

# Think Twice: A Human-like Two-stage Conversational Agent for Emotional Response Generation

Yushan Qian  
College of Intelligence and  
Computing, Tianjin University  
Tianjin, China  
yushanqian@tju.edu.cn

Wu Bin  
Quesoar Co. Ltd.  
Tianjin, China  
wubin@quesoar.com

Kun Huang  
Artificial Intelligence Laboratory,  
China Mobile Communication Group  
Tianjin Co., Ltd.  
Tianjin, China  
hknzh@126.com

Bo Wang  
College of Intelligence and  
Computing, Tianjin University  
Tianjin, China  
bo\_wang@tju.edu.cn

Shuo Zhang  
Quesoar Co. Ltd.  
Tianjin, China  
s@quesoar.com

Yuxian Hou  
College of Intelligence and  
Computing, Tianjin University  
Tianjin, China  
yxhou@tju.edu.cn

Shangzhao Ma  
College of Intelligence and  
Computing, Tianjin University  
Tianjin, China  
shangzhaoma@tju.edu.cn

Dongming Zhao  
Artificial Intelligence Laboratory,  
China Mobile Communication Group  
Tianjin Co., Ltd.  
Tianjin, China  
waitman\_840602@163.com

## ABSTRACT

Towards human-like dialogue systems, current emotional dialogue approaches jointly model emotion and semantics with a unified neural network. This strategy tends to generate safe responses due to the mutual restriction between emotion and semantics, and requires the rare large-scale emotion-annotated dialogue corpus. Inspired by the "think twice" behavior in human intelligent dialogue, we propose a two-stage conversational agent for the generation of emotional dialogue. Firstly, a dialogue model trained without the emotion-annotated dialogue corpus generates a prototype response that meets the contextual semantics. Secondly, the first-stage prototype is modified by a controllable emotion refiner with the empathy hypothesis. Experimental results on the DailyDialog and EmpatheticDialogues datasets demonstrate that the proposed conversational agent outperforms the compared models in the emotion generation and maintains the semantic performance in the automatic and human evaluations.

## KEYWORDS

Emotional Dialogue; Dialogue Systems; Human Interaction

### ACM Reference Format:

Yushan Qian, Bo Wang, Shangzhao Ma, Wu Bin, Shuo Zhang, Dongming Zhao, Kun Huang, and Yuxian Hou. 2023. Think Twice: A Human-like Two-stage Conversational Agent for Emotional Response Generation. In *Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, London, United Kingdom, May 29 – June 2, 2023, IFAAMAS, 10 pages.

*Proc. of the 22nd International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2023)*, A. Ricci, W. Yeoh, N. Agmon, B. An (eds.), May 29 – June 2, 2023, London, United Kingdom. © 2023 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). All rights reserved.

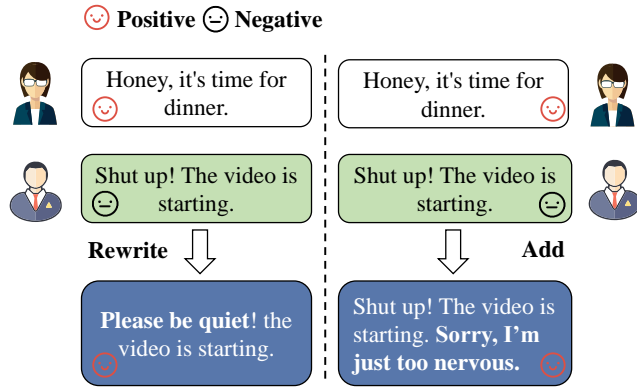
## 1 INTRODUCTION

In the task of open-domain dialogue generation, emotional dialogue aims to generate responses involving the perception and expression of proper emotions. A large number of studies [29, 33, 38] have demonstrated that emotional dialogue can significantly improve users' satisfaction in a human-machine conversation. Moreover, building a dialogue system with human emotions is one of the ultimate goals of artificial intelligence.

Towards emotional dialogue systems, in addition to the early methods of manually compiling rules by professionals [43], existing statistical approaches are mainly based on neural network models [1, 11, 20, 22, 24, 27, 44, 52–54]. With an end-to-end strategy, these neural network models jointly generate the semantics and emotions of the dialogue responses.

However, current end-to-end emotional dialogue models still face several challenges. Firstly, in deep neural networks, the input emotion signals are often weakened through the complex learning process. Secondly, in the joint generation model, the design to enhance emotions often restricts the semantic performance of generated responses (e.g., safe responses). Thirdly, large-scale emotion-annotated dialogue corpora are rare for joint training of semantics and emotions with deep neural networks.

In response to the above challenges, we propose to generate emotional responses with the idea of human intelligent dialogue behavior. When humans respond in a dialogue, the simultaneous processing of emotion and semantics can not ensure satisfying results [7]. The intuitive emotion generated simultaneously with semantics is often arduous to ensure a response in the appropriate emotion. One source of the appropriate emotional response comes from an independent emotion selection after determining the semantics, i.e., thinking twice about appropriate emotions. In this



**Figure 1: The real-life examples of emotion adjustment in the human dialogue. The "think twice" strategy can be observed in human intelligent behavior and effectively improves the quality of emotional responses by rewriting expressions or adding extra information. Bold tokens are the rewritten or added part.**

independent emotion selection, a paramount strategy for humans to determine the appropriate emotion is the empathy strategy, which makes the emotion of the response consistent with that of the context [2, 24, 40]. As visualized in Figure 1, each response firstly has a proper semantic to respond to the context. Then, by recognizing the context’s emotional state, we can adjust the response emotionally and achieve a certain degree of empathy by responding to the partner’s emotion.

Therefore, we design a human-like two-stage conversational agent for emotional response generation. Firstly, a prototype response with proper semantics is generated with a pre-trained model fine-tuned on dialogue corpus without emotion annotation. Then, the contextual emotional state is recognized by a dialogue emotion detector. According to the empathy hypothesis [24], the type of generated emotion is consistent with the contextual emotional state. Finally, the prototype is modified by a controllable emotion refiner to generate a final response that is both semantically relevant and emotionally appropriate.

