# MuSE: A Multi-scale Emotional Flow Graph Model for Empathetic Dialogue Generation

Deji Zhao[1], Donghong Han[1(✉)], Ye Yuan[2], Chao Wang[3], and Shuangyong Song[3]

[1] School of Computer Science and Engineering, Northeastern University, Shenyang, China
zhaodeji@stumail.neu.edu.cn, handonghong@cse.neu.edu.cn
[2] School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China
yuan-ye@bit.edu.cn
[3] Department of Big Data and AI, China Telecom, Beijing, China
{wangc17,songshy}@chinatelecom.cn

**Abstract.** The purpose of empathetic dialogue generation is to fully understand the speakers' emotional needs in dialogues and to generate appropriate empathetic responses. Existing works mainly focus on the overall coarse-grained emotion of the context while neglecting different utterances' fine-grained emotions, which leads to the inability to detect the speakers' fine-grained emotional changes during a conversation. However, in real-life dialogue scenarios, the speaker usually carries an initial emotional state that changes continuously during the conversation. Therefore, understanding a series of emotional states can help to better understand speakers' emotions and generate empathetic responses. To address this issue, we propose a **Mu**lti-**S**cale **E**motional flow model called **MuSE**, which simulates speakers' emotional flow. First, we introduce a fine-grained expansion strategy to transform context into an emotional flow graph that combines multi-scale coarse and fine-grained information. This emotional flow graph captures speakers' constant emotional changes at each turn of a conversation. And then, the emotion node and the situational node are introduced to the emotional flow graph respectively in order to extend the speakers' initial emotion into the ensuing conversation. Finally, we conduct experiments on the public EMPATHETIC DIALOGUES dataset. The experimental results demonstrate that the MuSE model achieves superior performance under both automatic evaluation and human evaluation metrics compared with the existing baseline models. Our code is available at https://github.com/DericZhao/MuSE.

**Keywords:** Empathetic Dialogue · Multi-scale · Emotional Flow · Dialogue Graph · Dialogue Generation


D. Koutra et al. (Eds.): ECML PKDD 2023, LNAI 14170, pp. 491–507, 2023.
https://doi.org/10.1007/978-3-031-43415-0_29

## 1 Introduction

In recent years, researchers have been increasingly interested in promoting more meaningful human-computer interactions in open-domain dialogue systems, such as the empathic dialogue system. The core of empathetic dialogue is to understand the speakers' emotional needs and to generate appropriate responses from their situation [5,13]. Empathetic dialogue can be leveraged for mental healthcare, emotional companionship, psychological counseling and other fields.

In order to improve the capability of empathic dialogue systems, existing works focus on recognizing emotion in the context and generating empathetic responses accordingly. Most existing approaches recognize contextual emotions from two directions. One method directly detects emotion through context [2,6,9,11,19], while the other method adopts external knowledge to indirectly understand emotional needs by identifying intent and emotional cause [20,23,25,26]. However, previous works tend to consider the conversation context as a whole coarse-grained emotion, without taking the subtle emotional changes of the speaker during each turn of the conversation into account. Li et al. [9] propose the EmpDG model that emphasizes the modeling of emotions during the conversation, but they don't incorporate speakers' emotional changes during the conversation. Although Wang et al. [27] devise the SEEK model to capture emotion dynamics, they still focus on the utterance level, ignoring speakers may say more sentences with different emotional states in a utterance.

We believe that the emotional changes generated between speakers are the essential difference between empathetic dialogues and ordinary multi-turn dialogues. As all the contexts of empathetic dialogue revolve around the changes in emotional flow, different speakers influence each other to different degrees through various emotional intensities. In the Emotion Recognition in Conversation (ERC) task [29], each turn of the conversation is characterized by different emotional states, which inspired us to introduce the concept of speaker emotional flow changes in empathetic dialogues. Furthermore, previous works [11,12,16,20,25] consider the role of the given situation information as a simple abstract of the conversation and do not leverage it. However, we contend that situation information can be utilized as supplementary knowledge to enrich the conversation context.

Figure 1 shows a dialogue extracted from the EMPATHETIC DIALOGUES dataset. The conversation revolves around a situation where the target predict emotion label is Afraid. The speakers start the conversation with an initial emotional state that changes continuously during the conversation. There are emotional changes during the conversation between Speaker1 and Speaker2, which include internal emotional changes and interactional emotional changes. As the conversation progresses, speaker1's subjective emotion changes from FEAR to NERVOUSNESS, which we consider as internal change. Speaker1's emotional change is also objectively influenced by another speaker, which we consider as interactional emotional change. Usually, Speaker2 would be the chatbot after training, and capturing the Speaker1's emotional changes can help to better understand the Speaker1's emotional needs.

