CARE: Causality Reasoning for Empathetic Responses by Conditional Graph Generation

Jiashuo WANG, Yi CHENG, Wenjie LI*
Hong Kong Polytechnic University
{csjwang,csycheng,cswjli}@comp.polyu.edu.hk

Abstract

Recent approaches to empathetic response generation incorporate emotion causalities to enhance comprehension of both the user's feelings and experiences. However, these approaches suffer from two critical issues. First, they only consider causalities between the user's emotion and the user's experiences, and ignore those between the user's experiences. Second, they neglect interdependence among causalities and reason them independently. To solve the above problems, we expect to reason all plausible causalities interdependently and simultaneously, given the user's emotion, dialogue history, and future dialogue content. Then, we infuse these causalities into response generation for empathetic responses. Specifically, we design a new model, i.e., the Conditional Variational Graph Auto-Encoder (CVGAE), for the causality reasoning, and adopt a multi-source attention mechanism in the decoder for the causality infusion. We name the whole framework as **CARE**[1], abbreviated for **CA**usality **R**easoning for **E**mpathetic conversation. Experimental results indicate that our method achieves state-of-the-art performance.

1 Introduction

Empathy is the capability to perceive, understand and respond to another individual's feelings, experiences and situation (Paiva et al., 2017; Decety and Jackson, 2004). It is composed of two aspects (Davis, 1983), which are (i) affection, i.e., emotion understanding and appropriate emotional reaction (Hoffman, 2001), and (ii) cognition, i.e., comprehension and reasoning of the other's experiences and situation (Preston and De Waal, 2002).

Earlier work on empathetic response generation merely pays attention to affection (Lin et al., 2019; Majumder et al., 2020; Li et al., 2020a). Consequently, their models lack understanding of the user's experiences, resulting in very weak empathy. Most recent studies begin to consider both affection and cognition by incorporating emotion cause and effect (Wang et al., 2021; Gao et al., 2021; Kim et al., 2021; Sabour et al., 2022). Despite notable improvement, their methods suffer from two critical problems. First, they only consider causalities between the user's emotion and the user's experiences, which are just part of cognition. Causalities between experiences also contribute to the comprehension of experiences. For the case in Figure 1, although *brother* does not cause *impressive* directly, it is the subject causing what impresses the user. Therefore, *brother* should be considered in the causal information for response generation. Second, these methods reason causalities independently and ignore interdependence among these causalities, leading to low-fidelity causality detection. As shown in Figure 1, GEE_{MIME}, one of these methods, fails to reason *killing* → *impressive*, since *killing* itself ordinarily is the cause or

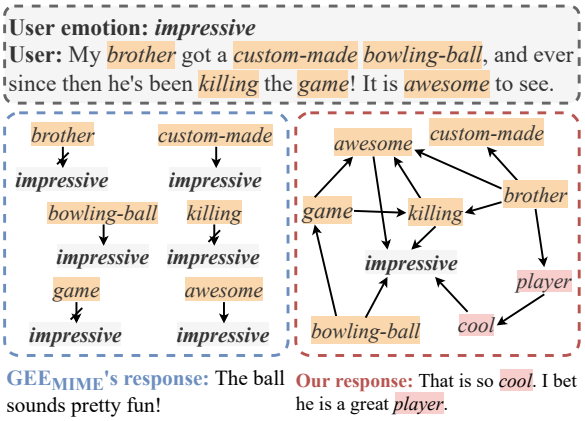

Figure 1: Causality reasoning results of GEE_{MIME} (Kim et al., 2021) and our proposed method in a real case. Arrows indicate relations from cause to effect, while strikeout arrows indicate no causal relations. GEE_{MIME} detects only direct causes and effects of the user's emotion independently, while ours extends the causality scope and reasons causalities interdependently.

*Corresponding author.
[1]The implementation of CARE is publicly available at https://github.com/wangjs9/CARE-master.

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 729–741
December 7-11, 2022

effect of a negative emotion. However, this causality is reasonable when simultaneously considering other causalities including *game* → *killing*, as our proposed method models. Due to the above two problems, these previous methods always misunderstand feelings and experiences of the user, impeding empathetic expression in responses.

To solve these problems, we propose to reason all plausible causalities, i.e., causalities stated explicitly in the dialogue history and probably in the future dialogue, interdependently and simultaneously by formulating the reasoning as a conditional graph generation task. Specifically, we aim to generate a causal graph[2] containing all plausible causalities conditioned on the user's emotion, dialogue history, and predicted future dialogue content. Inspired by the Variational Graph Auto-Encoder (VGAE) (Kipf and Welling, 2016), we design a Conditional Variational Graph Auto-Encoder (CVGAE), which uses latent variables for conditional structure prediction, to accomplish causality reasoning. Accordingly, the model is expected to have a deeper understanding of the user's feelings and experiences. In addition, some feelings and experiences, which are not explicitly stated in dialogue history but contribute to response generation, can be inferred in this process as shown in Figure 1.

In this paper, we propose a novel empathetic response generation model, called **CARE** (**CA**usality **R**easoning for **E**mpathetic conversation). CARE reasons all plausible causalities by CVGAE, and infuses them into response generation by a multi-source attention mechanism in the decoder. In addition, we adopt multi-task learning to integrate causality reasoning and response generation during training. The experimental results on the EMPATHETICDIALOGUES (Rashkin et al., 2019) benchmark suggest that our method improves the model's understanding of user's feelings and experiences, and **CARE** achieves state-of-the-art performance on empathetic response generation.

Our main contributions are three-fold:

1). We propose to reason all plausible causalities in empathetic conversation interdependently and simultaneously for a deep understanding of the user's feelings and experiences.

2). We turn causality reasoning into a conditional graph generation task, and introduce CVGAE, which uses latent variables for conditional structure prediction, to achieve the reasoning.

3). We design CARE, which augments empathetic response generation with causality reasoning, and prove its outstanding performance on the EMPATHETICDIALOGUES benchmark.

[2]Each node is a word to represent the user's feelings and experiences, and each edge indicates a causal relationship between two nodes.

2 Related Work

Since empathy is a critical character for social chatting systems (Sharma et al., 2020; Pérez-Rosas et al., 2017), many studies have contributed to empathetic response generation. Earlier work mainly focuses on the affective aspect of empathy. MoEL (Lin et al., 2019) adopts a mixture of experts architecture to combine outputs from different decoders, each of which represents one emotion. Based on the idea of emotion mixture, MIME (Majumder et al., 2020) takes emotion polarity (positive or negative) into account. Moreover, it uses emotion stochastic sampling and emotion mimicry to generate empathetic responses. Li et al. (2020a) propose to capture nuances of emotion at the token-level for decoding. Moreover, an adversarial learning framework is leveraged to involve user feedback.