Specifically, towards effective refining for an emotional response in the second stage, we also refer to two human pragmatic strategies. First, humans express the same information in different ways with different vocabulary choices [37]. Therefore, we involve expected emotional attributes in the response by replacing the original emotional words or phrases instead of constructing a new sentence from scratch, i.e., the "rewriting" strategy. Second, there are also some implicit emotions reflected through the whole sentence instead of specific words. Consequently, we also adjust the emotion by adding extra sentences to the response, i.e., the "adding" strategy.

In summary, our contributions are:

- Inspired by human intelligent dialogue behavior, we propose a human-like two-stage conversational agent for emotional

response generation. To the best of our knowledge, it is the first two-stage model specifically for emotional dialogue.

- The proposed method effectively alleviates two problems of existing emotional dialogue approaches, i.e., weakening the emotion effect during the complex learning process and restricting the semantic generation to meet the emotion demand.
- The proposed two-stage conversational agent reduces the demand for the sizeable emotion-annotated dialogue corpus. The training of the prototype response generator in the first stage only requires general dialogue corpora without emotion annotation, and the controllable emotion dialogue refiner is trained on non-dialogue and non-parallel emotion-annotated corpora.
- The proposed method can be generalized to other existing end-to-end emotion dialogue generation models as post-processing for emotionalization. Even if some sentences have poor emotional expressions, there is no need to retrain the whole model and build new sentences from scratch.

## 2 METHODS

### 2.1 Preliminaries

Formally, in this paper, the dialogue context is alternate utterances of two speakers, defined as  $C = \{U_1, U_2, \dots, U_n\}$ , where  $n$  denotes the number of utterances in a dialogue. The set of context emotions is  $E = \{e_1, e_2, \dots, e_n\}$ , which corresponds to the dialogue context  $C$ . Our goal is to generate the next utterance  $U_{n+1}$ , which is coherent to the context and contains the appropriate emotion.

As shown in Figure 2, our model consists of three parts: the Prototype Utterance Generator  $G$ , the Dialogue Emotion Detector  $D$ , and the Controllable Emotion Refiner  $R$ . The Controllable Emotion Refiner  $R$  has two modules, named "Rewrite" and "Add". The Prototype Utterance Generator  $G$  takes the dialogue context  $C$  as input and generates a prototype response  $U_m$ . The Dialogue Emotion Detector  $D$  takes the dialogue context  $C$  as input and obtains the emotion state set  $E$ , which dynamically determines the response emotion  $e_{n+1}$ . The Controllable Emotion Refiner  $R$  refines  $U_m$  according to  $e_{n+1}$  by rewriting  $U_m$  or adding extra sentences into  $U_m$  with Rewrite Module and Add Module, respectively, and generates the final response  $R_e$ , which is  $U_{n+1}$ .

### 2.2 Prototype Utterance Generator

We use DialoGPT [51] as the Prototype Utterance Generator to generate relevant, diverse, and contextually consistent responses. Large-scale pre-trained language models [5, 39, 51] have extensively promoted the research progress of the open-domain dialogue in recent years. The method of pre-training and fine-tuning can avoid training models from scratch, save computing resources, and achieve excellent results in downstream tasks.

DialoGPT has a 12-to-48 layer Transformer with layer normalization like GPT-2 [39], which is trained on the 147M large-scale dialogue dataset from Reddit. All utterances of the context are spliced into a long sentence with "`<|endoftext |>`" as input. The conditional distribution of the target prototype utterance  $U_m$  is the

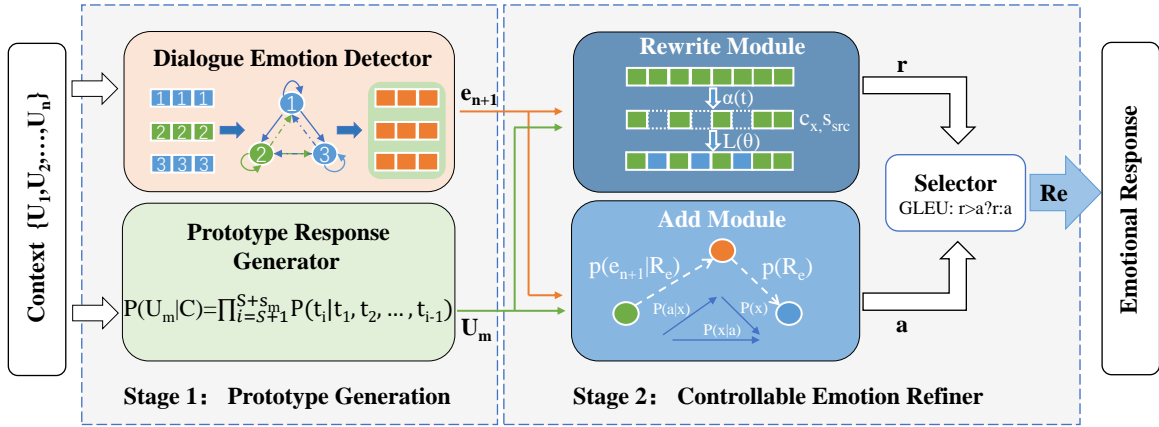


Figure 2: The overall architecture of the proposed two-stage conversational agent. The first stage includes Prototype Utterance Generator and Dialogue Emotion Detector, which generates prototype response  $U_m$  and detects the contextual emotion state as the expected emotion  $e_{n+1}$  of the final response, respectively. The second stage includes Rewrite Module, Add Module, and Selector. Rewrite Module and Add Module refine  $U_m$  according to  $e_{n+1}$ , and the Selector selects the final response from the outputs of Rewrite Module and Add Module based on the GLEU score.

product of a series of conditional probabilities:

$$P(U_m|C) = \prod_{i=S+1}^{S+s_m} P(t_i|t_1, t_2, \dots, t_{i-1}), \quad (1)$$

where  $C = \{U_1, U_2, \dots, U_n\}$ ,  $n$  is the number of utterances,  $U_i = \{t_1, t_2, \dots, t_{s_i}\}$ ,  $t_i$  is each token in the utterance,  $s_i$  is the length of each utterance,  $S = s_1 + s_2 + \dots + s_n$  represents the number of tokens in all utterances,  $s_m$  is the length of  $U_m$ .

We use the maximum mutual information scoring function (MMI) and the top-k sampling [6] to reduce the generation of meaningless responses. The specific implementation is based on tools provided by Hugging Face<sup>1</sup>. We conduct fine-tuning on the training set for 3 epochs with the batch size of 8 and the learning rate of 0.00001. Emotional labels in the training set will not be used. During the decoding process, we use the top-k (k=100) sampling and nucleus sampling (p=0.7) [14].