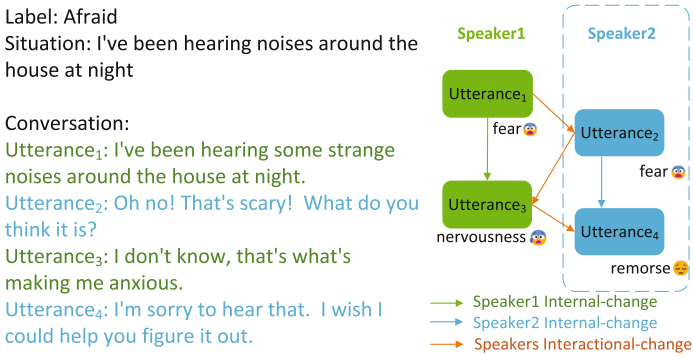


**Fig. 1.** The emotional change between speaker1 and speaker2 during a conversation.

In this paper, we propose a **Mu**lti-**S**cale **E**motional flow graph Dialogue Generation Model, called **MuSE** to simulate speakers' emotional flow. The MuSE model considers the changes in emotional flow as a graph structure, and utilizes graph neural networks to extract features. We first construct an oriented graph to better simulate emotional flow changes in empathetic dialogue. Furthermore, speakers may speak more than one sentence in a conversation, such as $Utterance_2$ and $Utterance_4$ in Fig. 1, and each sentence has a different emotional state. To address this, we introduce a fine-grained sentence expansion strategy to segment these sentences and thus capture more subtle emotional changes of the speaker, which combines multi-scale coarse and fine-grained information. To extend the speakers' initial emotion into the ensuing conversation, we add a key emotion node and key situation node into emotional graph as background information, called $KeyEmotion$ and $KeySituation$. The $KeyEmotion$ is the predicted emotion distribution, without giving away the real label information. The MuSE model first predicts the emotion label through the constructed emotional change graph and generates appropriate responses.

The main contributions of this work are summarized as follows:

- We propose an emotional flow dialogue model that can better capture the emotional changes of the speaker in an empathetic dialogue.
- An oriented emotional dialogue graph is constructed to simulate the changes of speakers' emotion states in empathetic context, in which key emotion and key situation nodes are introduced for the first time to extend the speakers' initial emotion into the ensuing conversation.
- We introduce a contextual fine-grained expansion strategy for empathetic dialogue, which can be combined with the emotional flow graph to better capture the subtle emotional changes of the speakers.
- We conduct experiments on the publicly available EMPATHETIC DIALOGUES dataset. The experimental results show that the MuSE model performs well on both automatic evaluation and human evaluation compared with the existing baseline models.

## 2 Related Work

Research on empathic dialogue in artificial intelligence starts in recent years. The empathic dialogue generation task is first proposed by Rashkin et al. [19]. In empathic dialogue, it is important to identify the emotional needs of speakers and then generate appropriate empathic responses accordingly. Existing work on perceived emotional needs is divided into two directions namely, directly recognizing the speakers' emotions or using external knowledge for indirect reasoning.

The first one is recognizing the speaker's emotion directly. The earliest Rashkin et al. [19] add an additional module for predicting emotions to the model, which can generate empathic responses for the first time. Next, Lin et al. [11] construct multiple decoders, using different decoders to respond to each contextual emotion depending on the speaker's emotional state. The MIME model is proposed by Majumder et al. [12] and they argue that the empathetic responses often mimic the emotion of the user to a varying degree, depending on its positivity or negativity. However, these coarse-grained ERC models lack the ability to capture fine-grained emotions, which may affect the performance of empathic responses. Li et al. [9] argue that the sensitive emotion expressed by the speaker is important. On the basis of these methods, we further considers the state transfer relationships between different fine-grained emotions, i.e., emotional flow changes.

The second one is to infer emotional needs indirectly with the help of external knowledge, which helps the model obtain some additional cues, including the identification of emotional causes, commonsense inferences, etc. Wang et al. [26] utilize the Concept Net external knowledge and construct an emotional causal map through a multi-hop strategy, which in turn generates empathic responses. The CEM model is proposed by Sabour et al. [20], they adopt ATOMIC [21] to access commonsense knowledge. For each sentence, ATOMIC infers six commonsense relations for the person involved in the event. The commonsense can also help to identify conversational emotion, Zhao et al. [29] use ATOMIC to inference each utterance's emotional state for the ERC task. And Wang et al. [27] devise the SEEK model, which utilizes the COMET [1] to detect the intent of each utterance and inference the emotion dynamics. Wang et al. [25] propose the state-of-the-art CARE model, which employs Cause Effect Graph external information to generate an empathetic response. Unlike previous work that used external knowledge to infer emotions indirectly, this paper uses external knowledge to identify speakers' fine-grained emotions directly.

In recent years, due to the powerful representation ability of graph networks, graph-based human-computer conversation models have received increasing attention. Ghosal et al. [3] propose the DialogueGCN model to recognize the emotions in a conversation. Qin et al. [18] use Co-Interactive Graph Attention Network to capture contextual information in conversations and mutual interaction information. Pang et al. [14] propose a MFDG model and construct a multi-factors dialogue graph to detect speakers' intent. Unlike previous works, our emotional flow dialogue graph employs a novel fine-grained strategy to construct a graph structure suitable for empathetic dialogue, capable of highlighting the unconcious emotional states of the speaker during the conversation.