Having realized that ignorance of cognition impedes empathy in conversation, some recent methods involve both affection and cognition by incorporating emotion causes and effects. Wang et al. (2021) incorporate emotion causes into empathetic response generation by multi-hop reasoning from emotion causes to emotion states. Gao et al. (2021) identify emotion causes from dialogue context, and use gates at the decoder to control the involvement of these emotion causes in the response generation. Kim et al. (2021) emphasize emotion causes in dialogue context by a rational speech act framework. These three methods identify emotion causes via a classifier, which detects whether there is a causal relationship between a conversation fragment and an emotion statement or word each time. CEM (Sabour et al., 2022) uses COMET, an if-then commonsense generator, to generate causes and effects of user experiences, and refines dialogue context with them for response generation. However, all these methods obtain causalities independently.

3 Preliminary

3.1 Transformer-based Response Generation

The response generation model is built upon the vanilla transformers (Vaswani et al., 2017), which

generates the response R given dialogue context C as input in an encoder-decoder manner. The encoder encodes the dialogue context and generates the context hidden state. That is:

$$E_{out} = \text{TRS}_{\text{enc}}(C), \tag{1}$$

$E_{out} \in \mathbb{R}^{|C| \times d}$, where d is the hidden size. The decoder takes the right shifted response as input and generates the response. Typically, the whole decoder includes L_{dec} decoder layers, each consisting of three sub-layers. The first one, i.e., the self-attention sub-layer, computes a representation of the input sequence:

$$\begin{aligned} \hat{H} &= \text{MultiHead}(H_{in}, H_{in}, H_{in}), \\ H_{out}^{(self)} &= \text{LayerNorm}(\hat{H} + H_{in}), \end{aligned} \tag{2}$$

where H_{in} is the embedding right shifted response for the first decoder layer, and is output of the $(l-1)$-th decoder layer for the l-th decoder layer. Then the decoder attends to the dialogue context by a cross-attention sub-layer:

$$\begin{aligned} H_{in} &= H_{out}^{(self)}, \\ \hat{H} &= \text{MultiHead}(H_{in}, E_{out}, E_{out}), \\ H_{out}^{(cross)} &= \text{LayerNorm}(\hat{H} + H_{in}). \end{aligned} \tag{3}$$

The output of the l-th decoder layer is obtained by the feed-forward sub-layer:

$$\begin{aligned} H_{in} &= H_{out}^{(cross)}, \\ H_{out}^{(ffn)} &= \text{LayerNorm}(\text{FFN}(H_{in}) + H_{in}). \end{aligned} \tag{4}$$

Finally, we apply linear transformation and a softmax operation on the output of the L_{dec} decoder layer to predict token probability distribution at each token position t:

$$P_t = softmax(H_{out,t}^{L} W_o + b_o), \tag{5}$$

where $H_{out,t}^{L}$ is the final output for the t-th token; $W_o \in \mathbb{R}^{d \times d_{vocab}}$ and $b_o \in \mathbb{R}^{d_{vocab}}$ are parameters, and d_{vocab} is the vocabulary size.

3.2 Variational Graph Auto-Encoder

Our proposed causality reasoning module, i.e., CVGAE, is based on VGAE (Kipf and Welling, 2016). Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with its adjacency matrix $\mathbf{A}$, VGAE generates graph latent variables by an inference model, and reconstructs the adjacency matrix by a generative model.

Inference Model The inference model encodes $\mathcal{G}$, and generates graph latent variables $\mathbf{Z} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{|\mathcal{V}|}\}$ by a recognition net $q(\mathbf{Z}|\mathcal{V}, \mathbf{A})$. Each graph latent variable z_i is obtained by:

$$\begin{aligned} & q(\mathbf{z}_i|\mathcal{V}, \mathbf{A}) = \mathcal{N}(\mathbf{z}_i|\mu_i, \sigma_i^2), \\ \text{with} \quad & \mu = \text{GCNLayer}_{\mu}(\mathcal{H}_{\mathcal{V}}, \mathbf{A}), \\ \text{and} \quad & \log\sigma = \text{GCNLayer}_{\sigma}(\mathcal{H}_{\mathcal{V}}, \mathbf{A}). \end{aligned} \tag{6}$$

Here, $\mathcal{N}$ is a sampling function, which follows the Gaussian distribution. μ is the matrix of the mean vectors μ_i; $\log\sigma$ is the matrix of log-variance vectors $\log\sigma_i$. In particular, $\mathcal{H}_{\mathcal{V}}$ is a shared hidden state obtained by:

$$\mathcal{H}_{\mathcal{V}} = \text{GCNLayer}_{\text{h}}(\mathcal{V}, \mathbf{A}). \tag{7}$$

Generative Model The generative model reconstructs the adjacency matrix by an inner product between latent variables:

$$\begin{aligned} & p(\hat{\mathbf{A}}|\mathbf{Z}) = \prod_{i=1}^{|\mathcal{V}|} \prod_{j=1}^{|\mathcal{V}|} p(\hat{\mathbf{A}}_{ij}|\mathbf{z}_i, \mathbf{z}_j), \\ \text{with} \quad & p(\hat{\mathbf{A}}_{ij} = 1|\mathbf{z}_i \mathbf{z}_j) = \text{sigmoid}(\mathbf{z}_i^{\top} \mathbf{z}_j). \end{aligned} \tag{8}$$

Inference Stage At the inference stage, adjacency matrix $\mathbf{A}$ is unavailable. Therefore, we replace $q(\mathbf{Z}|\mathcal{V}, \mathbf{A})$ with a prior net $p(\mathbf{Z})$, which is parameterized by a Gaussian distribution: $p(\mathbf{z}_i) = \mathcal{N}(\mathbf{z}_i|0, 1)$, to infer $\mathbf{Z}$. Then, we use the same generative model to generate the adjacency matrix.

Objective VGAE is optimized by maximizing:

$$\begin{aligned} \mathcal{L} = & \mathbb{E}_{q(\mathbf{Z}|\mathcal{V}, \mathbf{A})}[\log p(\hat{\mathbf{A}}|\mathbf{Z})] \\ & - \text{KL}[q(\mathbf{Z}|\mathcal{V}, \mathbf{A})||p(\mathbf{Z})], \end{aligned} \tag{9}$$

where $KL[q(\cdot)|p(\cdot)]$ is the Kullback-Leibler divergence between $q(\cdot)$ and $p(\cdot)$.

