Wish I Can Feel What You Feel: A Neural Approach for Empathetic Response Generation

Yangbin Chen and **Chunfeng Liang**
Suzhou Fubian Medical Technology Co., Ltd., China
{dongyiwu92, cfliang666}@gmail.com

Abstract

Expressing empathy is important in everyday conversations, and exploring how empathy arises is crucial in automatic response generation. Most previous approaches consider only a single factor that affects empathy. However, in practice, empathy generation and expression is a very complex and dynamic psychological process. A listener needs to find out events which cause a speaker's emotions (emotion cause extraction), project the events into some experience (knowledge extension), and express empathy in the most appropriate way (communication mechanism). To this end, we propose a novel approach, which integrates the three components - emotion cause, knowledge graph, and communication mechanism for empathetic response generation. Experimental results on the benchmark dataset demonstrate the effectiveness of our method and show that incorporating the key components generates more informative and empathetic responses.

1 Introduction

According to Hoffman (2000), empathy is an affective response more appropriate to another's situation than one's own, which is the spark of human concern for others and the glue that makes social life possible. It is a complex human trait and dynamic psychological process related to emotion and cognition, where emotional empathy refers to vicarious sharing of emotion and cognitive empathy refers to mental perspective taking (Smith, 2006). Since 1990s, the study of empathy has been widely applied to mental health support (Bohart and Greenberg, 1997; Fitzpatrick et al., 2017), quality of care improvement (Mercer and Reynolds, 2002), and intelligent virtual assistants (Shin et al., 2019).

Expressing empathy becomes more important in today's dialogue systems. However, there are challenges in developing an empathetic model, such as preparing a proper training corpus, learning to get a comprehensive understanding of the dialogue context, and designing an appropriate empathy expression strategy.

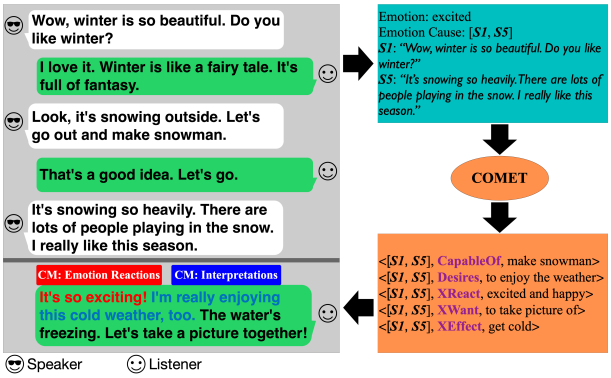

Figure 1: An example of empathetic response from EMPATHETICDIALOGUES dataset. In the teal box are emotion and causes detected from the dialogue context. In the orange box is extended knowledge via COMET. The colored texts in the final reply show two types of communication mechanisms.

Recently, there has been some work to address these issues. A standard benchmark containing large-scale empathetic conversations was proposed, laying the cornerstone of empathetic dialogue research (Rashkin et al., 2018). Some researchers try to gain a deeper understanding of contextual information. For example, Gao et al. (2021) applied an emotion cause extractor to conversations and used the extracted causes to guide the response generation process. Li et al. (2022) incorporated external commonsense information to enrich the context. During the language generation process, some researchers focus on controlling emotions of generated responses using emotional blending to imitate the speakers' emotions (Majumder et al., 2020; Lin et al., 2019).

All the above work considers only a single aspect that affects empathy. However, in practice, empathy generation and expression is a very complex and dynamic process. According to research work in the field of psychological science, we believe

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 922 - 933
December 7-11, 2022 ©2022 Association for Computational Linguistics

that three different but related factors matter in empathy: emotion (the automatic proclivity to share emotions with others), cognition (the intersubjectivity to interpret others' intentions and feelings while keeping separate self and other perspectives), and behavioral outcome (the actions to express empathy) (Decety and Meyer, 2008; Heyes, 2018). Consequently, we divide the entire empathy process into five functional modules: emotion perception, cause extraction, experience projection, dialogue reaction, and verbal expression. Specifically, emotion perception aims to sense emotions from others. Cause extraction is to determine detailed events corresponding to the emotions. Experience projection enriches the contextual information through knowledge extension from the emotion causes. Dialogue reaction decides the response strategies by learning from the contexts. Verbal expression is the final step in a dialogue system to generate responses in terms of languages.