### 2.3 Dialogue Emotion Detector

As an intuitive hypothesis of empathy, during emotional dialogues between two individuals, the listener usually tends to respond in a way that recognizes the speaker’s feelings [2, 24] and achieves a certain degree of empathy by calling the respondent’s emotions. In the dialogue scene of this work, there are not many turns in the dialogue context ( $<4$ ), so we adopt this empathy hypothesis to determine the emotion of the response according to the emotion state of the context.

To this end, the goal of the Dialogue Emotion Detector  $D$  is to detect emotions in the dialogue context. According to the empathy hypothesis,  $D$  determines the expected emotion of the Controllable Emotion Refiner by the recognized emotion distribution in the dialogue context.

The Dialogue Emotion Detector  $D$  is developed based on DialogueGCN [10], which regards each utterance in the dialogue as

a node of the graph network. There are directed edges between utterances, and the sequence order of utterances determines the direction of each edge. These directed edges can model the emotional impact of what the speaker or the other people has said before. We use Glove embedding and CNN to extract features of utterances and get the embedding of each utterance  $U_i$ , which is the vector of each node. There are  $b$  utterances before each utterance, and there are  $a$  utterances after it. The node of each utterance has edges with  $a + b + 1$  nodes (including itself). The weight  $a_{ij}$  of each edge is decided by the relationship between nodes as follows:

$$\alpha_{ij} = \text{softmax}(U_i^T W_u [U_{i-b}, \dots, U_{i+a}]), \quad (2)$$

for  $j = i - b, \dots, i + a$ .

Further details about DialogueGCN construction are available in [10]. Finally, the embedding from the sequence encoder  $sq$  and the speaker-level encoder  $sp$  are spliced together, and combined with the similarity-based attention mechanism to obtain the final embedding of the utterance node. Then we use a fully connected network to classify multiple emotion categories:

$$\begin{aligned} H &= [h_1, h_2, \dots, h_n], \\ h_i &= \text{softmax}([sq_i, sp_i]^T W H) H^T, \\ e_i &= \text{argmax}(\text{softmax}(\text{FFN}(h_i))). \end{aligned} \quad (3)$$

We use L2 regularization classification cross-entropy loss as the loss function and Adam [18] as the optimizer. We classify emotions in the emotion state set  $S$  into two groups of negative emotions and positive emotions as in [27]. Following [2, 24], we assume that empathetic responses may mimic the user’s emotions to some extent. Therefore, the target emotion  $e_{n+1}$  that we finally pass to

<sup>1</sup><https://huggingface.co/models>

the Controllable Emotion Refiner  $R$  is defined as follows:

$$\begin{aligned} e_{n+1} &= \text{positive,} \\ \text{if } \text{Num}_{\text{pos}}(E) &> \text{Num}_{\text{neg}}(E), \\ \text{otherwise } e_{n+1} &= \text{negative.} \end{aligned} \quad (4)$$

Where  $E = \{e_1, e_2, \dots, e_n\}$  is the set of emotions in each dialogue.  $\text{Num}_{\text{pos}}$  and  $\text{Num}_{\text{neg}}$  represent the number of positive emotions and negative emotions in  $E$ , respectively.

## 2.4 Controllable Emotion Refiner

The Controllable Emotion Refiner  $R$  takes the prototype response  $U_m$  and the target emotion  $e_{n+1}$  as input, and generates the final response  $R_e$ . The goal we need to learn is defined as:

$$P(R_e|U_m, e_{n+1}) \ \& \ \text{Stype}(R_e) = e_{n+1}, \quad (5)$$

where ‘‘Stype’’ represents the emotion type.

The Controllable Emotion Refiner  $R$  consists of two modules, ‘‘Rewrite’’ and ‘‘Add’’. The Rewrite Module transforms the emotion attribute of the  $U_m$  by replacing the original emotion symbols in the sentence with symbols that express the target emotion. The Add module adjusts the emotion type by adding extra sentences. We train the Controllable Emotion Refiner  $R$  by parts.

**Rewrite Module.** The Rewrite Module consists of two parts: the first one is the deletion part, which determines whether each token in the input is an emotion attribute word, learns the emotional part and non-emotional part in the input, and deletes the emotional part. We adopt the attention mechanism of Transformer to extract the attention score as the weight of each token [45]:

$$\alpha(t) = \text{softmax}(QK^T), \text{ for } t \in U_m, \quad (6)$$

where  $Q$  and  $K$  carry the original connotations of query and key vectors in the Transformer.

The second is the generating part, which generates sentences with target emotion attributes. The generating part adopts the Transformer structure, based on the Hugging Face [49]. The input of generating part is the prototype response and the target emotion. The output is a sentence that conforms to the target emotion. Without requiring the parallel corpus, the training goal of generating part is to minimize the following reconstruction loss:

$$L(\theta) = \sum_{x, s_{src} \in D} \log p(x|c_x, s_{src}; \theta), \quad (7)$$

where  $D$  is the training dataset. Given a sentence  $x$ , the Rewrite Module model learns to reconstruct  $y = x$  with  $c_x, s_{src}$ .  $c_x$  is the non-emotional content of  $x$ , and  $s_{src}$  is the original style of the sentence.

**Add Module.** The Add Module is developed based on the work of [4] to change the emotion polarity of the original sentence by adding extra sentences with the target emotion. Using Bayes’ theorem, we can use the model  $p(x)$  and the model  $p(a|x)$  to express the model  $p(x|a)$ :

$$p(x|a) \propto p(a|x)p(x). \quad (8)$$

In order to obtain the required  $p(x|a)$  to generate the sentence based on attribute  $a$ , we already have a language model  $p(x)$  that can generate fluent sentences. Furthermore, we build a classifier to

**Table 1: Two groups of emotions in the DailyDialog dataset according to positivity and negativity.**

Positive	Negative
happiness, surprise, other	anger, disgust, fear, sadness

**Table 2: Two groups of emotions in the EmpatheticDialogues dataset according to positivity and negativity.**

Positive	Negative
confident, joyful, grateful, impressed, proud, excited, trusting, hopeful, faithful, prepared, content, surprised, caring	afraid, angry, annoyed, anticipating, anxious, apprehensive, ashamed, devastated, disappointed, disgusted, embarrassed, furious, guilty, jealous, lonely, nostalgic, sad, sentimental, terrified

determine whether the text  $x$  generated by the language model has  $a$  attribute, that is,  $p(a|x)$ , then  $p(x|a)$  can be obtained.