4 Method

Figure 2 presents an overview structure of our proposed model CARE. It first reasons all plausible causalities interdependently by generating a causal graph. Specifically, we use CVGAE to generate this graph under the condition of the user's emotion, dialogue history, and predicted future dialogue content. Notably, CVGAE works differently at the training and inference stages: it reconstructs a posterior causal graph (by *R-Net*$_G$) with this posterior causal graph as input during training, while generates a posterior causal graph (by *P-Net*$_G$) with a

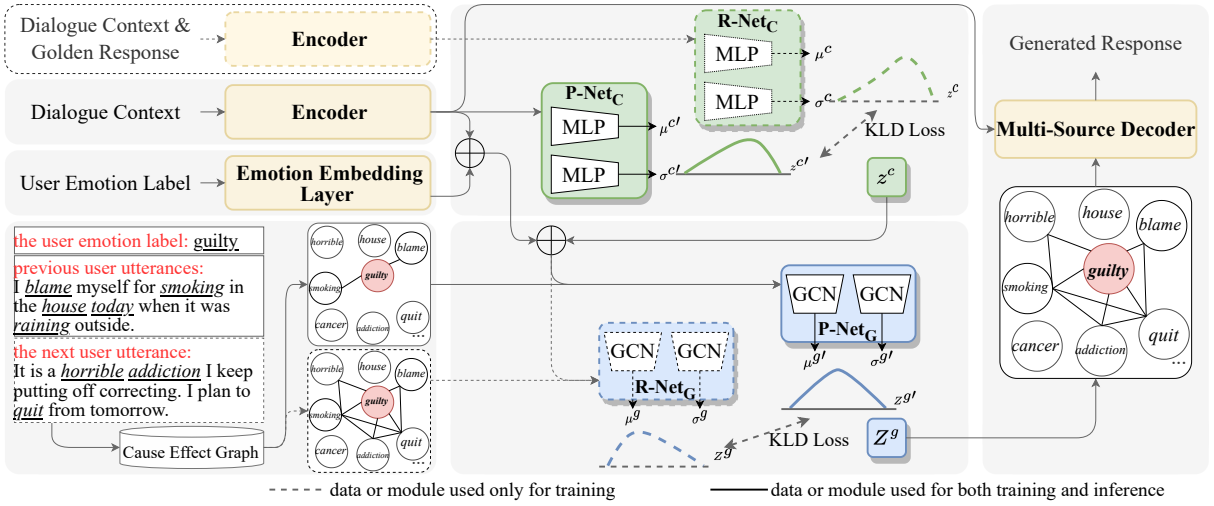

Figure 2: The overview of our proposed framework. The solid lines represent modules or data used for both posterior and prior computation, while the dot lines represent modules or data used only for posterior computation.

prior causal graph as input during inference. The prior causal graph contains causalities explicitly mentioned in previous user utterances, while the posterior one contains additional causalities in the next user utterance. Then, CARE infuses causalities in the reasoned causal graph into response generation by multi-source attention at the decoder.

4.1 Graph Construction

As mentioned, we need a prior causal graph $\mathcal{G}_{prior} = (\mathcal{V}, \mathcal{E}_{prior})$ and a posterior causal graph $\mathcal{G}_{post} = (\mathcal{V}, \mathcal{E}_{post})$, for causality reasoning at inference and training stage, respectively. We construct them with the assistance of a causal knowledge graph, i.e., Cause Effect Graph (CEG) (Li et al., 2020b). These two graphs share the same node set, which theoretically contains all nodes in CEG. However, for effectiveness, we only consider those among a certain set of nodes $\mathcal{V}$, which contains emotion label word, words appearing in previous user utterances, and one-hop neighbors of above two kinds of words. The edge sets of these two graphs are different. $\mathcal{E}_{prior}$ contains causal relationships in previous user utterances, while $\mathcal{E}_{post}$ also contains those in the next utterances. In specific, we collect $\mathcal{E}_{prior}$ and $\mathcal{E}_{post}$ according to following rules. For any couple nodes in $\mathcal{V}$ having a relationship in CEG, if both nodes are covered by the user emotion label word and words in previous user utterances, we add the relationship into $\mathcal{E}_{prior}$; if both nodes are covered by the user emotion label word and words in previous and next user utterances, we add the relationship into $\mathcal{E}_{post}$.

4.2 Conditional Variational Graph Auto-Encoder (CVGAE)

We design a novel structure CVGAE to generate a (posterior) causal graph for causality reasoning. As an extension of VGAE, CVGAE works in a similar manner (§ 3.2). In particular, it generates graphs latent variables for graph reconstruction under some conditions, including a context condition, an emotion condition, and a context latent variable.

Context and Emotion Conditions The context condition is expected to provide information of dialogue context C, thus it is derived from the encoder output. Following (Wang and Wan, 2019), we use multi-head attention to perform it. That is:

$$\mathbf{c}_{\text{ctx}} = \text{MultiHead}(v_{\text{rand}}, E_{out}, E_{out}), \quad (10)$$

where $E_{out} \in \mathbb{R}^{|C| \times d}$ is the encoder output computed by $\text{TRS}_{\text{enc}}(C)$ in Equation (1); $v_{\text{rand}} \in \mathbb{R}^{1 \times d}$ is a randomly initialized vector and is regarded as a single query for multi-head attention.

The emotion condition is expected to provide information of the user emotion e. Accordingly, we define the emotion embedding $E^{emo} \in \mathbb{R}^d$ which converts an emotion label into embeddings. The emotion condition is formulated as:

$$\mathbf{c}_{\text{emo}} = E^{emo}(e). \quad (11)$$

Context Latent Variable We use a context latent variable z_c to provide information from the future dialogue. This variable is generated by a contextual recognition net (*R-Net*$_C$ in Figure 2) with dialogue

context C and the golden response R as input:

$$q_c(\mathbf{z}^{\mathrm{c}}|C,R)=\mathcal{N}(\mathbf{z}^{\mathrm{c}}|\mu^c,\sigma^{c2}). \quad (12)$$

Here $\mu^c=\mathrm{MLP}_\mu(\mathbf{c}_{\mathrm{lant}})$ is the mean vector and $\log\sigma_c=\mathrm{MLP}_\sigma(\mathbf{c}_{\mathrm{lant}})$ is the log-variance vector, where $\mathbf{c}_{\mathrm{lant}}$ is accessed similar to Equation (10):

$$E_{rep}=\mathrm{TRS}_{\mathrm{enc}}(C\oplus R), \quad (13)$$
$$\mathbf{c}_{\mathrm{lant}}=\mathrm{MultiHead}(v_{\mathrm{rand}},E_{rep},E_{rep}). \quad (14)$$

