Non-Emotion-Centric Empathetic Dialogue Generation

Yuanxiang Huangfu, Peifeng Li*, Yaxin Fan and Qiaoming Zhu
School of Computer Science and Technology, Soochow University, Suzhou, China
yxhuangfu@stu.suda.edu.cn, pfli@suda.edu.cn
yxfansuda@stu.suda.edu.cn, qmzhu@suda.edu.cn

Abstract

Previous work on empathetic response generation mainly focused on utilizing the speaker's emotions to generate responses. However, the performance of identifying fine-grained emotions is relatively poor, which results in cascading errors in generating empathetic responses. Moreover, due to the conflict between the information in the dialogue history and the recognized emotions, previous work often generated general and uninformative responses. To address the above issues, we propose a novel framework **NEC** (**N**on-**E**motion-**C**entric empathetic dialogue generation) based on contrastive learning and context-sensitive entity and social commonsense, in which the frequent replies and sentences with incorrect emotions are punished through contrastive learning, thereby improving the empathy, diversity and information of the responses. The experimental results demonstrate that our NEC enhances the quality of empathetic generation and generates more diverse responses in comparison with the state-of-the-art baselines. The code will be available at https://github.com/huangfu170/NEC-empchat

1 Introduction

In social psychology theory, empathy is expressed in two dimensions (Gerdes et al., 2010): (1) the physiological experience of feeling what another person is feeling (Batson et al., 1987); and (2) the cognitive processing of these feelings (Hoffman, 2001). Empathic ability in human dialogue enables individuals to comprehend each other's experiences and emotions, thereby fostering more intimate interpersonal relationships (Keskin, 2014). Empathetic response generation aims to generate empathetic responses (e.g., comfort, sympathy, and understanding) to the speaker by thoroughly comprehending the speaker's background and emotional state. It is a widely used technique in open-domain dialogue systems, such as digital humans (Shen et al., 2021) and interactive entertainment (Shao et al., 2023), with the objective of enhancing the user-system interaction experience.

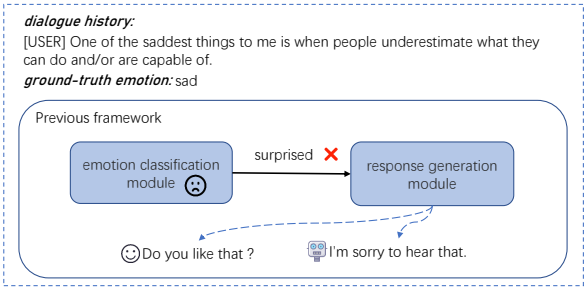

Figure 1: Two common errors in Empathetic Dialogue System: (1) Emotional Error: responses that deviate entirely from the speaker's situation due to errors in emotion recognition. (2) Uninformative Response: namely, the generation of responses that are overly generic.

Previous work (Lin et al., 2019; Xie and Pu, 2021; Pang et al., 2023) on empathetic response generation for neural dialogue systems has primarily focused on accurately identifying emotions and subsequently utilizing the identified emotions to guide the generation of empathetic responses. Nevertheless, the challenge of fine-grained emotion recognition (e.g., EMPATHETICDIALOGUES encompassed 32 distinct emotional categories) remains a significant barrier, with an accuracy rate below 50% (Gao et al., 2021; Zhao et al., 2023; Cai et al., 2023). Moreover, a dialogue may encompass various emotions, which further challenges the model of emotion identification, preventing it from accurately recognizing the appropriate emotions. This leads to cascading errors in the generation of empathetic responses. Figure 1 illustrates how the model incorrectly identifies a sad scenario as "surprised", resulting in cascading errors that affect the model's output. This leads to the model to ask the speaker *"Do you like that?"*, which is

* Corresponding author

Proceedings of the 31st International Conference on Computational Linguistics, pages 989–999
January 19–24, 2025.

inconsistent with the scenario described by the conversation history. Such errors are unacceptable in empathic response generation and may have more serious consequences for the speaker than those in the emotion-unaware dialogue system.

Furthermore, a discrepancy may arise when the incorrectly identified emotion does not align with the dialogue history. In other words, the information obtained from the dialogue history is contaminated with the emotion that is incorrectly identified. This ultimately leads to the model being unable to generate responses that are both informative and appropriate. Figure 1 illustrates an example where the model generates an uninformative generic sentence *"I'm sorry to hear that"* at a high frequency due to the conflict between the incorrectly identified emotion and the dialogue history. This will lead to less consistency and empathy, affecting the speaker's desire to continue the conversation.

In light of these considerations, a crucial issue emerges concerning the means of minimizing the influence of emotion recognition errors on the model, while concurrently reducing the prevalence of uninformative or non-empathetic sentences generated by the model. To address the above issue, we propose a novel framework **NEC** (**N**on-**E**motion-**C**entric empathetic response generation) which incorporates context-sensitive entity and social commonsense, as well as contrastive learning. Specifically, we first utilize contrastive learning to replace the explicit emotion recognition, which can constrain the language model's inferences in correct emotions. Furthermore, we introduce context-related entity and social commonsense knowledge to enhance the model's comprehension of dialogue history and the emotion information embedded within entities, which can also assist the model in making accurate emotion judgments. The experimental results indicate that our NEC enhances the quality of empathetic generation and generates more diverse responses in comparison with the SOTA baselines.