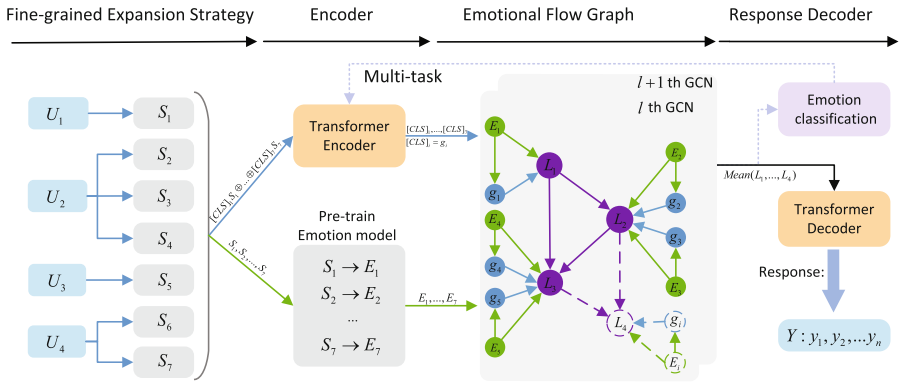


**Fig. 2.** In our proposed MuSE model, blue nodes denote sentences, green nodes denote emotions, and purple nodes denote abstracted aggregation nodes. (Color figure online)

## 3 Our Model: MuSE

In this section, we illustrate the MuSE model in detail, whose architecture is depicted in Fig. 2. The MuSE model contains several parts, which are the fine-grained expansion strategy, encoder, emotional flow graph, and response decoder. The input of the model first goes through the fine-grained expansion strategy to get fine-grained sentences, and the external knowledge of the pre-trained model is used to obtain the fine-grained sentences emotional states, and then the emotional flow graph is constructed to simulate the emotion changes of speakers. Then the graph neural network is used to obtain the contextual representation and the emotion labels, and finally, the learned information is fed into the decoder to generate empathic responses. The goal of the model's emotion classifier, as part of multi-task learning, is to attach emotions to the model's responses and become more empathetic.

We formulate the task of empathetic response generation as follows. Given the $contexts = \{U_1, U_2, ..., U_n\}$, where $n$ is the turns of a dialogue and there are two speakers $speaker1$ and $speaker2$. $\{S_1, S_2, ..., S_i\}$ are the sentences after fine-grained expansion strategy in the utterance, where $i$ represents the $i$th sentence after segment. $\{E_1, E_2, ..., E_i\}$ is the emotional state of the new utterance. $E_i$ is obtained from a pre-trained model with seven emotion classifications, which include $fear$, $sadness$, $neutral$, $joy$, $disgust$, $anger$, and $surprise$ [4]. The target response is $\boldsymbol{Y} = (y_1, y_2, ..., y_n)$.

### 3.1 Fine-Grained Expansion Strategy

In order to capture the subtle emotion states in context and encode the context, we introduce a fine-grained expansion strategy to segment the context and exploit transformer [24] encoder to encode each sentence. Unlike the previous direct segmentation approaches according to speakers, we further split the context based on punctuation marks, which are the period, question mark, and exclamation mark.

Taking the dialogue in Fig. 1 as an example, Table. 1 shows the results after adopting fine-grained expansion strategy. The MuSE model first splits the contexts by different speakers and obtains $U_1$ to $U_4$. The fine-grained expansion strategy further segments sentences based on the speakers' punctuation in the context to capture all the fine-grained speakers' emotions. The sentences $S_1$ to $S_i$ are obtained by fine-grained expansion strategy. $E_1$ to $E_i$ are the results predicted by emotional pre-trained model [4].

**Table 1.** An example from EMPATHETIC DIALOGUES after adopting fine-grained expansion strategy.

| Label | | | | Afraid |
|---|---|---|---|---|
| Situation | | | | I've been hearing noises around the house at night |
| Speaker1 | $U_1$ | $S_1$ | $E_1$:*fear* | I've been hearing some strange noise around the house at night. |
| Speaker2 | $U_2$ | $S_2$ | $E_2$:*surprise* | Oh no! |
| | | $S_3$ | $E_3$:*fear* | That's scary! |
| | | $S_4$ | $E_4$:*neutural* | What do you think it is? |
| Speaker1 | $U_3$ | $S_5$ | $E_5$:*neutural* | I don't know, that's what's making me anxious. |
| Speaker2 | $U_4$ | $S_6$ | $E_6$:*sadness* | I'm sorry to hear that |
| | | $S_7$ | $E_7$:*sadness* | I wish I could help you figure it out. |

The statistics in the table show that after fine-grained strategy segmentation, there is a substantial increase in the number of sentences obtained through fine-grained strategy to capture more subtle speaker emotion states.