Graph Latent Variables We generate graph latent variables $\mathbf{Z}^g$ by a recognition net (*R-Net*$_G$ in Figure 2): $q_g(\mathbf{Z}^g|\mathcal{V},\mathbf{A}_{post},c_{\mathrm{cond}})$, where $\mathbf{A}_{post}$ is the adjacency matrix of $\mathcal{G}_{post}$. This process is similar to that of VGAE, i.e., Equations (6) and (7).

$$\begin{aligned} q(\mathbf{z}_i^g|\mathcal{V},\mathbf{A}_{post},c_{\mathrm{cond}})&=\mathcal{N}(\mathbf{z}_i^g|\mu_i^g,\sigma^{\mathbf{g}2}_i),\\ \text{with}\quad \mu^g&=\mathrm{GCNLayer}_\mu(\mathcal{H}_\mathcal{V},\mathbf{A}_{post}),\\ \text{and}\quad \log\sigma^g&=\mathrm{GCNLayer}_\sigma(\mathcal{H}_\mathcal{V},\mathbf{A}_{post}). \end{aligned} \quad (15)$$

The shared hidden state $\mathcal{H}_\mathcal{V}$ is generated with attention to the concatenation of $\mathbf{c}_{\mathrm{ctx}}$, $\mathbf{c}_{\mathrm{emo}}$, and $\mathbf{z}^{\mathrm{c}}$:

$$\begin{aligned} c_{\mathrm{cond}}&=\mathbf{c}_{\mathrm{ctx}}\oplus\mathbf{c}_{\mathrm{emo}}\oplus\mathbf{z}^{\mathrm{c}},\\ \hat{\mathcal{H}}_\mathcal{V}&=\mathrm{GCNLayer}_{\mathrm{h}}(\mathcal{V},\mathbf{A}_{post}),\\ \mathcal{H}_\mathcal{V}&=\mathrm{MultiHead}(\hat{\mathcal{H}}_\mathcal{V},c_{\mathrm{cond}},c_{\mathrm{cond}}). \end{aligned} \quad (16)$$

Causal Relation Generation With graph latent variables $\mathbf{Z}^g$, we reconstruct the posterior causal graph, i.e., the matrix adjacency $\hat{\mathbf{A}}$ by Equation (8). Then we select top-k relationships from the reconstructed graph according to their probability, denoted as $\mathcal{R}=(r1,\ldots,r_k)$, where r_i is the sum of the head and tail node embeddings.

Inference Stage During inference, R (the golden response) and $\mathbf{A}_{post}$ are unavailable, thus we use a prior net $p_g(\mathbf{Z}^{g\prime}|\mathcal{V},\mathbf{A}_{prior},c'_{\mathrm{cond}})$ (*P-Net*$_G$ in Figure 2) to approach $q_g(\mathbf{Z}^g)$, i.e., Equations (15) and (16). $\mathbf{A}_{prior}$ is $\mathcal{G}_{prior}$'s adjacency matrix, and $c'_{\mathrm{cond}}=\mathbf{c}_{\mathrm{ctx}}\oplus\mathbf{c}_{\mathrm{emo}}\oplus\mathbf{z}^{\mathrm{c}\prime}$, where $\mathbf{z}^{\mathrm{c}\prime}$ is obtained by a contextual prior net (*P-Net*$_C$ in Figure 2):

$$p_c(\mathbf{z}^{\mathrm{c}\prime}|C)=\mathcal{N}(\mathbf{z}^{\mathrm{c}\prime}|\mu^{c\prime},\sigma^{c\prime 2}), \quad (17)$$

with $\mu^{c\prime}=\mathrm{MLP}_{\mu'}(\mathbf{c}_{\mathrm{ctx}})$, $\log\sigma^{c\prime}=\mathrm{MLP}_{\sigma'}(\mathbf{c}_{\mathrm{ctx}})$.

4.3 Graph-Infused Response Generation

To infuse the reasoned $\mathcal{R}$ into generation, we enable the decoder to attend to both dialogue context and the causal graph (*Multi-Source Decoder* in Figure 2). In particular, we slightly modify the cross-attention sub-layer of the original decoder, i.e., Equation (3), with our multi-source attention mechanism. Therefore, the output after this modified sub-layer is computed by:

$$\begin{aligned} \hat{H}^C&=\mathrm{MultiHead}(H_{in}^{(cross)},E_{out},E_{out}),\\ \hat{H}^\mathcal{R}&=\mathrm{MultiHead}(H_{in}^{(cross)},\mathcal{R},\mathcal{R}),\\ \hat{H}&=(\hat{H}^C\oplus\hat{H}^\mathcal{R})W_{multi},\\ H_{out}^{(cross)}&=\mathrm{LayerNorm}(\hat{H}+H_{in}^{(cross)}), \end{aligned} \quad (18)$$

where $E_{out}\in\mathbb{R}^{|C|\times d}$ is the encoder output, $W_{multi}\in\mathbb{R}^{2d\times d}$ is a group of linear transformation parameters, and H_{in} is the output of the self-attention sub-layer of the decoder computed by Equation (2). Notably, the reset of the original decoder, i.e, Equations (2), (4) and (5), remains the same. In this way, we generate the final response.

4.4 Training Objective

We optimize the model with multi-task learning to further integrate the causality reasoning and the graph-infused response generation. For the causality reasoning, we consider graph reconstruction accuracy and similarity between posterior and prior distribution. Similar to Equation (9), the corresponding loss can be calculated by:

$$\begin{aligned} \mathcal{L}_r=&\mathbb{E}_{q_g(\mathbf{Z}^g|\mathcal{V},\mathbf{A}_{post},c_{\mathrm{cond}})}[\log p(\hat{\mathbf{A}}|\mathbf{Z}^g)]\\ &-\mathrm{KL}[q_g(\mathbf{Z}^g)||p_g(\mathbf{Z}^{g\prime})]\\ &-\mathrm{KL}[q_c(\mathbf{z}^{\mathrm{c}})||p_c(\mathbf{z}^{\mathrm{c}\prime})]. \end{aligned} \quad (19)$$

The response generation loss is calculated by:

$$\mathcal{L}_g=\prod_{t=1}^{|R|}P_t, \quad (20)$$

where P_t is obtained by Equation (5). Finally, we train CARE by maximizing $(\mathcal{L}_r+\mathcal{L}_g)$.