The process of the Add Module has three steps:

1. First, a forward pass is performed through the language model to compute the likelihood of the desired attribute using an attribute model that predicts  $p(a|x)$ .
2. Second, a backward pass updates the internal latent representations of the language model, using gradients from the attribute model to increase the likelihood of the sentence having the desired attribute.
3. Third, re-sampling to generate a new word according to the obtained new output probability distribution.

To generate more diverse sentences that conform to the language model, two methods are adopted to ensure that the language model of the generated sentence is as close as possible to the original language model: Kullback–Leibler (KL) Divergence and Post-norm Geometric Mean Fusion. About the language model, we use GPT2. Regarding the specific attribute discriminator  $p(a|x)$ , we take the existing non-dialogue emotion-annotated corpus and pre-train a classifier.

**Selector.** A Selector is designed to determine whether the response is from the Rewrite Module or the Add Module is selected as the final output. The Selector uses GLEU [30] as a basis for judging the overall effect of responses, which compares with the prototype response. Sudhakar et al. [45] found that GLEU is more suitable for human score than BLEU score. The Selector selects the final response with a higher GLEU score.

## 3 EXPERIMENTS

In this section, we introduce the datasets, baselines and evaluation metrics. The proposed conversational agent is experimentally compared with baselines and the experimental results are discussed.

### 3.1 Datasets

We used the DailyDialog [23] and EmpatheticDialogues [40] datasets for the experiment. DailyDialog is a multi-round dialogue dataset

for daily chat scenes. There are a total of 12,218 dialogues and 103,607 utterances. The topic and emotion in each utterance are labeled. There are seven types of emotion: anger, disgust, fear, happiness, sadness, surprise, and others. Refer to Majumder et al. [27], we divided the 7 emotion types into two groups containing 3 positive and 4 negative emotions, respectively, as listed in Table 1. EmpatheticDialogues is a widely-used benchmark dataset for empathetic response generation, which is a large-scale multi-turn dataset containing 25k empathetic dialogues between crowdsourcing workers. EmpatheticDialogues also provides an emotion label for each dialogue from 32 available emotions. Following Majumder et al. [27], we divided the 32 emotion types into two groups containing 13 positive and 19 negative emotions, respectively, as listed in Table 2. We focus on positive and negative emotions because the consistency of polarity level emotion is more popular in emotion study and robust in the application. Since our method and baselines are under the same assumption and processed in the same way during evaluation, the results are competitive and convincing.

Considering the limited running space and to unify the number of rounds in each dialogue, we segment the original dialogues into sub dialogues having 4 rounds. Finally, for the DailyDialog dataset, the dialogue numbers of the training / validation / test set are 54,299 / 5,109 / 4,782, respectively. For the EmpatheticDialogues dataset, the dialogue numbers of the training / validation / test set are 18,383 / 2,810 / 3,320, respectively.

### 3.2 Compared Models

To the best of our knowledge, this is an early work in the two-stage generation of emotional dialogue. In view of the empathy hypothesis, we compare our approach with a range of models used in related tasks, including general dialogue, emotional dialogue, and empathetic dialogue.

**Transformer** [48]: The standard Transformer model that is trained to optimize NLL loss.

**Multi-TRS** [40]: A multi-task Transformer model jointly trained by predicting the emotion and generating the response.

**Mojitalk** [54]: An encoder-decoder based CVAE model incorporated with emotion embedding.

**MoEL** [24]: A Transformer-based model employs emotion-specific decoders whose outputs are aggregated and fed to a final decoder to generate the empathetic response.

**MIME** [27]: A Transformer-based model that leverages emotion groups and emotion mimicry, which effectively blends emotions in positive and negative emotion groups and generates the empathetic response.

**EmpDG** [20]: An interactive adversarial model consists of a generator and a discriminator. The discriminator requires user feedback. Besides, the model exploits both the coarse-grained dialogue-level and fine-grained token-level emotions. Referring to Sabour et al. [41], we only apply the empathetic generator to ensure consistent input and output in the test set for a fair comparison with other baselines.

### 3.3 Implementation Details

We use the official codes of all baselines, especially, EmpDG only applies the empathetic generator. We implement all the models

using PyTorch except Mojitalk. All the baselines were trained on a V100 GPU with the batch size of 16 and the early stopping strategy. About the Adam optimizer, we set  $\beta_1 = 0.9$  and  $\beta_2 = 0.98$ . For the emotion detection in the automatic evaluation, emotion pre-training model in Senta is “*ernie\_2.0\_skep\_large\_en*”.

### 3.4 Evaluation Metrics

**3.4.1 Automatic Evaluation.** We apply the following evaluation metrics in the automatic evaluation:

**BLEU:** Word-overlap scores with human responses [32]. We use BLEU-4, which is calculated with an NLG evaluation toolkit<sup>2</sup>.

**Diversity:** Dist-n measures the proportion of unique n-grams in the generated responses [25]. It is commonly used to evaluate whether the dialogue model can generate a diverse response as humans do. Low diversity often means the model tends to generate similar safe responses to different contexts. We refer to the work of [3] to calculate the Dist-1 and Dist-2 metrics.

**Emotion Accuracy (Acc):** The emotion accuracy is defined as the proportion of consistent emotion polarities between generated responses and the ground truth. We use the Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis (SKEP) model [46] proposed by Baidu as the emotion detector during the evaluation<sup>3</sup>. SKEP is a state-of-the-art emotion detector in 14 typical Chinese and English sentiment analysis tasks. We use it to automatically detect the emotion polarity of the responses generated by our proposed conversational agent and baselines.

**3.4.2 Human Evaluation.** We randomly sampled 100 dialogues and generated responses with our proposed conversational agent and baselines. We employed three human annotators to evaluate each response based on four aspects:

**Content(Con):** Whether the response is appropriate for the context in the current dialogue. It is rated on a Likert scale (1: not at all, 3: somewhat, 5: very much).

**Emotion(Emo):** Whether the response is appropriate for the context in the emotion polarity. 1 indicates the response is appropriate, and 0 indicates the response is inappropriate.