### 3.2 Encoder

Transformer block [24] is adopted as the encoder and different from the previous methods of directly splicing contexts [16,20,25], MuSE model concatenates each sentence $S_i$ with the special token $[CLS]$ to obtain the representation respectively. We use the hidden representation of $[CLS]_i$ as the representation of $sentence_i$ in a context. In the encoder, the representation of each sentence $[CLS]_i$ is obtained after transformer block:

$$[CLS]_i = TransformerBlock(S_i), \tag{1}$$

**Table 2.** Statistics of datasets under different splitting strategies.

| Dataset | Origin | Speaker level Split | Fine-grained Split |
|---|---|---|---|
| Train | 40250 | 84686 | 132944 |
| Valid | 5734 | 12188 | 19313 |
| Test | 5255 | 11127 | 18716 |

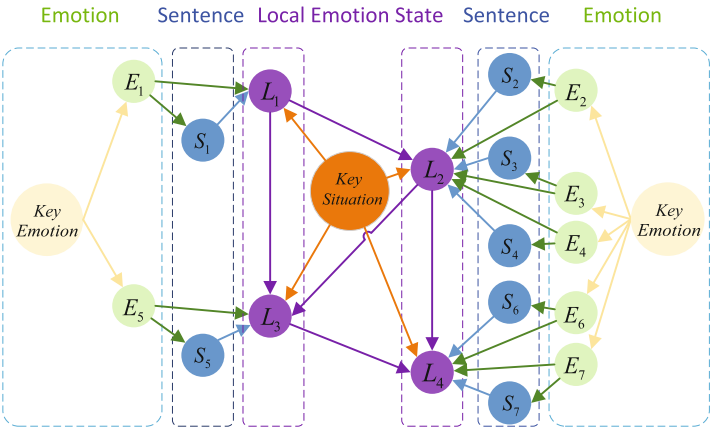


**Fig. 3.** Emotional flow graph is constructed according to Table 1. The left part represents speaker1 and the right part represents speaker2. Blue nodes denote sentences, green nodes denote emotions, and purple nodes denote abstracted aggregation nodes. (Color figure online)

The $[CLS]_i$ vector is used to initialize the nodes vector in Subsect. 3.3 and get the final context vector after Subsect. 3.3 graph network learning. A separate encoder is used to encode the situation and the $[CLS]_{KeySituation}$ can be obtained by the same method as Eq. 1.

### 3.3 Emotional Flow Graph

The speakers' emotional changes during the conversation are not completely chronological, the speakers' emotion state is often influenced by subjective features or objective features. So based on Table 1, we transform context into graph structure and employ Graph Convolutional Networks (GCN) to extract features.

**Graph Construction.** The emotional flow graph is constructed as shown in Fig. 3. Since the splitting strategy is used in Sect. 3.1, in order to integrate the speaker's clause emotion state and semantic information, we introduce $LocalEmotionState$ nodes $L_n$ to aggregate the semantic and emotion information, as the purple node shown in Fig. 3 is the local emotion state area. The number of $L_n$ is equal to the $U_n$. For the first time, we introduce the $KeyEmotion$ node and $KeySituation$ node in the graph to extend the speakers' initial emotion into the ensuing conversation, as different speakers have different pre-existing initial emotions at the beginning of the conversation. The $KeyEmotion$ node is considered as the background information of each segmented sentence, so this will be a uni-directed edge pointing from the $KeyEmotion$ node to the utterance emotion node $E_i$.

**Graph Initialization and Network.** The initialization of the graph embedding can have a significant impact on the model. The $[CLS]$ vector of the transformer block encoder in Subsect. 3.2 is used as the initialized semantic embedding of the sentence node. And the emotion $E_i$ is initialized as word embedding. The $KeyEmotion$ node vector can be calculated as followed:

$$V_{KeyEmotion} = Mean([CLS]_1, [CLS]_2, ..., [CLS]_i). \tag{2}$$

The $KeySituation$ node's initialized embedding is $[CLS]_{KeySituation}$. The sentence node $S_i$ can be represented as the sentence vector $[CLS]_i$, where $i$ is the turn of sentences after segmentation. The $LocalEmotionState$ is an aggregation node, where the average of the sentence vectors of all the partitioned sentences in the current turn is used as the initialization. Taking the $LocalEmotionState$ $L_2$ as an example, the initialization vector can be calculated as:

$$L_2 = Mean({S_2}^{CLS}, {S_3}^{CLS}, {S_4}^{CLS}). \tag{3}$$

Graph neural networks are very effective for modeling structured information like knowledge graphs, the MuSE model uses Graph Convolutional Networks (GCN) [7] to model the flow of emotions in a conversation. In order to ensure that the node vector update order in the directed emotional flow graph can fully simulate the actual emotional flow of the speaker, we design the update pattern of vectors between different node types. Unlike the traditional GCN that updates nodes randomly, the MuSE model controls the update order of different types of nodes, from emotion nodes to sentence nodes to local emotion state nodes. Equations 4, 5 and 6 show the update direction of emotion nodes, sentence nodes and local emotion state nodes respectively. The $^{(l)}$ denotes the node vector in the $l$th layer of the GCN network.