2 Related Work

Dataset Empathetic response generation has recently garnered significant attention, as it necessitates not only generating contextually relevant responses but also expressing empathy towards the recipient of the reply. Rashkin et al. (2019) released an empathetic dialog dataset EMPATHETICDIALOGUES in which the annotators were first provided with an emotional word and then were tasked with engaging in an empathic conversation with another individual based on the given emotion.

Empathetic response generation The majority of prior research has primarily focused on two principal areas: firstly, the accurate identification of emotions and secondly, the enhancement of the interaction between emotions and responses. For example, Lin et al. (2019) devised distinct decoders for different emotions and subsequently aggregated the output via a meta-listener. Li et al. (2020) underscored the pivotal role of emotion recognition, which employed emotion lexicons to identify emotions and incorporated a feedback mechanism to detect the consistency between the generated content and the context. Majumder et al. (2020) utilized emotion embedding to discern emotions and employed the mimic approach to incorporate emotional information into the decoder. However, all of them designed their models based on the recognized emotions. Due to the poor capacity of emotion recognition, the above models are instead affected by cascading errors in emotion detection.

In recent years, the advancement of knowledge graphs and commonsense models have prompted researchers to recognize the potential of utilizing commonsense knowledge to enhance the performance of open-domain dialogue models (Tu et al., 2022; Lee et al., 2022). Two notable examples are ConceptNet (Speer et al., 2017) and ATOMIC (Hwang et al., 2021), which have been widely used to enhance the capacity of natural language understanding (NLU) (Ghosal et al., 2020) and natural language generation (NLG) (Liu and Kilicoglu, 2023; Strathearn and Gkatzia, 2021). For example in empathetic response generation, CEM (Sabour et al., 2022) utilized commonsense knowledge to enhance the model's capacity, leveraging the underlying social commonsense reasoning embedded in the conversational history. DCKS (Cai et al., 2023) dynamically selected commonsense knowledge based on emotional states and context, establishing a dynamic connection between emotional states and commonsense.

3 Methodology

Figure 2 illustrates the framework of our NEC, in which we leverage a Transformer-based pretrained model, i.e., BART, as the backbone. Given a conversation history, we employ various prompt templates to extract entities and social commonsense

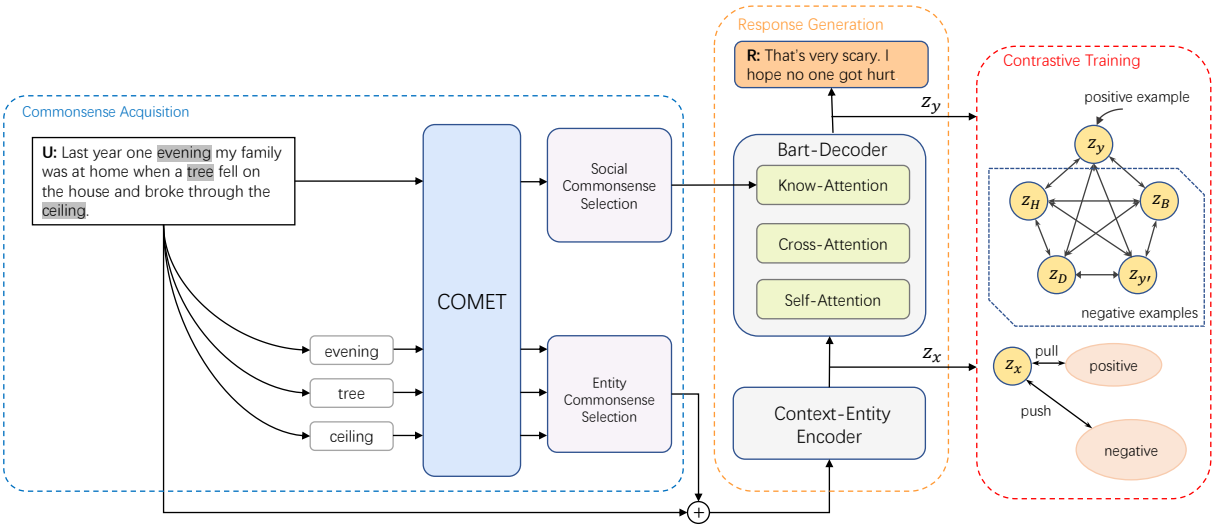

Figure 2: The architecture of our framework NEC, where $\mathbf{z_x}$, $\mathbf{z_y}$ are the hidden states of encoder and the ground truth y, and $\mathbf{z_B}$, $\mathbf{z_D}$, $\mathbf{z_H}$, $\mathbf{z_{y'}}$ denote the representations of from-batch samples, the different emotion samples, the high-frequency samples and the self-generated samples, respectively.

from COMET (Hwang et al., 2021). After a selecting process, they are then injected into the encoder and decoder, respectively (Section 3.2 and 3.3). Subsequently, we further train the model based on the contrastive learning training strategy that we designed, which can avoid the cascade errors associated with emotion recognition (Section 3.4).