**Emotion-intensity(Int):** What the emotion intensity of the response is. 0 represents no emotion, 1 represents slight intensity, and 2 represents strong intensity.

**Fluency(Flu):** Whether the response is readable and understandable. It is rated on a Likert scale (1: not at all, 3: somewhat, 5: very much).

### 3.5 Main Results

Both automatic and human evaluation results are shown in Table 3 and Table 4 on the DailyDialog and EmpatheticDialogues datasets, respectively.

For the performance of emotion generation, it can be observed that the proposed conversational agent outperforms baselines in **Acc** for the automatic evaluation and **Emo** for the human evaluation in two datasets, indicating the outstanding performance of our conversational agent in emotion generation. In addition, our conversational agent also achieves the best and second-best results

<sup>2</sup><https://github.com/Maluuba/nlg-eval>

<sup>3</sup><https://github.com/baidu/Senta>

**Table 3: Automatic and Human evaluations in the DailyDialog dataset (The significant improvement with p-value < 0.05 (t-test). Fleiss’s kappa for Human evaluation is 0.526, indicating "moderate agreement").**

Model	Automatic Evaluation				Human Evaluation			
	BLEU-4	Dist-1	Dist-2	Acc(%)	Con	Emo	Int	Flu
Transformer	0.44	1.10	5.84	54.04	3.54	0.69	0.49	4.60
Multi-TRS	0.58	1.17	6.35	54.29	3.49	0.75	0.52	4.35
Mojitalk	0.58	4.99	22.40	53.99	3.14	0.79	0.61	4.21
MoEL	0.46	1.51	8.39	55.44	3.35	0.77	0.73	<b>4.67</b>
MIME	0.54	0.48	1.79	53.43	3.42	0.78	0.68	4.43
EmpDG	0.57	1.12	5.73	56.02	3.40	0.79	0.77	4.47
Our	<b>0.67</b>	<b>9.71</b>	<b>42.77</b>	<b>57.36</b>	<b>3.82</b>	<b>0.84</b>	<b>0.79</b>	4.45

**Table 4: Automatic and Human evaluations in the EmpatheticDialogues dataset (The significant improvement with p-value < 0.05 (t-test). Fleiss’s kappa for Human evaluation is 0.462, indicating "moderate agreement").**

Model	Automatic Evaluation				Human Evaluation			
	BLEU-4	Dist-1	Dist-2	Acc(%)	Con	Emo	Int	Flu
Transformer	0.35	0.64	2.40	62.29	3.24	0.62	0.84	<b>4.86</b>
Multi-TRS	0.35	0.73	2.77	61.36	3.30	0.76	0.98	4.80
Mojitalk	0.23	<b>6.99</b>	<b>33.52</b>	61.81	2.74	0.76	1.03	4.54
MoEL	0.34	1.15	7.28	62.47	3.44	0.77	1.04	4.82
MIME	0.37	0.89	3.93	61.87	3.32	0.80	<b>1.12</b>	<b>4.86</b>
EmpDG	<b>0.39</b>	0.75	2.50	62.62	3.33	0.81	1.05	4.78
Our	<b>0.39</b>	5.24	22.37	<b>65.42</b>	<b>3.65</b>	<b>0.82</b>	1.11	4.72

**Table 5: Ablation Analysis.**

Model	DailyDialog				EmpatheticDialogues			
	BLEU-4	Dist-1	Dist-2	Acc(%)	BLEU-4	Dist-1	Dist-2	Acc(%)
Our	<b>0.67</b>	<b>9.71</b>	<b>42.77</b>	<b>57.36</b>	<b>0.39</b>	<b>5.24</b>	<b>22.37</b>	65.42
w/o Add	0.59	7.48	33.79	56.04	0.29	4.07	19.63	65.33
w/o Rewrite	0.49	7.93	35.74	55.88	0.35	4.95	20.27	<b>65.78</b>
w/o DED	0.63	8.77	39.27	56.84	—	—	—	—

in **Int** on the DailyDialog and EmpatheticDialogues datasets, respectively, which clearly verifies that the emotional effect of our model is more significant and sufficient compared with the SOTA end-to-end systems. This is because in the process of the two stages, the emotional effect of the response is separately determined and refined in the second stage without being influenced by the semantic generation in the first stage.

For the performance of semantic generation, the proposed conversational agent reaches the highest level in **BLEU-4** and **Con**. In terms of **Dist-1** and **Dist-2**, our conversational agent also scores moderately. These results confirm that the proposed conversational agent significantly improves emotional expression while maintaining appropriate semantics. We also note that our proposed conversational scores are ordinary in **Flu**, which is possibly due to the "adding" strategy increasing the sentence length and affecting the reading difficulty. This is a small limitation given that all models score above 4, which we will explore in the future.

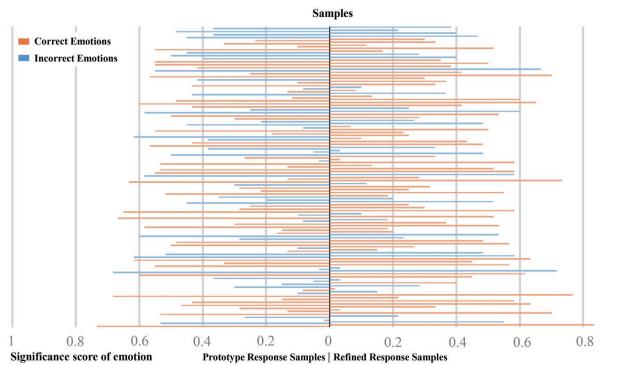
For the performance of compared baselines, MoEL has a low **BLEU-4** score and a high emotion accuracy score, which shows that this existing model loses semantic information when pursuing

emotion features. Furthermore, all baselines except Mojtalk have low scores on diversity metrics, which indicates that there are a large number of safe responses in the generated responses. MoEL, Mojtalk, MIME, and EmpDG also have low scores in **Con** in the DailyDialog dataset, which is lower than Transformer. This may be primarily due to the mutual restriction of semantics and emotions, which reduces the output space.

Since there are 8 metrics (Automatic and Human Evaluation) evaluated, which are less than 30, we choose t-test. Specifically, we use "ttest\_ind" in dcdd"scipy.stats" package to calculate metric values of our method with the metric values of baselines. All p-values are less than 0.05, accept the assumption, which means it is statistically significant.