$${E_i}^{(l+1)} = GCN({E_i}^{(l)} | [CLS]_{KeyEmotion}) \tag{4}$$

$${S_i}^{(l+1)} = GCN({S_i}^{(l)} | E_i) \tag{5}$$

$${L_n}^{(l+1)} = L_{n-1}^{(l)} + GCN({L_n}^{(l)} | E_i) + GCN({L_n}^{(l)} | S_i) + GCN({L_n}^{(l)} | [CLS]_{KeySituation}) \tag{6}$$

The GCN layer is calculated as followed:

$$f\left(X^{(l)}, A\right) = \sigma\left(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}X^{(l)}W^{(l)}\right), \tag{7}$$

with $\hat{A} = A + I$, where $A$ denotes the adjacency matrix and $I$ denotes the identity matrix, $\hat{D}$ refers to the diagonal node degree matrix of $\hat{A}$ and $W^{(l)}$ denotes a trainable weight matrix. $\sigma$ refers to a non-linear activation.

The average of aggregated local emotion state nodes $\{L_1, ..., L_n\}$ are used as representatives of the overall context, so the context vector $V_{context}$ can be computed as follows, where $n$ is the number of sentences in a conversation:

$$V_{context} = Mean(L_1, L_2, ..., L_n) \tag{8}$$

We exploit the $V_{context}$ for contextual emotion prediction, calculated as follows, where $FFN$ is feed-forward network:

$$EmotionLable = Softmax(FFN(V_{context})) \tag{9}$$

### 3.4 Decoder

In the decoding process, for each word $y_t$ in $Y$, we employ the mask operation during the training process to avoid the model from seeing the correct response labels in advance. The negative log-likelihood loss can be calculated as:

$$\mathcal{L}_1 = -\sum_{i=1}^{n} \log p(y_n | \{y_1, ..., y_{i-1}\}, v_{context}) \tag{10}$$

And the emotion classification loss $\mathcal{L}_2$ is calculated by cross-entropy loss. All the parameters for our proposed model are trained and optimized based on the weighted sum of the mentioned losses:

$$\mathcal{L}(\theta) = \gamma_1 * \mathcal{L}_1 + \gamma_2 * \mathcal{L}_2, \tag{11}$$

where $\gamma_1$ and $\gamma_2$ are hyper-parameters and $\theta$ is all learnable parameters.

## 4 Experiment

### 4.1 Datasets

We conduct our experiments on the public dataset EMPATHETIC DIALOGUES[1] [19]. It contains 25k dialogues grounded in situations prompted by specific emotion labels. There are 32 evenly distributed categories of emotion labels in this dataset, representing the main emotions in the context of conversation.

### 4.2 Baselines

The following strong baseline models are selected for comparison.

- **Transformer** [24]: An original transformer model based on the seq2seq structure which is the classical generative model.
- **Multi-TRS** [19]: A generative model based on transformer with multi-task for emotion prediction. They built an emotion predictor to capture the speaker's emotions.
- **MoEL** [11]: MoEL can capture the user emotions distribution and softly combine the output states of the appropriate Listener(s). It's also a transformer-based model, which can react to certain emotions and generate an empathetic response.

[1] https://github.com/facebookresearch/EmpatheticDialogues.

- **EmpDG** [9]: EmpDG introduces an interactive adversarial learning framework that exploits user feedback and identifies whether the generated responses evoke emotion perceptivity in dialogues.
- **CEM** [20]: CEM leverages commonsense external knowledge to obtain more information about the user's situation and further enhance the empathy expression in generated responses.
- **CARE** [25]: A Conditional Variational Graph Auto-Encoder model considers the interdependence among causalities and reasons independently. The model utilizes Cause Effect Graph external knowledge to construct a graph.

### 4.3 Experiment Settings

To ensure the fairness of the experiment, we set the parameters of all models to a uniform standard, and the performance of the CEM model [20] after our tuning parameters is higher than the results reported in the original paper. The pre-trained GloVE vector [17] is used to initialize the word embeddings. During the training process, we adopt the adam optimizer with 16 batch size and the learning rate is 0.0001. All the models are trained on NVIDIA RTX 3090 GPU.

### 4.4 Automatic Evaluation

In order to compare with strong baseline models such as CARE [25], CEM [20] and SEEK [27] etc., we use the evaluation metrics from most papers and take Perplexity (PPL), Distinct-$n$(Dist-$n$) [8], BLEU [15], Rouge [10], Bert Score [28] as main automatic metrics. However, previous works including the CEM model [20] only focus on the Dist-$n$ & Acc indicators and the CARE's authors believe that the BLEU and BertScore are more important, so they don't use the Dist-$n$ & Acc indicators. We believe that all metrics matter because there are limitations to using distinct indicators alone to evaluate models. And we will introduce details in Subsect. 4.6.