3.1 Task Formulation

Empathetic response generation involves a dialogue model processing historical information and subsequently producing a language modeling probability distribution. Formally, let $U = \{u_1, \cdots, u_{n-1}\}$ denote a dialogue history composed of n-1 utterances, and $Y = \{y_1, \cdots, y_M\}$ denotes the empathetic response of M tokens. Let $K_E = \{ke_1, \cdots, ke_m\}$ and $K_S = \{ks_1, \cdots, ks_l\}$ represent the entity and the social knowledge set, respectively. The goal is to generate appropriate and informative responses using the dialogue history U and the commonsense knowledge K_E and K_S as follows. A complete training example can be found at Appendix A.

$$\arg\max P(Y|U, K_E, K_S) \tag{1}$$

3.2 Entity Commonsense Integration

Entities are words that frequently appear in conversations and can also reflect the speaker's emotions. For example, calling someone "my baby" indicates feelings of love, while "bastard" indicates feelings of disgust in most cases. Moreover, the same entity exhibits different properties in different contexts. Consequently, inferring the meaning of an entity from its context necessitates human-like commonsense. In order to address this issue, we employ the dialogue history to generate context-aware commonsense for entities within the dialogue. In particular, for each entity and relation, we utilize the conversation history $U = u_1 \oplus u_2 \oplus \cdots \oplus u_{n-1}$ as the head, and the input of COMET, i.e., IE_i, is as follows.

$$IE_i = U \oplus P_c \oplus ent_j \oplus [rel_i] \tag{2}$$

where P_c represents the prompt connecting context and entity, which is referred to the phrase "in this case". ent_j is the j-th entity and rel_i is one of the following entity relations in COMET: 1) the entity's usage ($ObjectUse$), 2) the location of the entity ($AtLocation$), 3) the composition of the entity ($MadeUpOf$), or 4) the inherent properties of the entity ($HasProperty$), respectively. The corresponding special relation tokens are appended to the entity to prompt COMET to generate the triplets (i.e., <$ke_{head}, ke_{rel}, ke_{tail}$> where $head$, rel and $tail$ refer to entity, relation and commonsense knowledge, respectively) for each relation rel_i. All triplets extracted by COMET are denoted as K_{all}.

Inspired by SAKDP (Wu et al., 2022), we construct positive/negative sample pairs to select entity commonsense that is conducive to response generation. Subsequently, we train a BERT model as a scorer to discern the relevance of the current knowledge triplet to the reply. The final selection is made

from the top m[1] triplets in K_{all} with the highest scores, denoted as K_E.

Finally, the pre-trained BART encoder is used to jointly encode the context and the selected entity knowledge, resulting in context-entity hidden states. The input to the encoder is as follows.

$$I_{enc} = U \oplus [SEP] \oplus P_r \oplus L(K_E) \quad (3)$$

where P_r represents the prompt sentence "*The following knowledge facts are highly relevant to the left query:*" to instruct the model on the relation between dialogic history and subsequent commonsense. $L(K_E)$ represents the linearization method. Specifically, we linearized each obtained entity's common knowledge triplet <ent, rel, $tail$> in the selected set K_E into "ent, rel, $tail$". The output of the encoder is as follows.

$$\mathbf{z_x} = BART\_Encoder(I_{enc}) \quad (4)$$

3.3 Social Commonsense Injection

We obtain social commonsense knowledge following Cai et al. (2023), and use the last utterance of the dialogue history as input to acquire relevant knowledge as follows.

$$I_{rel} = u_{n-1} \oplus [SR] \quad (5)$$

where SR represents five types of relations in social commonsense used in Cai et al. (2023) as follows: 1) the impact of an event on the involved person X ($xEffect$), 2) X's reaction to the event ($xReact$), 3) X's intentions before the occurrence of the event ($xIntent$), 4) what X needs for the event to happen ($xNeed$), and 5) what X desires after the event has occurred ($xWant$). We also use the special tokens ([$xEffect$], [$xReact$], [$xIntent$], [$xNeed$], [$xWant$]) concatenated at the end for replacement and then prompt COMET as mentioned in the previous section to obtain unfiltered social commonsense set $\{ks_1, ..., ks_i, ...\}$.

Given the existence of multiple candidate inferences for each relation, we employ MPNet (Song et al., 2020), which has been pre-trained on the sentence similarity task[2], to calculate the cosine similarity between the hidden vectors of commonsense and the context as follows, which allows us to select the knowledge with the highest score per relation.

$$\mathbf{h_{ctx}}, \mathbf{h_{ks,i}} = MPNet(U), MPNet(ks_i) \quad (6)$$

$$Score(ks_i) = \frac{\mathbf{h_{ctx}} \cdot \mathbf{h_{ks,i}}}{\|\mathbf{h_{ctx}}\| \cdot \|\mathbf{h_{ks,i}}\|} \quad (7)$$

where $\mathbf{h_{ctx}} \in \mathbb{R}^{1\times d}$, $\mathbf{h_{ks}}^T \in \mathbb{R}^{d\times 1}$ represent the hidden vectors of the dialogue history and each candidate social commonsense ks_i, respectively. Social commonsense, which can be defined as a coarse-grained form of knowledge, is integrated into a standard Transformer Decoder by the addition of a multi-head attention layer. This layer facilitates the fusion of contextual information and social commonsense, thereby enhancing the overall performance. The collection of social commonsense K_S is comprised of selected knowledge of various relations with highest similarity. The decoder representation during training stage is denoted as follows.

$$\mathbf{z_y} = BART\_Decoder(K_S, \mathbf{z_x}, Y) \quad (8)$$

3.4 Contrastive Training

Although there has been an enhancement from the incorporation of commonsense knowledge, it remains insufficient as a control signal for generating empathetic responses. Contrastive learning (Chen et al., 2020; Wu et al., 2020; Giorgi et al., 2021) is a widely used method in representation learning, which has been widely applied to various NLP tasks, including dialogue response generation (Dai et al., 2021; Tang et al., 2023).