Towards this end, we propose a novel approach **IMAGINE**, a.k.a. **I**ntegrating e**M**otion c**A**uses, knowled**G**e, and commun**I**catio**N** m**E**chanisms for empathetic dialogue generation. Using these components improves cognitive understanding of contexts and enhances empathy expression in the generated responses. Our framework involves three stages – emotion cause extraction, knowledge-enriched communication, and response generation. We evaluate our approach on the EMPATHETICDIALOGUES dataset. Extensive experimental results demonstrate the effectiveness of **IMAGINE** in automatic and human evaluations, showing that our approach generates more informative and empathetic responses (An example is shown in Figure 1).

Our contributions can be summarized as follows:

1) We propose a new approach **IMAGINE** which integrates emotion causes, knowledge, and communication mechanisms into a dialogue system, demonstrating that they are significant factors in the generation and expression of empathy.

2) We divide relationships within a knowledge graph into several categories, including Affect, Behaviour, Physical, and Events. Meanwhile, we design a three-stage process of emotion cause extraction, knowledge-enriched communication, and response generation based on the dialogue history.

3) Experimental results show that our proposed approach significantly outperforms other comparison methods, with more informative and empathetic responses.

2 Related Work

2.1 Empathetic dialogue generation

Empathetic response generation is a sub-task of emotion-aware response generation. Rashkin et al. (2018) first proposed a standard benchmark containing large-scale empathetic conversations. Some researchers focus on understanding the dialogue context. Li et al. (2021) and Gao et al. (2021) identified the emotion causes of the conversation to understand the context related to emotions better. Sabour et al. (2021) and Li et al. (2022) leveraged external knowledge, including commonsense knowledge and emotional lexical knowledge, to explicitly understand and express emotions. Some researchers focus on the language generation process, for example, controlling emotions of generated responses through mixture model (Lin et al., 2019), adversarial framework (Li et al., 2019), and mimicking the emotions of the speaker (Majumder et al., 2020). Sharma et al. (2020) and Zheng et al. (2021) explore the expressive factors that elicit empathy. Moreover, as big models are popular today, Lin et al. (2020) adapted GPT2 (Radford et al., 2019) to produce empathetic responses via transfer learning, active learning, and negative training.

2.2 What affects empathy?

Emotion Cause The emotion cause (also called antecedents, triggers, or stimuli) (Ellsworth and Scherer, 2003) is a stimulus for human emotions. Recognizing the emotion cause helps understand human emotions better to generate more empathetic responses. The cause could also be a speaker's counterpart reacting towards an event cared for by the speaker(inter-personal emotional influence). For example, understanding the sentence, "I like summer as it is a great time to surf," is not only to detect the positive emotion, HAPPY, but also to find its cause – "it is a great time to surf." The emotion cause recognition method (Poria et al., 2021) is used in our work.

External Knowledge A major part of cognitive empathy is understanding the situations and feelings of others. Conversations are limited in time and content. Therefore, using our experience (e.g., external knowledge) is important to connect what is explicitly mentioned and what is associated with it. In this work, we use the ATOMIC-2020 dataset (Hwang et al., 2020) as our commonsense

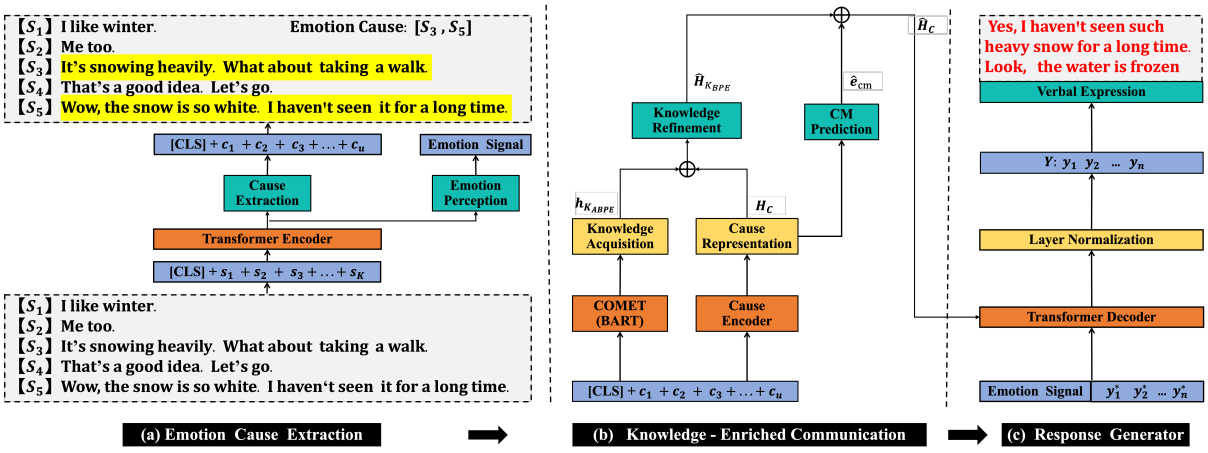

Figure 2: An overall framework of **IMAGINE**.

knowledge base, which is a collection of commonsense reasoning inferences about everyday if-then contexts. Detailed information about ATOMIC is covered in Appendix A.