The indicators are introduced as followed: (1) Perplexity (PPL) focuses on the model's confidence in response generation. (2) Distinct-$n$[2] measures the generated response's diversity. (3) BLEU[3] estimates the matching between n-grams of the generated response and those of the golden response. (4) The Rouge-L[4] indicator is very similar to the BLEU indicator, which is used to measure the matching degree between the generated results and the standard results. The difference is that ROUGE is based on the recall rate, while BLEU pays more attention to the accuracy rate. (5) BertScore[5] is based on the pre-trained model, uses context embedding to describe sentences, and calculates the semantic similarity between two sentences. BertScore has precision, recall and F1 metrics, and F1 value is influenced by precision and recall. We use the F-BERT to evaluate

[2] https://github.com/Sahandfer/CEM/blob/master/src/scripts/evaluate.py.
[3] https://github.com/mjpost/sacrebleu.
[4] https://github.com/pltrdy/rouge.
[5] https://github.com/Tiiiger/bert_score.

the model. On the PPL metric, a smaller value indicates a better model, and on the other metrics, a larger value indicates a better model.

Admittedly, BLEU and ROUGE can effectively evaluate the performance of models, and many models have adopted such evaluation metrics, but their experimental results are difficult to be compared under the same criteria due to the different calculation methods. We find several convenient, fast, and fair calculation methods from highly recognized repositories in GitHub and believe that they can significantly reduce the workload of researchers in evaluating indicators. The evaluation code is available at MuSE[6].

**Table 3.** Results of automatic evaluation on EMPATHETIC DIALOGUES.

| Model | PPL ↓ | Dist-1↑ | Dist-2↑ | Acc↑ | BLEU↑ | Rouge-L↑ | F-BERT↑ |
|---|---|---|---|---|---|---|---|
| Transformer | 34.7083 | 0.4918 | 2.3134 | – | 2.6616 | 16.7549 | 22.7056 |
| Multi-TRS | 34.8442 | 0.4882 | 2.3594 | 35.85 | 2.7565 | 17.3173 | 22.7507 |
| MoEL | 35.5586 | 0.5632 | 2.8986 | 34.60 | 2.8129 | 17.1865 | 23.8777 |
| EmpDG | 34.4143 | 0.5693 | 3.1470 | 33.15 | 2.8297 | 18.1459 | 23.5411 |
| CEM | 34.9705 | 0.6180 | 3.058 | 39.07 | 2.5781 | 17.2128 | 23.2403 |
| CARE | 33.8397 | 0.5776 | 2.3096 | – | 2.8300 | 18.2122 | 23.2610 |
| MuSE | **33.5451** | **0.6476** | **3.4380** | **42.99** | **2.8397** | **18.3105** | **24.2781** |

Unlike previous models that excel in a single aspect, the MuSE model demonstrates superior performance on multiple evaluation metrics. From Table 3, we can find that the MuSE model with emotional dialogue graph achieves superior results in the majority of indicators. In terms of Dist-$n$ metric, the emotional dialogue graph significantly improves the word richness of responses. Unlike previous works that utilized external knowledge, the MuSE model fuses fine-grained external knowledge of speakers' subtle emotions by constructing a dialogue graph and interacts with coarse-grained information at the contextual utterance level. We believe this is a new approach for empathetic dialogue that can capture speakers' emotional state changes at multiple scales. Emotion recognition accuracy (Acc) is a measure of how accurately the model captures the emotion state of the context. It can be seen from the table that the MuSE model can better capture the emotional states of the user due to the emotional flow graph structure. Because of the special structure of the CARE model, they don't provide Acc value from their paper.

The BLEU, Rouge-L, and F-BERT indicators focus more on the difference between the generated responses and golden truth sentences. The MuSE model achieves the best results compared with all baseline models in the BLEU, Rouge-L, and F-BERT metrics. The MuSE model achieves a significant superiority on both rigorous tests of evaluation metrics. We believe that this enhancement

[6] https://github.com/DericZhao/MuSE/evaluate.py.

is influenced by the constructed emotional dialogue graph, during which the speakers' emotional changes are important and the emotional flow dialogue graph captures this subtle characteristic information.

### 4.5 Human Evaluation

Besides the automatic evaluation, we also conduct human evaluation at the same time. We follow the evaluation method introduced in [9,11,20,25] from three perspectives. (1) Empathy, which measures the level of understanding of speakers' emotions. (2) Relevance, which measures the consistency of the topic, and relevance of responses to the context. (3) Fluency, which measures whether the response is linguistically sound and grammatically accurate. Each sentence corresponds to a 5-level score, where 5 is the best. We recruit 5 evaluators to judge the response from three aspects and each evaluator has a research interest in natural language processing and has obtained a master's degree. Then we compute the average value for each metric.

**Table 4.** Results of human evaluation on EMPATHETIC DIALOGUES

| Model | Empathy | Relevance | Fluency |
|---|---|---|---|
| Multi-TRS | 2.58 | 2.27 | 3.99 |
| MoEL | 2.66 | 2.29 | 4.24 |
| EmpDG | 2.35 | 2.43 | 4.18 |
| CEM | 2.52 | 2.41 | 4.80 |
| CARE | 2.99 | 3.09 | 4.75 |
| MuSE | **3.29** | **3.12** | **4.88** |

We evaluate the above classical models and the strong baseline models. As shown in Table 4, our model shows significant improvement over the other models in empathy, relevance, and fluency metrics. In terms of empathy degree, the MuSE model can better understand the speakers' intentions through emotional changes. From the relevance metric, the model captures fine-grained emotional changes between speakers, with more accurate control over whole contextual emotion, thus generating empathic responses. The fluency metric reflects model's convergence degree and it can be found that the MuSE model can answer more fluently than the previous models.

