *starting to look for a job*. RecEC may understand *getting a job* as *getting a promotion*, but this is acceptable. CEM only conveys the emotion of congratulations to the user without further development around dialogue situation, which is disadvantageous to the continuation of human-computer communication. In contrast, EmpDGM has a better understanding of the dialogue situation and responds with an informative sentence (i.e., "*What kind of job is it?*"). The second case shows the performance of different models in a multi-turn scenario. As can be seen, EmpDGM successfully portrays the user as a kindhearted person (i.e., "*That's so sweet of you.*") while responding positively (i.e., "*I'm sure she is a great friend.*") to the rapport between them (i.e., "*my beloved cat*").

6 CONCLUSION AND FUTURE WORK

In this paper, we propose a novel empathetic dialogue generation model named EmpDGM, which demonstrates the effectiveness of semantics decoupling for understanding the dialogue situation and the stimulative effect of the gated fusion mechanism on understanding and expressing context emotion. Extensive experimental results show that our EmpDGM can generate more empathetic responses, which are affective and informative.

Using content semantics to enhance dialogue situation understanding is a coarse-grained approach at the sentence level. In future work, we suggest that some dialogue situation-related fine-grained factors can be introduced to explicitly and purposefully facilitate the model's understanding of the dialogue situation.

REFERENCES

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. arXiv:1607.06450

[2] John S. Callender. 2015. Being amoral: Psychopathy and moral incapacity. *British Journal of Psychiatry* 207, 3 (2015), 274–275.

[3] Mark H Davis. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology* 44, 1 (1983), 113–126.

[4] Robert Elliott, Arthur C Bohart, Jeanne C Watson, and David Murphy. 2019. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy* 55, 4 (2019), 399.

[5] Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. *JMIR Ment Health* 4, 2 (2017), e19.

[6] 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 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 807–819.

[7] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR'15)*. OpenReview.net, San Diego, USA.

[8] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 15th Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT'16)*. Association for Computational Linguistics, San Diego, USA, 110–119.

[9] 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'2020)*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 4454–4466.

[10] 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'19)*. Association for Computational Linguistics, Hong Kong, China, 121–132.

[11] 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'16)*. Association for Computational Linguistics, Austin, USA, 2122–2132.

[12] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP'21)*. Association for Computational Linguistics, Virtual Event, 3469–3483.

[13] 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 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. Association for Computational Linguistics, Online, 8968–8979.

[14] 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 (ACL'02)*. Association for Computational Linguistics, Philadelphia, USA, 311–318.

[15] 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'14)*. Association for Computational Linguistics, Doha, Qatar, 1532–1543.

[16] 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 Annual Meeting of the Association for Computational Linguistics (ACL'19)*. Association for Computational Linguistics, Florence, Italy, 5370–5381.

[17] 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'22)*. AAAI Press, Virtual Conference, 11229–11237.

[18] 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 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP'21)*. Association for Computational Linguistics, Punta Cana, Dominican Republic, 3124–3134.

[19] 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 31st Conference on Neural Information Processing Systems (NeurIPS'17)*. Curran Associates, Inc., Long Beach, USA, 5998–6008.

[20] Oriol Vinyals and Quoc Le. 2015. A neural conversational model. arXiv:1506.05869

[21] Liuping Wang, Dakuo Wang, Feng Tian, Zhenhui Peng, Xiangmin Fan, Zhan Zhang, Mo Yu, Xiaojuan Ma, and Hongan Wang. 2021. CASS: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–31.

[22] 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'20)*. OpenReview.net, Addis Ababa, Ethiopia.

[23] 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'21)*. Association for Computational Linguistics, Virtual Event, 813–824.