5 Experiments

5.1 Dataset

We conduct our experiments on EMPATHETICDIALOGUES[3] (Rashkin et al., 2019). It contains 25k crowdsourced one-on-one conversations, each of which is developed based on a particular emotion. There are 32 emotion categories distributed in a balanced way. Following its original division, we adopt approximately 80%, 10%, and 10% of the dataset for training, validation, and testing.

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

5.2 Comparison Models

We select seven models for comparison according to some special considerations. Three models that merely consider the affective aspect of the empathy are selected. They are:

MoEL [4] (Lin et al., 2019): This model leverages a mixture of expert architecture to combine outputs from several decoders, each of which pays attention to a unique emotion type.

MIME [5] (Majumder et al., 2020): Based on MoEL's idea of emotion mixture, this model takes emotion polarity into account. Moreover, it considers emotion mimicry during generation.

EmpDG [6] (Li et al., 2020a): This model detects nuanced emotion at word-level as a part of decoder inputs, and uses adversarial learning framework to involve user's feedback.

In addition, four models that considers both the affection and cognition of empathy are selected:

KEMP [7] (Li et al., 2022) This model leverages external commonsense knowledge and emotional lexicon to understand and express emotion for empathetic response generation.

CEM [8] (Sabour et al., 2022) This model generates causes and effects of the user's latest mentioned experiences, and uses them to refine the context encoding for a better understanding of the user's situations and feelings.

RecEC$_{\text{soft}}$ [9] (Gao et al., 2021): This model pays more attention to emotion causes, detected from dialogue context, at word-level by a soft gated attention mechanism in the decoder.

GEE$_{\text{MIME}}$ [10] (Kim et al., 2021): This model uses a rational speech act framework to update the response generated by MIME to obtain the final response that focuses more on the emotion cause words in dialogue context.

All above models, as well as ours, are built upon transformer backbone for a fair comparison.

[4] https://github.com/HLTCHKUST/MoEL
[5] https://github.com/declare-lab/MIME
[6] https://github.com/qtli/EmpDG
[7] https://github.com/qtli/KEMP
[8] https://github.com/Sahandfer/CEM
[9] https://github.com/A-Rain/EmpDialogue_RecEC
[10] https://github.com/skywalker023/focused-empathy

5.3 Implementation Details

Our Model: We implemented our model using PyTorch[11], and trained it on a GPU of Nvidia GeForce RTX 3090. The token embeddings are initialized with 300-dimensional pre-trained Glove vectors (Pennington et al., 2014), and shared between between the encoder, the CVGAE model, and the decoder. The hidden size d is set as 300. The number of node number $|\mathcal{V}|$ is 800, and the number of selected relationships k is 512 (0.16%). Both the encoder layer number and the decoder layer number are 2. The batch size is set as 16. When training the model, we use Adam optimizer (Kingma and Ba, 2015) and vary the learning rate following Vaswani et al. (2017).

Comparison Models: We implement GEE$_{\text{MIME}}$ under its official instructions, since only testing codes and instructions are provided by the authors. For the rest of the comparison models, we utilize their official codes released on GitHub.

5.4 Automatic Evaluation

Metrics: Three kinds of metrics are applied for automatic evaluation: (1) Perplexity (**PPL**), which measures the model's confidence in the response generation. (2) BLEU (Papineni et al., 2002), which estimates the matching between n-grams of the generated response and those of the golden response. We adopt **BLEU-3** and **BLEU-4**. (3) BERTScore (Zhang et al., 2020), which computes the similarity for each token in the generated response with that in the golden response. We use its matching precision, recall and F1 score (**P$_{\text{BERT}}$**, **R$_{\text{BERT}}$**, and **F$_{\text{BERT}}$**). For perplexity, a lower score indicates a better performance; while, for the rest metrics, higher scores indicate better performances.

Annotation Statistics: Table 1 presents the automatic evaluation results, and the highest score in terms of each metric is in bold. For each model, we repeat five runs with different seeds, and compute the average values and standard deviations. In addition, values that are statistically significant with $p < 0.05$ are marked with $*$.

Results: According to Table 1, our proposed model CARE outperforms the other models in terms of all metrics. The lowest perplexity score suggests that our proposed architecture is more confident in its generated responses than other models.

[11] https://pytorch.org/

Model		PPL	BLEU-3	BLEU-4	P_{BERT}	R_{BERT}	F_{BERT}
Affection	MoEL	$36.87_{\pm 0.97}$	$4.53_{\pm 0.53}$	$2.80_{\pm 0.32}$	$.499_{\pm .008}$	$.467_{\pm .007}$	$.480_{\pm .006}$
	MIME	$37.88_{\pm 0.49}$	$4.48_{\pm 0.15}$	$2.71_{\pm 0.09}$	$.490_{\pm .004}$	$.466_{\pm .002}$	$.475_{\pm .002}$
	EmpDG	$55.64_{\pm 3.78}$	$3.64_{\pm 0.38}$	$1.99_{\pm 0.22}$	$.475_{\pm .007}$	$.458_{\pm .008}$	$.465_{\pm .004}$
Affection+Cognition	KEMP	$36.59_{\pm 0.45}$	$4.13_{\pm 0.29}$	$2.43_{\pm 0.15}$	$.484_{\pm .005}$	$.460_{\pm .004}$	$.470_{\pm .005}$
	CEM	$36.70_{\pm 0.44}$	$3.55_{\pm 0.42}$	$2.24_{\pm 0.24}$	$.498_{\pm .001}$	$.461_{\pm .006}$	$.477_{\pm .004}$
	$RecEC_{soft}$	$149.3_{\pm 15.9}$	$3.02_{\pm 0.15}$	$1.62_{\pm 0.12}$	$.491_{\pm .004}$	$.461_{\pm .002}$	$.473_{\pm .002}$
	GEE_{MIME}	-	$2.76_{\pm 0.18}$	$1.50_{\pm 0.14}$	$.472_{\pm .002}$	$.443_{\pm .002}$	$.456_{\pm .001}$
	CARE	$\mathbf{32.84}^{*}_{\pm 0.23}$	$\mathbf{4.88}^{*}_{\pm 0.13}$	$\mathbf{2.95}^{*}_{\pm 0.06}$	$\mathbf{.501}_{\pm .004}$	$\mathbf{.475}^{*}_{\pm .002}$	$\mathbf{.486}^{*}_{\pm .003}$

Table 1: Automatic evaluation results in terms of *PPL*, *BLEU* and *BERTScore*. For each method, we repeat five runs with different seeds. We display the average values of the results along with the standard deviations. The values marked with $*$ mean the results are statistically significant with $p < 0.05$. The highest score in terms of each metric is in bold. The full automatic evaluation results can be found in Appendix A.1.