The above human evaluation results can also prove that capturing the speakers' fine-grained emotional changes is important to improve the performance of empathy dialogue.

### 4.6 Ablation Experiments

In the ablation experiments, we conduct some experiments separately to investigate the importance of different modules of the MuSE model. First, we change

the oriented emotional flow graph to an undirected graph to verify the correctness of our proposed oriented emotional flow graph. Second, we remove the *KeyEmotion* node from the graph and in the third experiment, we remove the *KeySituation* node from the graph to evaluate the performance of the key nodes we proposed. Finally, we validate it is necessary to employ multiple metrics to evaluate the empathetic model through MuSE(GoEmotion).

**Table 5.** Results of ablation studies on EMPATHETIC DIALOGUES.

| Model | PPL ↓ | Dist-1↑ | Dist-2↑ | Acc↑ | BLEU↑ | Rouge-L↑ | F-BERT↑ |
|---|---|---|---|---|---|---|---|
| MuSE | 33.5451 | 0.6476 | 3.4380 | **42.99** | **2.8397** | **18.3105** | **24.2781** |
| w/o Directional | 33.7078 | **0.6490** | 3.3116 | 41.92 | 2.6090 | 17.9899 | 24.1727 |
| w/o *KeyEmotion* | **33.4750** | 0.6324 | **3.4897** | 42.09 | 2.7902 | 18.1466 | 24.2596 |
| w/o *KeySituation* | 33.9445 | 0.5620 | 2.8363 | 33.40 | 2.8267 | 17.4461 | 23.8502 |
| MuSE(GoEmotion) | 33.9068 | **0.7286** | **3.8021** | 42.82 | 2.3294 | 17.2796 | 23.4009 |

From Table 5, we can find that the results of undirected graphs have a decrease in most of the metrics. After the case study, we find that the improvement in the Dist-$n$ metric is due to generate more context-irrelevant words, which can be confirmed by the decrease in BLEU values. After removing the *KeyEmotion* node, there is a slight decrease in the evaluation metrics. However, after removing the *KeySituation* node, the model's performance drop sharply. After the above ablation experiments, we believe that the background contextual word embedding information at the emotion word level has not as much impact on the whole model as the situation node sentence embedding information at the sentence level. The key situation contains richer and more accurate preexisting information, which also proves that it is very effective to introduce key nodes to the emotional dialogue graph.

In Table 5, MuSE (GoEmotion) replaces the original external knowledge pre-trained model with 'GoEmotion' [22] for emotion recognition from $E_1$ to $E_n$. The GoEmotion model has 27 emotion types, while the pre-trained model used by MuSE has seven emotion types. As can be seen from the Table 5, the MuSE(GoEmotion) model performs much better than the original MuSE model proposed in this paper if Dist-$n$ is used as the evaluation metric. Compared with other strong baseline models, there is also an improvement in Dist-$n$ metric according to the data in Table 3. But according to the case study in Table 6, it can be found that the improvement in Dist-$n$ values is due to the fact that the MuSE(GoEmotion) generates context-irrelevant response, resulting in a certain improvement in richness of sentences. This also demonstrates that complex external information leads to poor model generation quality. Therefore, we choose all metrics to better measure the empathic dialogue generation task in this paper.

**Table 6.** Case study of the MuSE model.

| Emotion: | Sentimental |
|---|---|
| Context: | |
| Speaker1: | Over the summer my family went to alabama to visit family and I felt a lot of emotions. |
| Speaker2: | I guess you do not live in alabama? was it overall a good trip .... or challenging? or both! |
| Speaker1: | No, I do not live in alabama<br>We drove and trip was overall very good as sometimes it can<br>be challenging when driving such a long distance. |
| Golden: | Well, If the drive was the most challenging it sounds like a good trip!<br>Too often it is family dynamics that are more challenging<br>and not so easily taken care of . |
| **MuSE:** | That is so cool!<br>I am glad you were able to get to go back to see them. |
| MuSE (Go Emotion): | I am glad you are able to find it! (Context-irrelevant) |

## 5 Conclusion and Future Work

In this paper, we argued that modeling the constantly changing emotional flow of speakers is crucial and propose the MuSE model to simulate speaker interactions. To capture speakers' subtle emotions, we introduced a new fine-grained expansion strategy that enriches the emotional flow graph through multi-scale fusion. By controlling the direction and order of node update information, we simulated emotion transmission between speakers on the directed emotional flow graph. We added the emotion and situation nodes into the emotional flow graph that extends the speakers' initial emotion into the ensuing conversation. Experimental results demonstrated the superior performance of the MuSE model. In the future, we will further explore the strategy of constructing fine-grained conversation graphs in conversation graphs.