In light of recent research indicating that language models can be directed towards the generation of more informative sentences through the construction of judicious positive and negative examples (Kalkstein et al., 2020; Li et al., 2022), we employ contrastive training to enhance the model's performance in empathy and diversity for two primary purposes: (1) Using the sentences with high frequency as negative examples can correct the model's tendency to generate generic and repetitive sentences. (2) Utilizing the responses with emotions as negative samples, which are different from the current example, can guide the model to generate responses related to emotions. This approach mitigates the cascading errors introduced by the models centered around emotion recognition.

We employed a similar contrastive learning approach to that described in An et al. (2022) to penalize the output of the model. The constructed negative examples are categorized into four parts as follows, which denoted as D, y', B, and H, respectively.

[1]We selected top 3 triplets during the test stage.

[2]https://huggingface.co/microsoft/mpnet-base

Different emotions (D) To enhance emotional expression, the responses are sampled with emotions that differ from the current target sentence. This is used to replace the emotion recognition module, thereby constraining the model to the appropriate emotion output.

Self-generated (y') As the model frequently generates more generalized sentences, the predictions of beam search can serve as implicit penalties for universally applicable sentences.

From-batch (B) Other samples in the same mini-batch.

High-frequency (H) Explicit universal sentence penalties are introduced by selecting the most frequent sentences from the training set.

In all settings, the target response y^+ is employed as positive example for contrastive learning. During the training stage, the model is augmented by incorporating negative examples y^- to refine the model. The training objective of contrastive learning, L_{total}, is as follows.

$$L_{total} = L_{NLL} + L_{CL} \tag{9}$$

$$L_{NLL} = -\sum_{t=1}^{N} \log P(y_t|I_{enc}, K_S, y_{<t}) \tag{10}$$

$$L_{CL} = \sum_{(y^+,y^-)\in\mathcal{P}} \max\{0, D(\mathbf{z_x}, y^+, y^-) + \xi\} \tag{11}$$

$$D(\mathbf{z_x}, y^+, y^-) = \cos(\mathbf{z_x}, \mathbf{z_{y^-}}) - \cos(\mathbf{z_x}, \mathbf{z_{y^+}}) \tag{12}$$

where $y_t \in Y$, and $\mathcal{P}$ denotes the sample set of contrastive learning. $\mathbf{z_{y^-}}$ and $\mathbf{z_{y^+}}$ represent the hidden states of $y^- \in \{D, y', B, H\}$ and y^+, respectively. We further set $\xi = \gamma * (rank(y^-) - rank(y^+))$ following (An et al., 2022), which reflects the quality difference in these pairs, where γ is a hyperparameter controlling the strength of contrast. We set 0.01 to it for all experiment settings following (An et al., 2022). cos means cosine similarity to the representation.

During the inference process, we leverage the acquired similarity method to augment the beam search, and the ultimate decoding target is to find the sequence y^* from beam search results that maximizes the amalgamation of the acquired similarity score S_{sim} and the generative probability S_{lm} derived from the language model as follows.

$$S_{sim} = \cos(\mathbf{z_x}, \mathbf{z_{\hat{y}}}) \tag{13}$$

$$S_{lm} = \prod_{t=0}^{n} P(\hat{y}_t|\mathbf{z_x}, \hat{y}_{<t}) \tag{14}$$

$$y^* = \arg\max_{\hat{y}} \{\alpha S_{sim} + (1-\alpha) S_{lm}\} \tag{15}$$

where $\mathbf{z_{\hat{y}}}$ is the hidden vector of the search results $\hat{y}$, and α is the balance hyperparameter of contrastive learning and language model likelihood.

4 Experimentation

In this section, we first describe the dataset, and automatic/human evaluation methods, then introduce seven strong baselines. Finally, we report the overall experimental results.

4.1 Dataset

We evaluate our NEC on the EMPATHETICDIALOGUES dataset (Rashkin et al., 2019), which is composed of 25K open-domain dialogues in empathy expression. In a conversation, the speaker conceptualizes a personal experience based on a given emotion and engages in an empathic dialogue with the listener based on that experience, and each conversation is labeled with only one emotion word, which is not utilized in this paper. Following previous work (Cai et al., 2023), we split the train/valid/test set by 8:1:1 and use the same preprocessing functions.

4.2 Implementation Details

The PyTorch framework and the base version of BART are employed as the encoder and decoder[3], respectively. The AdamW optimizer with an initial learning rate of 0.00001 is utilized, with betas parameters set to 0.9 and 0.999. The 2020 version of COMET is employed for the generation of commonsense knowledge [4]. We use NLTK[5] package to extract entities for entity commonsense integration. In the preliminary training phase, the loss function is chosen to plateau at a value indicative of the end of the training process, with a batch size of 8. In contrastive learning, the batch size is 8 during training and 16 during testing, with the parameter α set to 0.7. An early stopping mechanism is employed to prevent overfitting, and the optimal model results on the test set are reported. The decoding strategy employed is beam search, with a maximum decoding step of 30 during inference. The training and testing of the model are conducted on a NVIDIA A100-PCIE (40GB).