Model		Emp.	Rel.	Flu.
Affection	MoEL	2.73	2.63	4.82
	MIME	2.30	2.24	4.88
	EmpDG	2.31	2.27	4.52
Affection +Cognition	KEMP	2.26	2.18	4.81
	CEM	2.77	2.70	**4.93**
	$RecEC_{soft}$	2.16	2.21	4.74
	GEE_{MIME}	1.75	1.75	4.78
	CARE	**2.83**	**2.79**	4.86

Table 2: Results of human ratings in terms of *Empathy*, *Relevance* and *Fluenct* on a $5-$point likert scale, where 5 is the best. The highest scores are in bold. The fleiss's kappa is 0.41 indicating a moderate level of agreement.

The table does not present the perplexity score of GEE_{MIME}. This is because its generated token probability distribution depends on the mediate results of MIME and its emotion cause detector, and therefore PPL is less relevant to its core structure, i.e., rational speech act framework. Highest BLEU and BERTScore scores indicate that our approach can generate more human-like responses by incorporating causality reasoning. Especially, all the above advantages are significant and stable, evident in high degrees of statistical significance and small standard deviations, respectively.

5.5 Human Ratings

Metrics: Although the automatic evaluation has provided useful information about models' performances, it cannot capture some features, such as empathy expression and contextual relevance. Therefore, following previous practices, we randomly sample 128 conversations, and corresponding responses generated by different models for human ratings. We ask three human annotators to score each generated response from the following three aspects: (1) Empathy (**Emp.**), which measures whether the response understands user feelings and experiences. (2) Relevance (**Rel.**), which measures whether the response is on-topic and appropriate given the previous conversation. (3) Fluency (**Flu.**), which measures whether the response is fluent and its language is accurate. Each is on a $5-$point likert scale, where 5 is the best. Then we compute the average value for each metric.

Annotation Statistics: Table 2 displays the human rating results, and the highest scores are in bold. We calculate Fleiss's kappa to measure inter-evaluator agreement of the human ratings. The result is 0.41, indicating a moderate level of agreement among three annotators.

Results: From these results, we can draw two conclusions. First, compared with most previous models, CARE achieves the highest scores in terms of **Emp.** and **Rel.**, and obtains relatively high **Flu.**. It indicates that our causality reasoning in an interdependent and simultaneous way indeed benefits empathetic expression and content relevance as we expect. Thanks to the reasoned causalities, CARE improves the understanding of user feelings and experiences. In addition, the reasoning process enables the model to identify some reasonable user's feelings and experiences that are not explicitly mentioned in the previous conversation. With such information, the model can show strong empathy in response, which is manifest in the case study. Second, models considering both affection and cognition (bottom half of the table) do not always outperform models merely considering affection (upper half of the table). This is also evident in Table 3, i.e., the automatic evaluation results. Although causality reasoning intuitively contributes to the understanding of user's feelings and experiences, inconsiderate reasoning can lead to one-sided understanding and low empathy.

Model Variant	PPL	BLEU-3/4	P/R/F_{BERT}
w/o reasoning	33.34	4.74/2.83	.493/.473/.481
w/o condition	33.23	4.74/2.83	.501/.473/.485
Full model	32.84	4.88/2.95	.501/.475/.486

Table 3: Automatic evaluation results of the ablation study for CARE. The metrics are the same as those in Table 1. Similarly, we repeat five runs with different seeds, and display the average values. Its full automatic evaluation results can be found in Appendix A.1.

Model Variant	Emp.	Rel.	Flu.
w/o reasoning	2.38	2.23	4.86
w/o condition	2.60	2.47	4.87
Full model	2.83	2.79	4.86

Table 4: Human rating results of the ablation study for CARE. The metrics are the same as those in Table 2.

5.6 Model Analysis

In § 5.4 and § 5.5, CARE has shown its superior performance. For deeper analyses of our model, we investigate its inner structures and functions.

Ablation Study We propose two variant models to verify the contribution of reasoning and the reasoning condition in CARE:

- **w/o reasoning**: We remove the CVGAE structure, and directly incorporate the prior causal graph into response generation.
- **w/o condition**: We replace CVGAE with VGAE to eliminate the effect of the reasoning condition.

Results are shown in Table 3 and Table 4, respectively. From Table 3, both variants achieve relatively high automatic evaluation metric scores. Moreover, the variant models surpass previous comparison models in Table 1. It indicates that causalities can help models respond more like humans, given that both variants consider additional causalities between the user's experiences. However, both variants' performances in terms of human evaluation are relatively low. Accordingly, we can draw the following three conclusions:

- Not all information in the golden response contributes to empathy. Although two variants have high automatic evaluation scores, they fail to achieve equally high human ratings. Such a phenomenon can also be clearly observed when comparing the performance of EmpDG and KMEP.

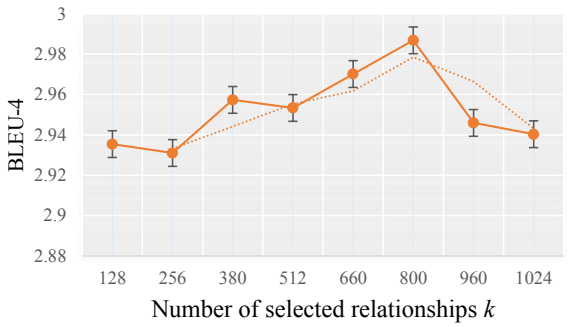

Figure 3: Model performs (**BLEU-4**) when we gradually increase the number of selected relationships k. The solid line and dot line represent BLEU-4 and two period moving average, respectively. For each k, we repeat five runs and compute the average BLEU-4.

- Generated responses considering causalities not mentioned in dialogue history are more empathetic and relevant, which is supported theoretically by Preston and De Waal (2002). It is reflected by higher **Emp.** and **Rel.** of **w/o condition** than **w/o reasoning**, given that **w/o condition** reasons plausible causalities not mentioned in dialogue history compared with **w/o reasoning**.
- Emotional and contextual information guides the model to reason causalities contributing to empathetic expression, given that the full model has higher Empathy and Relevant than **w/o condition**.

Analysis of #SelectedRelationships k As shown in Figure 3, the performance of CARE with regard to BLEU-4 first rises and then drops as we increase the number of relationships infused in response generation k. It indicates that sufficient causalities benefit empathetic expression, but excess ones could involve noise and hurt empathy.

5.7 Case Study

Table 5 presents a case along with responses generated by our models and comparison models. From the table, CARE can respond more empathically to the user when compared with other models. Notably, CARE is able to show deep and considerate comprehension of the user's feelings and experiences in the response. For instance, it understands that the "apprehensive" emotion comes from lack of *confidence* and the user has already proposed a quite effective solution (*great idea*).