**Acknowledgments.** This work is supported by the National Natural Science Foundation of China (61672144, 61872072).

**Ethics Statement.** We do not observe direct ethical and security issues arising from the emotional dialogue itself. The public dataset used in this paper may contain user privacy, but it has been made harmless in the earliest published dataset papers.

## References


1. Bosselut, A., Rashkin, H., Sap, M., Malaviya, C., Celikyilmaz, A., Choi, Y.: Comet: commonsense transformers for automatic knowledge graph construction. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 4762–4779 (2019)
2. Gao, J., et al.: Improving empathetic response generation by recognizing emotion cause in conversations. In: Findings of the association for computational linguistics: EMNLP 2021 (2021)
3. Ghosal, D., Majumder, N., Poria, S., Chhaya, N., Gelbukh, A.: Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 154–164 (2019)
4. Hartmann, J.: Emotion english distilroberta-base (2022). https://huggingface.co/j-hartmann/emotion-english-distilroberta-base/
5. Keskin, S.C.: From what isn't empathy to empathic learning process. Procedia Soc. Behav. Sci. **116**, 4932–4938 (2014)
6. Kim, H., Kim, B., Kim, G.: Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 2227–2240 (2021)
7. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks. In: International Conference on Learning Representations (2017). https://openreview.net/forum?id=SJU4ayYgl
8. Li, J., Galley, M., Brockett, C., Gao, J., Dolan, W.B.: A diversity-promoting objective function for neural conversation models. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 110–119 (2016)
9. Li, Q., Chen, H., Ren, Z., Ren, P., Tu, Z., Chen, Z.: Empdg: Multi-resolution interactive empathetic dialogue generation. In: Proceedings of the 28th International Conference on Computational Linguistics, pp. 4454–4466 (2020)
10. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: Text summarization branches out, pp. 74–81 (2004)
11. Lin, Z., Madotto, A., Shin, J., Xu, P., Fung, P.: Moel: mixture of empathetic listeners. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 121–132 (2019)
12. Majumder, N., et al.: Mime: Mimicking emotions for empathetic response generation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 8968–8979 (2020)
13. Paiva, A., Leite, I., Boukricha, H., Wachsmuth, I.: Empathy in virtual agents and robots: a survey. ACM Trans. Interact. Intell. Syst. (TiiS) **7**(3), 1–40 (2017)

14. Pang, J., Xu, H., Song, S., Zou, B., He, X.: Mfdg: A multi-factor dialogue graph model for dialogue intent classification. In: Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2022, Grenoble, France, September 19–23, 2022, Proceedings, Part II. pp. 691–706. Springer (2023). https://doi.org/10.1007/978-3-031-26390-3_40
15. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pp. 311–318 (2002)
16. Peng, W., Hu, Y., Xing, L., Xie, Y., Sun, Y., Li, Y.: Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. In: Raedt, L.D. (ed.) Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23–29 July 2022, pp. 4324–4330. ijcai.org (2022). https://doi.org/10.24963/ijcai.2022/600, https://doi.org/10.24963/ijcai.2022/600
17. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)
18. Qin, L., Li, Z., Che, W., Ni, M., Liu, T.: Co-gat: A co-interactive graph attention network for joint dialog act recognition and sentiment classification. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 13709–13717 (2021)
19. Rashkin, H., Smith, E.M., Li, M., Boureau, Y.L.: Towards empathetic open-domain conversation models: A new benchmark and dataset. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5370–5381 (2019)
20. Sabour, S., Zheng, C., Huang, M.: Cem: Commonsense-aware empathetic response generation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 36, pp. 11229–11237 (2022)
21. Sap, M., et al.: Atomic: An atlas of machine commonsense for if-then reasoning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, pp. 3027–3035 (2019)
22. Savani, B.: Bert base go emotion. https://huggingface.co/bhadresh-savani/bert-base-go-emotion (2021)
23. Tu, Q., Li, Y., Cui, J., Wang, B., Wen, J.R., Yan, R.: Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 308–319 (2022)
24. Vaswani, A., et al.: Attention is all you need. Adv. Neural Inform. Process. Syst. **30** (2017)
25. Wang, J., Cheng, Y., Li, W.: Care: Causality reasoning for empathetic responses by conditional graph generation. In: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022 (2022)
26. Wang, J., Li, W., Lin, P., Mu, F.: Empathetic response generation through graph-based multi-hop reasoning on emotional causality. Knowl.-Based Syst. **233**, 107547 (2021)
27. Wang, L., et al.: Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection. In: Findings of the Association for Computational Linguistics: EMNLP 2022, pp. 4634–4645. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Dec 2022), https://aclanthology.org/2022.findings-emnlp.340

28. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with bert. In: International Conference on Learning Representations (2020). https://openreview.net/forum?id=SkeHuCVFDr
29. Zhao, W., Zhao, Y., Lu, X.: Cauain: Causal aware interaction network for emotion recognition in conversations. In: Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI, pp. 4524–4530 (2022)