[3] huggingface.co/facebook/bart-base
[4] github.com/allenai/comet-atomic-2020/
[5] https://www.nltk.org/

4.3 Automatic Evaluation Metrics

We use Perplexity (**PPL**), BLEU-n (**B-n**) (Papineni et al., 2002), Distinct-n (**Dist-n**) (Li et al., 2016) to automatically evaluate the model. PPL represents the confidence level of the model in the response. A lower PPL indicates a higher probability that the model will predict the response. BLEU-n is used to measure the n-gram similarity between the model-generated responses and the ground truth. A higher BLEU-n value indicates a closer approximation. Following previous work, we also use BLEU 1-4 to illustrate model performance. Distinct-n represents the proportion of unique n-grams generated by the model. It provides insight into the word-diversity of the model, which is particularly significant for open-domain dialogue systems. Furthermore, we introduce a new metric, **Sent-Std**, to measure sentence-level diversity instead of n-gram level as follows.

$$\text{Sent-Std} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\text{count}(s_i) - \mu)^2} \quad (16)$$

where μ represents the average frequency of sentences generated across the test dataset. s_i and N represent the certain sentence and the number of different sentences that have appeared. A small Sent-Std score indicates a more evenly distributed sentence occurrence in the generated responses, which will be used to measure the diversity of different models at the sentence level, that none of the previous work has explored.

4.4 Human Evaluation Methods

Given that automated evaluation results can only assess model's consistency with the ground truth, they are incapable of gauging the level of empathy conveyed in the responses. Consequently, as in the approach employed by CEM (Sabour et al., 2022), we adopt a human evaluation to complement the shortcomings of automatic metrics by considering the model's performance across four dimensions: (1) Coherence (**Coh.**): the relevance between response and context; (2) Empathy (**Emp.**): the capacity to comprehend the speaker's situation and exhibit adequate empathy; (3) Informativeness (**Inf.**): the extent of contextual information present in response; and (4) Continuity (**Con.**): which reply fosters a greater desire in the speaker to sustain the conversation. After shuffling the one hundred randomly selected responses' order in each sample, we assign three annotators who are master's students specializing in NLP to label each pair on a scale of 1 to 5 as in Li et al. (2020), where scores of 1, 3 and 5 represent unacceptable, moderate and excellent performance respectively, while 2 and 4 fall in between.

4.5 Baselines

To verify the effectiveness of our proposed NEC, we selected the following strong models as baselines for comparison:

- **Transformer**(Vaswani et al., 2017), the vanilla Transformer for response generation.
- **Multi-TRS** (Rashkin et al., 2019), a model combining emotion classification for multi-task training with Transformer.
- **MoEL** (Lin et al., 2019), a Transformer-based model with a single encoder and multiple decoders, where each decoder is dedicated to a specific emotion and the outputs of the different decoders are aggregated to generate a response.
- **MIME** (Majumder et al., 2020), another Transformer-based model that mimics the emotion in both negative and positive way to achieve an appropriate balance.
- **EmpDG** (Li et al., 2020), a multi-granularity generation framework introducing interactive discriminators to identify the semantic and emotional accuracy of generated content.
- **CEM** (Sabour et al., 2022), an empathetic response generation model involving affective and cognitive commonsense knowledge.
- **DCKS** (Cai et al., 2023), a SOTA model using BART-based generation with dynamic commonsense selection through emotions.

4.6 Experimental Results

Table 1 presents the results of the automatic evaluation conducted on the EMPATHETICDIALOGUES dataset. The avoidance of cascading errors on emotion recognition has led to significant improvements in all metrics, in comparison with seven strong baselines that require prior emotion recognition. This result provides evidence of the effectiveness of our NEC.

Specifically, the BLEU-n scores of our NEC have increased by 20% in comparison with the SOTA model DCKS, suggesting that our NEC aligned better with real people's responses than previous best models. Additionally, our NEC surpasses the baselines in three metrics Distinct-n,

Model	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	Dist-1↑	Dist-2↑	Sent-Std↓
Transformer	37.62	18.07	8.34	4.57	2.86	0.36	1.35	47.35
Emotion methods								
Multi-TRS	37.50	18.78	8.55	4.70	2.95	0.35	1.27	100.11
MoEL	36.60	18.07	8.30	4.37	2.65	0.59	2.64	40.15
MIME	37.24	18.60	8.39	4.54	2.65	0.47	1.66	216.08
EmpDG	37.43	19.96	9.11	4.74	2.80	0.46	1.99	26.94
Commonsense methods								
CEM	36.33	16.12	7.29	4.06	2.03	0.62	2.39	61.26
DCKS	16.08	21.73	10.62	6.24	4.09	2.19	9.61	36.89
NEC(Ours)	**12.89**	**23.62**	**12.14**	**7.27**	**4.78**	**3.15**	**13.91**	**18.61**
Human	—	—	—	—	—	19.49	43.55	0.71

Table 1: Performance comparison between our NEC and the baselines (cited from (Cai et al., 2023)) where ↑ indicates good performance for large numbers, while ↓ indicates good performance for small numbers.