In Figure 3, for 100 human evaluation samples, we compared the correctness and significance of the emotion of the prototype response and the refined response generated by our proposed conversational agent. The red and blue columns indicate correct (i.e., coherent to the contextual emotion) and incorrect emotions, respectively. The length of the columns indicates the significance of the emotion. We can note that the number of red columns in the refined

responses is more, and the length is longer, which illustrates that the refined responses improve the correctness and significance of the emotion in the prototype response.



**Figure 3: Compare the prototype response and refined response with respect to the correctness and significance of the emotion. The left and right columns indicate prototype and refined responses, respectively. The red and blue columns indicate correct (i.e., coherent to the contextual emotion) and incorrect emotions, respectively. The length of the columns indicates the significance of the emotion.**

### 3.6 Human A/B Test

We conducted the human A/B test, which is shown in Table 7. We randomly sampled 100 response pairs and asked 3 annotators to choose the preferred response based on the dialogue context. A tie is allowed if both are good or bad. The inter-annotator agreement is measured by Fleiss’s kappa. One example needs to be judged six times. Pay about 43.12\$ for every 100 examples. We can observe that responses generated by the proposed conversational agent are preferred by annotators over those generated by other models, which indicates that responses with appropriate emotions and diversity are more attractive to users.

### 3.7 Ablation Analysis

In order to verify the effectiveness of our proposed conversational agent, we also conducted ablation studies. **1) w/o Add:** The Add module is removed in the Controllable Emotion Refiner. We only consider using "rewriting" to refine the prototype response; **2) w/o Rewrite:** The Rewrite module is removed in the Controllable Emotion Refiner, and we consider using "adding" to refine the prototype response; **3) w/o DED:** The Dialogue Emotion Detector is removed and replaced by emotion recognition of a single utterance. This is only conducted in the DailyDialog dataset because the emotion annotation of the EmpatheticDialogues dataset is dialogue-level.

As shown in Table 5, we can observe that removing the Add Module or the Rewrite Module both causes a drop in most metrics. This suggests that combining the "rewriting" and "adding" strategy is beneficial to generating appropriate responses in line with the human language characteristics of both explicit and implicit

expression. However, the Selector of the second stage can be improved, such as selecting which module to use in advance based on the prototype response and emotion. How to refine the prototype response by the "Rewrite" or "Add" module more effectively and reasonably is a problem worth exploring. Moreover, the dialogue emotion detector also plays an important role in emotional response generation, which is superior to concatenating the context into a long sentence or identifying a single utterance.

### 3.8 Case Study

We sampled some generated responses from all models in Table 6. We can observe that responses generated by other baselines have emotional expressions, but the semantics are less appropriate in general. Although responses generated by Transformer are fluent, they often do not conform to the context. In contrast, the response of the proposed conversational agent not only inherits the contextual semantics but also involves rich and appropriate emotions at the same time. For example, the proposed conversational agent coherently transforms "It is a problem. I am not sure how to solve it." to "It is a problem. I am not sure how to solve it. I feel very sorry."

## 4 RELATED WORK

**Dialogue Emotion Recognition.** Different from the emotion recognition of independent sentences, emotions in dialogue should be recognized by the context. Used context typically includes history utterances [36], history emotions [28] and mutual influence of speakers [13]. To model the context, utterances and speakers can be independently [13] or interactively [12] modeled by GRU. [10] uses GCN to solve the problem of context propagation in existing GRU-based methods. Commonsense knowledge [9], psychological knowledge [19], and cognitive theory of emotion [15] are also used to enhance dialogue emotion recognition.

**Emotional Dialogue.** Emotional dialogue aims to generate emotional responses with two main strategies. One strategy is to specify a target emotion in advance [22, 44, 53, 54]. The advantage of this advantage is that the generated emotions are flexible and controllable, and its disadvantage is that large-scale emotion-annotated dialogue corpora are required. The other strategy is to utilize the dialogue context to learn emotions by itself [1], which is close to empathetic dialogue [8, 20, 21, 24, 27, 42] supposing that listeners can infer speakers’ emotions [40]. The advantage of this strategy is that it can utilize the existing large-scale dialogue corpora, and its disadvantage is that the emotions of generated responses are challenging to control. Furthermore, a promising task emotional support dialogue [26] has recently emerged, which provides valuable assistance to people in need [34, 35, 47].

**Controllable Text Generation.** Controllable text generation aims to generate texts with controllable styles. Style is defined as tokens belonging to a specific category or label [37]. Typical processes include training a large-scale conditional generation model from scratch, fine-tuning from a pre-trained language model, and replacing the key n-tuple to adjust the style of the generated sentence [4]. As a kind of style, emotion has good practical significance for its controllable generation. Emotion controlled text generation is to redefine the text to contain the specific emotion without changing the contextual intention [16, 17, 31, 45, 50].

**Table 6: Examples of generated responses by our proposed conversational agent and baselines. The corresponding contexts did not appear in the training set. Bold tokens are the refined part.**

Context	Ground Truth Response	Models	Generated Response
DailyDialog			
No, honey, why didn't you go to cram school last night? Dad, I don't want to talk about it now. Honey, if you don't want to go to cram school, you should tell me the reason why.	I'm sorry, dad. But I would rather stay at school than go to cram school.	Transformer	I do not think I can do that .
		Multi-TRS	Yeah, I know, but I am sure I will.
		Mojitalk	Well, be able to be a good point.
		Moel	I do not know. I want to go to school.
		MIME	That is true. But I do not know.
		EmpDG	I am sorry. I can not wait to get out of the internet.
		<b>Ours</b>	<b>Sorry</b> dad. I can't do that <b>again</b> . I'm so busy.
I could have Suzy stay at my place. I don't want to do that to you, Karen. Why not?	Your own kids would get Suzy's cold. I think it would be better to have a good babysitter here.	Transformer	Yes, I think I would love to.
		Multi-TRS	I am sorry. I must have to give her a girl.
		Mojitalk	Well, I have a cold.
		Moel	I do not know. I would like to. You?
		MIME	I am not sure. I am not sure.
		EmpDG	I am sorry, but I have no idea.
		<b>Ours</b>	I'm sorry. I just can't. <b>I've got nothing to lose. I just want to make the best of what I have, and the best I can.</b>
EmpatheticDialogues			
Why does McDonalds always get my order wrong! I said no lettuce! Report to their management. Sometimes I think the management don't even care!	Just give it a trial.	Transformer	I am sorry to hear that. I hope they win.
		Multi-TRS	I hope you get a better next time.
		Mojitalk	That's a great idea! I'm sure you'll get his family, he has done issues as well.
		Moel	I hope everything works out!
		MIME	I am sorry to hear that. I hope you find a new job.
		EmpDG	I do not know what I do.
		<b>Ours</b>	It is a problem. I am not sure how to solve it. <b>I feel very sorry.</b>
I visited an orphan once and it was such a surreal moment. It made me realise how thankful I should be for a family. That must have been an intense and emotional moment for you. Yes it definitely was. I went with intentions to teach the kid a few things but I ended up learning lots more from him.	Uau, what a surreal moment. I always believe kids have a lot to teach us.	Transformer	I am sure he'll be a good person.
		Multi-TRS	I am glad you are able to find a new one.
		Mojitalk	Me too. The original kids is really much now.
		Moel	I am glad you are on the right?
		MIME	That is great. I am glad you are happy for you.
		EmpDG	I am sure you will do it.
		<b>Ours</b>	I'm glad you had such a <b>great</b> experience.