User Emotion Type: apprehensive **User**: I had some hesitations when I was at the gym today, thought I wouldn't be able to handle the weights. **Bot**: How'd it turn out? Congrats on deciding to lead a healthier life, by the way. **User**: I decided to focus on the music that was blasting, just tried switching up what I play in the gym, it sounds like vikings are bringing me into battle	
MoEL	That is great. I am sure you will do great!
MIME	I am sure you will be fine.
EmpDG	That is a great time to have! You should always feel the same way.
KEMP	I am glad you are going through that.
CEM	I am sure it will be fun!
$\text{RecEC}_{\text{soft}}$	That is a lot of money. How long did you do?
GEE	I hate that. I hope you are okay.
CARE	That is a *good idea*. I hope you have a lot of *confidence*.

Table 5: Responses generated by our method and previous empathetic response generation models. The content showing comprehension of feelings and experiences is highlighted in italic.

6 Conclusion

In this paper, we propose to reason all plausible causalities in conversation interdependently and simultaneously for a deep understanding of the user's feelings and experiences in empathetic dialogue. Further, we turn the causality reasoning problem into a conditional graph generation task. Correspondingly, we design CVGAE, which uses latent variables for conditional structure prediction, and predicted future conversation content, to implement the reasoning. The reasoned causalities are infused into response generation for the final empathetic responses by a multi-source attention mechanism in the decoder. This whole structure is named as CARE (**CA**usality **R**easoning for **E**mpathetic conversation). Experimental results show that CARE outperforms prior methods in terms of both automatic and manual evaluations.

Limitations

In this paper, we improve the model's empathy from the aspect of affection and cognition, especially the latter one. For this purpose, we incorporate reasoned causal knowledge into response generation. However, other knowledge, such as sentiment knowledge and commonsense knowledge, can also contribute to affection and cognition. KEMP (Li et al., 2022), one of the comparison models in our experiment, has explored incorporating commonsense knowledge and sentiment knowledge into response generation. However, according to its model design, its use of knowledge is universal in chitchat conversations and is not aimed at empathetic expression. Therefore, it has low **Emp.** score as shown in Table 2. Therefore, it is worth exploring the connection between empathy and different types of knowledge. Besides, how to fuse different knowledge in one model for more empathetic responses is also a valuable problem.

Ethical Considerations

The widely-used open-sourced EMPATHETICDIALOGUES (Rashkin et al., 2019) benchmark used in our experiment is collected through interaction with Amazon Mechanical Turk (MTurk). In this process, user privacy is protected, and no personal information is contained in the dataset. Therefore, we believe that our research work meets the ethics of EMNLP.

Acknowledgements

This work was supported by the Research Grants Council of Hong Kong (PolyU/5204018, PolyU/15207920, PolyU/15207122) and National Natural Science Foundation of China (62076212).

References

Richárd Csáky, Patrik Purgai, and Gábor Recski. 2019. Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

Mark H Davis. 1983. Measuring individual differences in empathy: evidence for a multidimensional approach. *Journal of personality and social psychology*, 44(1):113.

Jean Decety and Philip L Jackson. 2004. The functional architecture of human empathy. *Behavioral and cognitive neuroscience reviews*.

Jun Gao, Yuhan Liu, Haolin Deng, Wei Wang, Yu Cao, Jiachen Du, and Ruifeng Xu. 2021. Improving empathetic response generation by recognizing emotion cause in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 807–819, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Thomas N Kipf and Max Welling. 2016. Variational graph auto-encoders. *ArXiv preprint*, abs/1611.07308.

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

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhongyang Li, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme. 2020b. Guided generation of cause and effect. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 3629–3636. ijcai.org.

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

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)*, pages 8968–8979, Online. Association for Computational Linguistics.

Ana Paiva, Iolanda Leite, Hana Boukricha, and Ipke Wachsmuth. 2017. Empathy in virtual agents and robots: A survey. *ACM Transactions on Interactive Intelligent Systems (TiiS)*.

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*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. 2017. Understanding and predicting empathic behavior in counseling therapy. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1426–1435, Vancouver, Canada. Association for Computational Linguistics.

Stephanie D Preston and Frans BM De Waal. 2002. Empathy: Its ultimate and proximate bases. *Behavioral and brain sciences*, 25(1):1–20.

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

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2022. Cem: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11229–11237.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Jiashuo Wang, Wenjie Li, Peiqin Lin, and Feiteng Mu. 2021. Empathetic response generation through graph-based multi-hop reasoning on emotional causality. *Knowledge-Based Systems*.

Tianming Wang and Xiaojun Wan. 2019. T-CVAE: transformer-based conditioned variational autoencoder for story completion. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5233–5239. ijcai.org.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

A Appendix

A.1 Automatic Evaluation

For our automatic evaluation, we modify codes[12] for dialogue evaluations provided by Csáky et al. (2019). In addition, Table 6 shows the full automatic evaluation results.

A.2 Human Evaluation

We implemented a system, as shown in Figure 4, for fair human ratings. For each case, we provide the previous dialogue turns, and the user emotion to the annotator. In addition, all responses generated by different models are displayed in a random order, thus the annotator cannot distinguish the source of each single response.

Since human ratings are subjective, we provide some statements and classic examples as the reference for human evaluation.

- *Empathy*. We prefer responses with the following features: (1). Emotions, e.g., care, concern, and encourage. (2). Content, which shows interests in what the user cares. For instance, we prefer *"Did you call the police?"* instead of *"What movie?"* when the user says *"It was stolen after the movie."*.
- *Relevance*. We prefer responses, based on which we can infer the topics in the previous dialogue content.
- *Fluency*. We reduce the marks if the following appears in a response: (1). Inappropriate (obvious features due to bad training) repetition, such as *"I am sorry. I am sorry. I am sorry. I am"*. (2). Grammar mistakes, e.g., misuse of personal pronouns and tense. (3). Conflicting contents, such as *"I can understand you. I cannot understand you."*.

Moreover, we encourage annotators to compare different responses in mind before grading each response.