Models	**Coh.**	**Emp.**	**Inf.**	**Cont.**
EmpDG	2.98	2.87	2.81	2.81
MIME	3.27	3.25	3.09	3.14
MoEL	3.55	3.66	3.42	3.53
Transformer	3.28	3.34	3.25	3.23
Multi-TRS	3.35	3.33	3.3	3.25
CEM	3.87	3.82	3.39	3.24
DCKS	3.95	3.90	4.05	**4.16**
NEC(Ours)	**4.34**	**4.32**	**4.08**	4.12
w/o Neg	4.21	4.11	4.01	4.09

Table 2: Results of human evaluation. Fleiss kappa fo the results is 0.40, indicating a moderate level of consistency among evaluators.

Model	Proportion(%)
Transformer	32.38
Multi-TRS	46.61
MIME	54.58
EmpDG	29.94
MoEL	25.97
CEM	36.82
DCKS	34.41
NEC(Ours)	**20.32**

Table 3: Proportion of top five frequent sentences.

Sent-Std, and PPL, which indicates that our NEC can generate more informative responses at the n-gram and sentence levels. This verifies that the incorporation of the contrastive learning strategies enables the model to avoid being erroneously guided by the emotion recognition task and to better leverage context for empathetic response generation.

In addition, the results of the human evaluation are presented in Table 2, demonstrating that our NEC outperforms all baselines, with no decline in performance observed after the removal of explicit emotion recognition. After removing the negative samples with different emotions (*w/o Neg*), our NEC exhibited a significant decline in empathy performance. This indicates that the negative samples with emotions in the contrastive learning strategies are effective in enhancing the empathic ability of the model.

Furthermore, Table 3 shows the proportion of the top five most frequent sentences among all different sentences generated by our and other models, in which most of them are uninformative generic sentences like "I'm sorry to hear that". This result demonstrates the diversity of responses of our NEC, which can generate more informative responses.

Further analysis is conducted through the sampling and examination of the generated sentences, with the objective of assessing whether a significant proportion of sentences have been rephrased from the top five results manually. Following the consolidation of sentences with similar semantics, it was observed that the proportion of occurrences for the top five and Sent-Std increased to 26.02% (+5.7%) and 27.15 (+8.54), respectively. Notwithstanding this increase, our method continues to outperform amlost all baselines with respect to both metrics. In the future, we intend to integrate BERTScore in order to more accurately calculate semantic similarity and cluster similar sentences, which may further enhance this metric.

4.7 Ablation Study

We conduct an ablation study to elucidate the roles of different modules in our NEC and the results are shown in Table 4. Two main components were designed: (1) Different Commonsense Knowledge: i) Entity commonsense was removed (*w/o Entity*) and dialogue history is utilized as the input of the encoder , and ii) Social commonsense was removed

Model	PPL↓	B-1↑	B-2↑	B-3↑	B-4↑	Dist-1↑	Dist-2↑	Sent-Std↓
NEC(Ours)	12.89	23.62	12.14	7.27	4.78	3.15	13.91	18.61
w/o Entity	12.71	22.56	11.57	6.96	4.61	2.41	10.11	19.31
w/o Social	12.63	23.54	12.15	7.08	4.57	2.11	8.02	19.22
w/o SelfGen	11.91	23.01	11.73	6.84	4.60	2.15	10.64	20.33
w/o HighFreq	11.81	23.42	12.01	7.22	4.69	3.02	12.06	21.12
w/o DiffEmo	12.97	22.91	11.66	6.81	4.62	2.74	11.51	19.66

Table 4: Results of ablation study.

(*w/o Social*) and the corresponding knowledge attention module in the decoder was also removed. (2) Different types of negative examples in contrastive learning: i) the self-generated negative examples (*w/o SelfGen*), ii) high-frequency statement negative examples (*w/o HighFreq*), and iii) negative examples with different emotions (*w/o DiffEmo*) were removed, respectively.

Commonsense knowledge The removal of commonsense knowledge results in a notable decline in the model's performance, particularly in terms of Dist-n and Sent-Std. This is due to the fact that entity knowledge enriches the model with a more comprehensive contextual understanding, extending beyond the confines of the conversation history itself. Furthermore, it effectively transfers knowledge from a commonsense model to the generative model through integration. Moreover, social commonsense guides the direction of the model's responses, ensuring that the model's output aligns with the human response to the context. Upon the removal of this guidance, the model is unable to process this kind of information, despite its comprehensive understanding of the context, aided by other modules.

Negative samples Emotion negative samples are designed as guidance emotion in the training stage, which don't exhibit a significant decrease in Sent-Std. However, Table 2 reveals a substantial decline in empathy of human evaluation. Removing self-generated and high-frequency negative examples results in a noticeable deterioration of the Sent-Std metric. This indicates that the model is inclined to generate more frequent generic sentences without these components.