Communication Mechanism (CM) For empathy generation, both conveying cognitive understanding (Truax and Carkhuff, 1967) and expressing stimulated emotions (Davis et al., 1980) are essential. Sharma et al. (2020) presented a computational approach to understanding empathy expressed in textual, asynchronous conversations and addressing both emotional and cognitive aspects of empathy. They developed components of an empathetic expression, consisting of three communication mechanisms - **Emotional Reaction** (expressing emotions such as warmth, compassion, and concern), **Interpretation** (conveying an understanding of feelings and experiences), and **Exploration** (improving understanding of the seeker by exploring the feelings and experiences).

2.3 Task Formulation

We formulate the task of empathetic response generation as follows. Given dialogue transcripts $\mathbf{S} = \{\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_k\}$ with k utterances, we firstly detect the emotion and extract emotion causes $\mathbf{C} = \{\mathbf{c}_0, \mathbf{c}_1, ..., \mathbf{c}_u\}$ which are a subset of $\mathbf{S}$. Each utterance $\mathbf{c}_i = \{\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, ..., \mathbf{c}_{i,l_i}\}$ is a sequence of tokens, where l_i denotes the length. Then, our goal is to generate an empathetic response $\mathbf{Y} = \{\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n\}$ given the sequence $\mathbf{C}$, with the assistance of external knowledge and communication mechanisms.

3 Approach

Our proposed model, **IMAGINE**, is built upon the standard Transformer (Vaswani et al., 2017) and its overview is illustrated in Figure 2. It has three stages consisting of five functional modules: emotion cause extraction (emotion perception, cause extraction), knowledge-enriched communication (dialogue reaction, experience projection), and response generation (verbal expression). Emotion perception predicts emotions of the input. Cause extraction extracts causes related to the emotions from the input. Experience projection acquires knowledge based on the causes mentioned above. Dialogue reaction decides the response strategies by learning from the contexts. Verbal expression integrates the information obtained from the above four modules and generates appropriate responses.

3.1 Emotion Cause Extraction

Given a dialogue context consisting of k utterances with the context emotion, the goal of emotion cause extraction is to identify which utterances in the dialogue context contain the emotion cause. We leverage an existing model which is trained on an open-domain emotional dialogue dataset named RECCON, for identifying emotion causes at utterance level in conversations (Poria et al., 2021). Gao et al. (2021) has verified the model's validity, and we follow the method in the first stage of our work.

3.1.1 Emotion Perception

It is a classification problem aiming at predicting the emotion ε within the dialogue context. Given the dialogue context $\mathbf{S} = \{\mathbf{s}_0, \mathbf{s}_1, ..., \mathbf{s}_k\}$ as the input, the tokens are then fed into a transformer-based encoder to obtain a sequence of contextualized rep-

resentations $\mathbf{H}_S$. Hence, we pass $\mathbf{H}_S$ through a linear layer followed by a softmax operation to produce the emotion category distribution:

$$\hat{\mathbf{e}}_{emo} = \mathbf{W}_e\mathbf{H}_S[0]+\mathbf{b}_e, \quad (1)$$

$$\hat{\mathbf{P}}(\varepsilon|\mathbf{S}) = \mathbf{softmax}(\hat{\mathbf{e}}_{emo}), \quad (2)$$

where $\mathbf{W}_e$ and $\mathbf{b}_e$ are trainable parameters. During training, we employ negative log-likelihood as the emotion perception loss:

$$\mathbf{L}_{emo} = -\mathbf{log}(\hat{\mathbf{P}}(\varepsilon = \mathbf{e}^*|\mathbf{S})), \quad (3)$$

where $\mathbf{e}^*$ denotes the emotion label, and ε denotes the predicted output. Emotional vectors $\hat{\mathbf{e}}_{\mathbf{emo}}$ will be fed into the decoder as a crucial emotional signal to guide the empathetic response generation.

3.1.2 Cause Extraction

Given the dialogue context $\mathbf{S}$ and its emotion ε, we extract emotion causes $\mathbf{C} = \{\mathbf{c}_0, \mathbf{c}_1, ..., \mathbf{c}_u\}$ according to the approach in Poria et al. (2021). The causes $\mathbf{C}$ are a subset of $\mathbf{S}$ and will be used as the input of the next two stages. Following previous work (Lin et al., 2019; Majumder et al., 2020; Sabour et al., 2021), we concatenate the utterances indicating emotion causes and prepend a special token $[CLS]$ to obtain the cause input $\mathbf{C} = [CLS] + \mathbf{c}_0 + \mathbf{c}_1 + ... + \mathbf{c}_u$. Each utterance $\mathbf{c}_i$ contains a sequence of tokens: $\mathbf{c}_i = \{\mathbf{c}_{i,1}, \mathbf{c}_{i,2}, ..., \mathbf{c}_{i,l_i}\}$, where l_i is the length of $\mathbf{c}_i$.