[12] https://github.com/ricsinaruto/dialog-eval

Model	PPL	BLEU-3	BLEU-4	P_{BERT}	R_{BERT}	F_{BERT}
MoEL	36.42	4.74	2.89	.489	.469	.477
MoEL	37.43	5.23	3.23	.503	.476	.487
MoEL	37.94	4.60	2.82	.511	.470	.487
MoEL	37.12	3.61	2.25	.490	.456	.470
MoEL	35.45	4.48	2.83	.486	.469	.475
MIME	37.46	4.43	2.62	.491	.464	.475
MIME	38.36	4.21	2.60	.485	.464	.472
MIME	37.74	4.63	2.84	.496	.465	.478
MIME	37.40	4.60	2.79	.490	.469	.477
MIME	38.43	4.53	2.70	.486	.469	.475
EmpDG	59.01	3.42	1.76	.477	.465	.469
EmpDG	59.40	3.70	2.05	.463	.459	.459
EmpDG	56.41	3.03	1.72	.483	.444	.461
EmpDG	51.86	3.92	2.19	.476	.460	.466
EmpDG	51.54	4.13	2.25	.478	.464	.469
KEMP	36.87	4.44	2.61	.477	.465	.469
KEMP	37.24	3.86	2.27	.463	.459	.459
KEMP	36.26	3.70	2.22	.483	.444	.461
KEMP	36.17	4.28	2.50	.476	.460	.466
KEMP	36.42	4.36	2.54	.478	.464	.469
CEM	36.47	3.99	2.47	.498	.468	.480
CEM	37.12	2.93	1.87	.497	.451	.470
CEM	36.61	4.03	2.54	.500	.468	.482
CEM	36.13	3.40	2.18	.500	.460	.477
CEM	37.15	3.39	2.15	.497	.460	.475
$RecEC_{soft}$	139.78	3.25	1.80	.493	.463	.475
$RecEC_{soft}$	136.33	3.08	1.70	.494	.461	.475
$RecEC_{soft}$	173.85	3.03	1.62	.484	.462	.470
$RecEC_{soft}$	157.11	2.88	1.55	.492	.458	.472
$RecEC_{soft}$	139.60	2.85	1.46	.491	.462	.475
GEE_{MIME}	-	2.85	1.44	.473	.445	.457
GEE_{MIME}	-	2.83	1.61	.475	.443	.457
GEE_{MIME}	-	2.86	1.62	.474	.442	.456
GEE_{MIME}	-	2.40	1.26	.470	.441	.453
GEE_{MIME}	-	2.88	1.58	.470	.445	.455
CAER	32.64	5.03	2.99	.507	.477	.490
CAER	32.70	5.03	3.05	.497	.478	.485
CAER	33.16	4.87	2.95	.503	.476	.487
CAER	32.99	4.77	2.92	.495	.473	.482
CAER	32.69	4.70	2.86	.503	.473	.486
w/o reasoning	33.47	4.57	2.81	.493	.469	.479
w/o reasoning	33.39	4.80	2.82	.484	.472	.476
w/o reasoning	33.36	4.96	2.95	.490	.476	.481
w/o reasoning	33.18	4.70	2.79	.500	.474	.485
w/o reasoning	33.28	4.68	2.80	.496	.474	.483
w/o condition	32.56	4.70	2.85	.497	.475	.484
w/o condition	33.42	4.71	2.86	.496	.470	.480
w/o condition	32.86	4.97	2.96	.508	.475	.489
w/o condition	33.39	4.67	2.76	.502	.474	.486
w/o condition	33.94	4.65	2.75	.503	.470	.484

Table 6: All automatic results from different methods with seed 0, 42, 1024, 1234 and 4096.

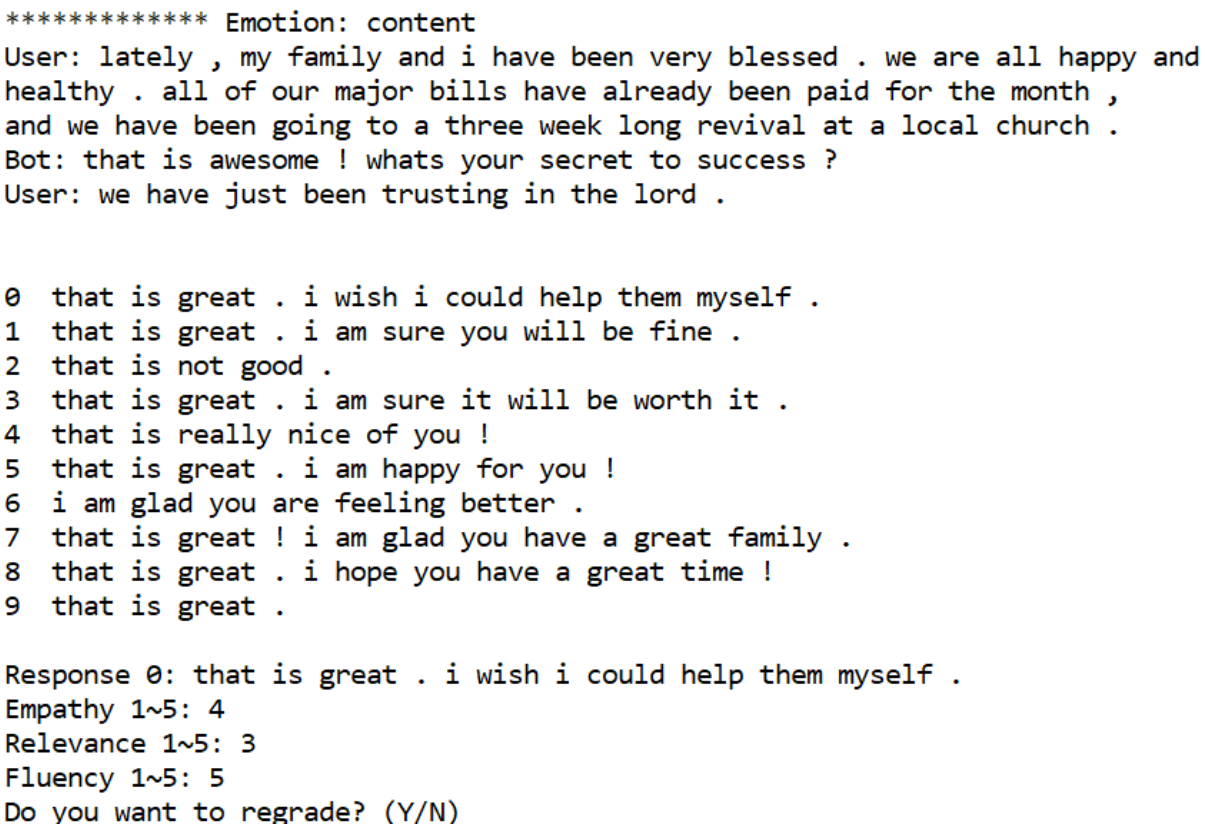

Figure 4: This is the user interface of the system for human ratings.