4.8 Case study

Table 5 presents two cases from EMPATHETICDIALOGUES. In the first example, the speaker emphasizes the breakup with his girlfriend of eight years, indicating a highly pessimistic emotion. However, most models fail to exhibit sufficient empathetic abilities and context relevance. Transformer, MoEL and Multi-TRS perceive this as a matter deserving of a positive response, and they reply in a cheerful manner, lacking expressions of consolation or topic transition. This not only fails to achieve empathetic responses but may potentially lead to more serious consequences. MIME employs relatively generic sentences, with the phrase "*I am sorry to hear that*" constituting 30.70% of all different sentences. EmpDG expresses regret about the speaker's breakup, but does not offer an effective statement that could keep the conversation going. For dialogue models that incorporated commonsense, we found they did not show emotions that were completely opposite to the dialog history. Instead, they displayed passive, uninformative responses. In contrast, our model and DCKS expressed comfort to the speaker in an active way.

In the second example, the speaker expresses confusion about some people not responding or confirming emails and seeks to vent to another person. The majority of models provide responses with less informativeness, offering generic phrases like "*What happened?*" or contextually indifferent phrases like "*I do not blame you*" by EmpDG and CEM. DCKS uses "I'm sorry to hear that" in response to the speaker's expression of oddity, which should correspond to the context of sadness. In contrast, our NEC acknowledges the author's confusion in "*I agree*" and expresses its own perplexity regarding people not responding to emails in " *I do not know why people don't respond to emails*".

4.9 Error Analysis

Uninformative sequences As evidenced by Table 3, NEC has significantly reduced the proportion of high-frequency sentences in comparison with previous work. However, when compared to human, there is still considerable room for improvement. We observe that both our model and previous models often adopt a pattern of one empathy-generic utterance followed by one informational

Context	I there, don't know what to do, jst broke up with my girlfirend, we were 8 years together.
EmpDG	Oh no ! that is a shame !
MIME	I am sorry to hear that. what happened?
MoEL	Oh wow, that must have been very exciting!
Multi-TRS	That sounds like a lot of fun!
Transformer	That sounds like a lot of fun !
CEM	I am sorry to hear that.
DCKS	I am sorry to hear that. i hope everything works out for you!
Ours	Oh no! I'm so sorry to hear that. I hope you find someone better.
Human	Sorry to hear! Do you have any idea about the break up? did you think about it?
Context	I find it odd and annoying when people do not respond to or even acknowledge emails .
EmpDG	I do not blame you.
MIME	I am sorry to hear that. what happened?
MoEL	Oh no. I hate that. i hope they do not have to deal with that.
Multi-TRS	That is terrible. I hope you find them soon.
Transformer	I know what you mean, it is not good.
CEM	I do not blame you.
DCKS	I am sorry to hear that. do you have a problem with that?
Ours	I agree. I also do not know why people don't respond to emails.
Human	Well, sometimes people are busy and do not have the time to check emails.

Table 5: Case study of generated responses of EMPATHETICDIALOGUES.

and empathy-special sentence (first sample in Table 5). In contrast, human responses frequently demonstrate empathy without directly using sentences as "*I'm sorry to hear that.*" to express empathy. For example, in the case of a speaker experiencing a car accident due to the other party's drunkenness, the model's output might be: "*That is horrible! I'm glad you are OK.*" In contrast, a human response could be "*Did you suffer any injuries?*". It expresses concern by directly asking whether the speaker was hurt. This might suggest that a significant proportion of high-frequency sentences generated by models potentially due to their reliance on a predefined pattern. This conversational pattern, compared to human interaction, appears more mechanistic, potentially inducing fatigue among participants and diminishing their engagement and continuity of the dialogue.

5 Conclusion

In this paper, we propose a novel empathetic dialogue generation framework NEC grounded in contrastive learning and multiple knowledge. The framework constructs various negative examples to train the model, mitigating the risk of cascading errors arising from poor emotion recognition. Simultaneously, it constrains the model's output on high-frequency statements, enabling it to produce more informative responses. The experimental results demonstrate the effectiveness of our NEC. In the future, we will consider how to improve the efficiency of the training stage.

Limitations

In our work, we utilize various types of negative samples during training. While this approach does not affect the decoding speed, it results in relatively longer training times in comparison with the original training method. This is a common challenge in many contrastive learning algorithms.

Acknowledgements

The authors would like to thank the three anonymous reviewers for their comments on this paper. This research was supported by the National Natural Science Foundation of China (Nos. 62376181 and 62276177), and Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions.

References

Chenxin An, Jiangtao Feng, Kai Lv, Lingpeng Kong, Xipeng Qiu, and Xuanjing Huang. 2022. Cont: Contrastive neural text generation. In *NeurIPS*, volume 35, pages 2197–2210.

C Daniel Batson, Jim Fultz, and Patricia A Schoenrade. 1987. Distress and empathy: Two qualitatively distinct vicarious emotions with different motivational consequences. *Journal of personality*, 55(1):19–39.

Hua Cai, Xuli Shen, Qing Xu, Weilin Shen, Xiaomei Wang, Weifeng Ge, Xiaoqing Zheng, and Xiangyang Xue. 2023. Improving empathetic dialogue generation by dynamically infusing commonsense knowledge. In *ACL*, pages 7858–7873.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *ICML*, pages 1597–1607.

Shuyang Dai, Guoyin Wang, Sunghyun Park, and Sungjin Lee. 2021. Dialogue response generation via contrastive latent representation learning. In *NLP4ConvAI*, pages 189–197.

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 *EMNLP*, pages 807–819.