**Table 7: Results of human A/B test. Fleiss' kappa result for DailyDialog and EmpatheticDialogues is 0.612 and 0.496, indicating "substantial agreement" and "moderate agreement", respectively.**

Models	DailyDialog			EmpatheticDialogues		
	Win	Lose	Tie	Win	Lose	Tie
Our vs Transformer	45.0%	34.7%	20.3%	56.3%	27.3%	16.3%
Our vs Multi-TRS	46.0%	31.0%	23.0%	52.3%	28.0%	19.7%
Our vs Mojitalk	56.3%	26.0%	17.7%	53.3%	26.0%	20.7%
Our vs MoEL	44.7%	31.0%	24.3%	46.7%	19.3%	14.0%
Our vs MIME	49.0%	31.7%	19.3%	51.0%	25.7%	23.3%
Our vs EmpDG	42.0%	32.0%	26.0%	48.0%	31.7%	26.3%

The differences between the proposed conversational agent and existing methods are:

(1) As far as we know, we are the first to study a two-stage emotional response generation paradigm in the field of emotional dialogue specially.

(2) We refine the response with dynamically recognized dialogue context emotions. However, current rewriting methods do not consider the dynamic acquisition of emotions.

## 5 CONCLUSIONS

This paper designed a two-stage conversational agent in the field of emotional dialogue that generates content-related and emotional

responses. The proposed conversational agent generates a semantically coherent prototype response in the first stage and emotionally refines the prototype response in the second stage. Extensive automatic and human evaluations have demonstrated that the proposed conversational agent can generate high-quality emotional responses of appropriate semantics, and reduce the demand for the sizeable emotion-annotated dialogue corpus.

In the future, to improve the proposed conversational agent, we will explore the prediction of explicit and implicit expressions, and the flexible enhancement of other specific features besides the emotion of the dialogue system, such as domain and style adaptation of existing dialogue models.



## ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (62272340, 61876128, 61876129, 62276187, 61976154, 61402323), the Artificial Intelligence for Sustainable Development Goals (AI4SDGs) Research Program, State Key Laboratory of Communication Content Cognition (Grant No.A32003).

## REFERENCES

- [1] Nabihah Asghar, P. Poupart, J. Hoey, Xin Jiang, and Lili Mou. 2018. Affective Neural Response Generation. *ArXiv abs/1709.03968* (2018).
- [2] Laurie Carr, Marco Iacoboni, Marie-Charlotte Dubeau, John C. Mazziotta, and Gian Luigi Lenzi. 2003. Neural mechanisms of empathy in humans: A relay from neural systems for imitation to limbic areas. *Proceedings of the National Academy of Sciences* 100, 9 (2003), 5497–5502. <https://doi.org/10.1073/pnas.0935845100> arXiv:<https://www.pnas.org/content/100/9/5497.full.pdf>
- [3] Chaotao Chen, Jinhua Peng, Fan Wang, Jun Xu, and Hua Wu. 2019. Generating Multiple Diverse Responses with Multi-Mapping and Posterior Mapping Selection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, 4918–4924. <https://doi.org/10.24963/ijcai.2019/683>
- [4] Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2019. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. arXiv:1912.02164 [cs.CL]
- [5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. <https://doi.org/10.18653/v1/N19-1423>
- [6] Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. *arXiv preprint arXiv:1805.04833* (2018).
- [7] Nico H. Frijda. 1987. Emotion, cognitive structure, and action tendency. *Cognition and emotion* 1, 2 (1987), 115–143.
- [8] Jun Gao, Yuhua 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: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 807–819. <https://doi.org/10.18653/v1/2021.findings-emnlp.70>
- [9] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: CommonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics, Online, 2470–2481. <https://doi.org/10.18653/v1/2020.findings-emnlp.224>
- [10] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. 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)*. Association for Computational Linguistics, Hong Kong, China, 154–164. <https://doi.org/10.18653/v1/D19-1015>
- [11] Sayan Ghosh, Mathieu Chollet, Eugene Laksana, Louis-Philippe Morency, and Stefan Scherer. 2017. Affect-LM: A Neural Language Model for Customizable Affective Text Generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 634–642. <https://doi.org/10.18653/v1/P17-1059>
- [12] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2594–2604. <https://doi.org/10.18653/v1/D18-1280>
- [13] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational Memory Network for Emotion Recognition in Dyadic Dialogue Videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2122–2132. <https://doi.org/10.18653/v1/N18-1193>
- [14] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751* (2019).
- [15] Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. DialogueCRN: Contextual Reasoning Networks for Emotion Recognition in Conversations. 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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 7042–7052. <https://doi.org/10.18653/v1/2021.acl-long.547>
- [16] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward Controlled Generation of Text. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 1587–1596. <http://proceedings.mlr.press/v70/hu17e.html>
- [17] N. Keskar, Bryan McCann, L. Varshney, Caiming Xiong, and R. Socher. 2019. CTRL: A Conditional Transformer Language Model for Controllable Generation. *ArXiv abs/1909.05858* (2019).
- [18] Diederik P. Kingma and Jimmy Ba. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980 [cs.LG]
- [19] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, 1204–1214. <https://doi.org/10.18653/v1/2021.findings-emnlp.104>
- [20] 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, Barcelona, Spain (Online), December 8-13, 2020*, Donia Scott, Núria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, 4454–4466. <https://doi.org/10.18653/v1/2020.coling-main.394>
- [21] Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge Bridging for Empathetic Dialogue Generation. (2022).
- [22] Shifeng Li, Shi Feng, Daling Wang, Kaisong Song, Yifei Zhang, and Weichao Wang. 2020. EmoElicitor: An Open Domain Response Generation Model with User Emotional Reaction Awareness. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, Christian Bessiere (Ed.). International Joint Conferences on Artificial Intelligence Organization, 3637–3643. <https://doi.org/10.24963/ijcai.2020/503> Main track.
- [23] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Asian Federation of Natural Language Processing, Taipei, Taiwan, 986–995. <https://www.aclweb.org/anthology/I17-1099>
- [24] Zhaoyang 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)*. Association for Computational Linguistics, Hong Kong, China, 121–132. <https://doi.org/10.18653/v1/D19-1012>
- [25] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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*. Association for Computational Linguistics, Austin, Texas, 2122–2132. <https://doi.org/10.18653/v1/D16-1230>
- [26] 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 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (Eds.). Association for Computational Linguistics, 3469–3483. <https://doi.org/10.18653/v1/2021.acl-long.269>
- [27] Navonil Majumder, Pengfei Hong, Shanshan Peng, Jiankun Lu, Deepanway Ghosal, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. MIMe: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8968–8979. <https://doi.org/10.18653/v1/2020.emnlp-main.721>
- [28] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. DialogueRNN: An Attentive RNN for Emotion Detection in Conversations. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01 (Jul. 2019), 6818–6825. <https://doi.org/10.1609/aaai.v33i01.33016818>
- [29] Bilyana Martinovski and David R. Traum. 2003. Breakdown in Human-machine interaction: the error is the clue. In *ISCA tutorial and research workshop on Error handling in dialogue systems*. [http://people.ict.usc.edu/~traum/Papers/ehsd\\_011.pdf](http://people.ict.usc.edu/~traum/Papers/ehsd_011.pdf)
- [30] Courtney Napoles, Keisuke Sakaguchi, Matt Post, and Joel Tetreault. 2016. GLEU Without Tuning. arXiv:1605.02592 [cs.CL]

- [31] Kartikey Pant, Yash Verma, and R. Mamidi. 2020. SentInc: Incorporating Sentiment Information into Sentiment Transfer Without Parallel Data. *Advances in Information Retrieval* 12036 (2020), 312–319.
- [32] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*. ACL, 311–318. <https://doi.org/10.3115/1073083.1073135>
- [33] Timo Partala and V. Surakka. 2004. The effects of affective interventions in human-computer interaction. *Interact. Comput.* 16 (2004), 295–309.
- [34] Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control Globally, Understand Locally: A Global-to-Local Hierarchical Graph Network for Emotional Support Conversation.
- [35] Wei Peng, Ziyuan Qin, Yue Hu, Yuqiang Xie, and Yunpeng Li. 2023. FADO: Feedback-Aware Double COntrolling Network for Emotional Support Conversation. *Knowledge-Based Systems* (2023), 110340.
- [36] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-Dependent Sentiment Analysis in User-Generated Videos. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 873–883. <https://doi.org/10.18653/v1/P17-1081>
- [37] Shrimai Prabhunoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring Controllable Text Generation Techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Barcelona, Spain (Online), 1–14. <https://doi.org/10.18653/v1/2020.coling-main.1>
- [38] Helmut Prendinger, Junichiro Mori, and Mitsuru Ishizuka. 2005. Using human physiology to evaluate subtle expressivity of a virtual quizmaster in a mathematical game. *International Journal of Human-Computer Studies* 62, 2 (2005), 231–245. <https://doi.org/10.1016/j.ijhcs.2004.11.009> Subtle expressivity for characters and robots.
- [39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).
- [40] 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*. Association for Computational Linguistics, Florence, Italy, 5370–5381. <https://doi.org/10.18653/v1/P19-1534>
- [41] Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: Commonsense-aware Empathetic Response Generation. *CoRR* abs/2109.05739 (2021). [arXiv:2109.05739](https://arxiv.org/abs/2109.05739) <https://arxiv.org/abs/2109.05739>
- [42] Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. CEM: Commonsense-aware Empathetic Response Generation. *CoRR* abs/2109.05739 (2021). [arXiv:2109.05739](https://arxiv.org/abs/2109.05739) <https://arxiv.org/abs/2109.05739>
- [43] M. Skowron. 2009. Affect Listeners: Acquisition of Affective States by Means of Conversational Systems. In *COST 2102 Training School*.
- [44] Zhenqiao Song, Xiaoqing Zheng, Lu Liu, Mu Xu, and Xuanjing Huang. 2019. Generating Responses with a Specific Emotion in Dialog. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 3685–3695. <https://doi.org/10.18653/v1/P19-1359>
- [45] A. Sudhakar, Bhargav Upadhyay, and A. Maheswaran. 2019. Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. *ArXiv* abs/1908.09368 (2019).
- [46] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, and feng wu. 2020. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4067–4076. <https://www.aclweb.org/anthology/2020.acl-main.374>
- [47] 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. *CoRR* abs/2203.13560 (2022). <https://doi.org/10.48550/arXiv.2203.13560>
- [48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. <https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf>
- [49] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45. <https://doi.org/10.18653/v1/2020.emnlp-demos.6>
- [50] Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Mask and Infill: Applying Masked Language Model for Sentiment Transfer. In *IJCAI*.
- [51] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-Scale Generative Pre-training for Conversational Response Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Association for Computational Linguistics, Online, 270–278. <https://doi.org/10.18653/v1/2020.acl-demos.30>
- [52] Peixiang Zhong, D. Wang, and C. Miao. 2019. An Affect-Rich Neural Conversational Model with Biased Attention and Weighted Cross-Entropy Loss. In *AAAI*.
- [53] Hao Zhou, Minlie Huang, T. Zhang, Xiaoyan Zhu, and Bing Liu. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *AAAI*.
- [54] Xianda Zhou and William Yang Wang. 2018. MojiTalk: Generating Emotional Responses at Scale. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 1128–1137. <https://doi.org/10.18653/v1/P18-1104>