Karen E. Gerdes, Elizabeth A. Segal, and Cynthia A. Lietz. 2010. Conceptualising and Measuring Empathy. *The British Journal of Social Work*, 40(7):2326–2343.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: COmmonSense knowledge for eMotion identification in conversations. In *EMNLP*, pages 2470–2481.

John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader. 2021. Declutr: Deep contrastive learning for unsupervised textual representations. In *AAAI*, pages 879–895.

Martin L Hoffman. 2001. *Empathy and moral development: Implications for caring and justice*. Cambridge University Press.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: on symbolic and neural commonsense knowledge graphs. In *AAAI*, volume 35, pages 6384–6392.

David A Kalkstein, David A Bosch, and Tali Kleiman. 2020. The contrast diversity effect: Increasing the diversity of contrast examples increases generalization from a single item. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 46(2):296.

Sevgi Coşkun Keskin. 2014. From what isn't empathy to empathic learning process. *Procedia - Social and Behavioral Sciences*, 116:4932–4938.

Jing Yang Lee, Kong Aik Lee, and Woon Seng Gan. 2022. Improving contextual coherence in variational personalized and empathetic dialogue agents. In *ICASSP*, pages 7052–7056.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and William B Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *NAACL*, pages 110–119.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2020. EmpDG: Multi-resolution interactive empathetic dialogue generation. In *COLING*, pages 4454–4466.

Weizhao Li, Junsheng Kong, Ben Liao, and Yi Cai. 2022. Mitigating contradictions in dialogue based on contrastive learning. In *ACL*, pages 2781–2788.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. MoEL: Mixture of empathetic listeners. In *EMNLP-IJCNLP*, pages 121–132.

Yiren Liu and Halil Kilicoglu. 2023. Commonsense-aware prompting for controllable empathetic dialogue generation. *arXiv preprint arXiv:2302.01441*.

Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Mime: Mimicking emotions for empathetic response generation. In *EMNLP*, pages 8968–8979.

Xiaobing Pang, Yequan Wang, Siqi Fan, Lisi Chen, Shuo Shang, and Peng Han. 2023. Empmff: A multi-factor sequence fusion framework for empathetic response generation. In *WWW*, pages 1754–1764.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *ACL*, pages 5370–5381.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *AAAI*, volume 36, pages 11229–11237.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-LLM: A trainable agent for role-playing. In *EMNLP*, pages 13153–13187.

Tong Shen, Jiawei Zuo, Fan Shi, Jin Zhang, Liqin Jiang, Meng Chen, Zhengchen Zhang, Wei Zhang, Xiaodong He, and Tao Mei. 2021. Vida-man: Visual dialog with digital humans. In *ACM MM*, page 2789–2791.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. In *NeurIPS*, volume 33, pages 16857–16867.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *AAAI*, volume 31, pages 4444–4451.

Carl Strathearn and Dimitra Gkatzia. 2021. Chefbot: A novel framework for the generation of commonsense-enhanced responses for task-based dialogue systems. In *INLG*, pages 46–47.

Yihong Tang, Bo Wang, Miao Fang, Dongming Zhao, Kun Huang, Ruifang He, and Yuexian Hou. 2023. Enhancing personalized dialogue generation with contrastive latent variables: Combining sparse and dense persona. *arXiv preprint arXiv:2305.11482*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: A mixed strategy-aware model integrating comet for emotional support conversation. In *ACL*, pages 308–319.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*, volume 30, pages 5998–6008.

Sixing Wu, Ying Li, Ping Xue, Dawei Zhang, and Zhonghai Wu. 2022. Section-aware commonsense knowledge-grounded dialogue generation with pre-trained language model. In *COLING*, pages 521–531.

Zhuofeng Wu, Sinong Wang, Jiatao Gu, Madian Khabsa, Fei Sun, and Hao Ma. 2020. Clear: Contrastive learning for sentence representation. *arXiv preprint arXiv:2012.15466*.

Yubo Xie and Pearl Pu. 2021. Empathetic dialog generation with fine-grained intents. In *CONLL*, pages 133–147.

Weixiang Zhao, Yanyan Zhao, Xin Lu, and Bing Qin. 2023. Don't lose yourself! empathetic response generation via explicit self-other awareness. In *ACL*, pages 13331–13344.

A Training Example

Table 6 shows a training example of the entity knowledge, the social knowledge and the constructed negative examples for contrastive learning.

Context	*Speaker:*I remember going to the fireworks with my best friend. there was a lot of people, but it only felt like us in the world. *Listener:* Was this a friend you were in love with, or just a best friend? *Speaker:*this was a best friend. i miss her.
gold emotion	sentimental
Entity Knowledge	[ObjectUse]friend:have a good time [ObjectUse]fireworks:have a good time. [AtLocation]friends:the fireworks
Social Knowledge	xReact:Sad; xNeed:Have a best friend. xIntent:Have a friend xWant:Talk to her xEffect:Is missed
Different Emotion	I am sorry to hear that. Did it happen out of the blue? That is great. I would have been scared to death. Was it recent?
Self-generate	I hope you can get better! I'm sorry to hear that.
From batch	I saw that game. They won on penalty kicks! Hello! I am in such a good mood since I got my new home.
High frequency	I am sorry to hear that. What happend?
Human Response	Sorry to hear! Do you have any idea about the break up? did you think about it?

Table 6: A training example.