Each token is represented from three aspects: its semantic meaning, its position in the sequence, and who said it. Suppose that the token ID and the position ID of $\mathbf{c}_{i,j}$ are $w_{\mathbf{c}_{i,j}} \in [0, |\mathbf{V}|)$ ($\mathbf{V}$ is the vocabulary) and $p_{\mathbf{c}_{i,j}}$, respectively. Additionally, in multi-turn dialogue settings, distinguishing a listener from a speaker is helpful. So we incorporate the dialogue state embedding into our input sequence. Specifically, each utterance $\mathbf{c}_i$ is labeled with its corresponding role $s_{\mathbf{c}_i} \in \{0, 1\}$ (0 for speaker and 1 for listener).

The token $\mathbf{c}_{i,j}$ is represented by summing up the word embedding, positional embedding, and dialogue state embedding:

$$\mathbf{E}_{\mathbf{c}_{i,j}} = \mathbf{E}_W[w_{\mathbf{c}_{i,j}}] + \mathbf{E}_P[p_{\mathbf{c}_{i,j}}] + \mathbf{E}_S[s_{\mathbf{c}_i}], \quad (4)$$

where $\mathbf{E}_W \in \mathbb{R}^{|V|\times d}$, $\mathbf{E}_P \in \mathbb{R}^{1024\times d}$, $\mathbf{E}_S \in \mathbb{R}^{2\times d}$ denote the embedding matrices of word, position, and state. $[\cdot]$ denotes the indexing operation, and d is the dimensionality of embeddings. We feed the entire sequence of token embeddings $\mathbf{E}_C$ organized by $\mathbf{E}_{\mathbf{c}_{i,j}}$ to a cause encoder to produce the contextual representation:

$$\mathbf{H}_C = \mathbf{Cause\text{-}Encoder}(\mathbf{E}_C), \quad (5)$$

where $\mathbf{H}_C \in \mathbb{R}^{|L|\times d}$, L is the length of the sequence, and d is the hidden size of the cause encoder.

Next, we use the hidden state at the $[CLS]$ of the cause encoder, $\mathbf{h}_c = \mathbf{H}_C[0]$, to predict CM strategies in the following stage.

3.2 Knowledge - Enriched Communication

3.2.1 Dialogue Reaction

CM Prediction While no empathetic conversation corpora provide annotations of diverse empathy factors, there are abundant publicly available resources that make automatic annotation feasible. We use two corpora annotated with CM provided by Sharma et al. (2020). There are three communication factors named Emotion Reaction (ER), Interpretation (IP), and Exploration (EX). Each mechanism has different degrees. In our work, we merge "weak" and "strong" into "yes" and differentiate each mechanism's degree into two types: "no" and "yes".

We pass $\mathbf{h}_c$ through a linear layer followed by a softmax operation to produce the CM category distribution:

$$\mathbf{e}_{cmi} = \mathbf{W}_{cmi}\mathbf{h}_c+\mathbf{b}_{cmi}, \quad cmi \in \{er, ip, ex\} \quad (6)$$

$$\hat{\mathbf{P}}_{cmi} = \mathbf{softmax}(\mathbf{e_i}), \quad (7)$$

The negative log-likelihood loss is calculated:

$$\mathbf{L}_{cm} = \sum_{cmi\in\{er,ip,ex\}} -\mathbf{log}(\hat{\mathbf{P}}_{cmi}), \quad (8)$$

Finally, $\mathbf{e}_{er}$, $\mathbf{e}_{ip}$, $\mathbf{e}_{ex}$ are summed up, weighted by their predicted degree, as a crucial CM signal:

$$\hat{\mathbf{e}}_{cm} = \hat{\mathbf{P}}_{er} \cdot \mathbf{e}_{er} + \hat{\mathbf{P}}_{ip} \cdot \mathbf{e}_{ip} + \hat{\mathbf{P}}_{ex} \cdot \mathbf{e}_{ex}, \quad (9)$$

3.2.2 Experience Projection

Knowledge Acquisition We extend the contexts by selecting from the knowledge graph those that are speaker-centered and contribute positively to the speaker. Finally, we split ATOMIC-2020 (Hwang et al., 2020) into four types: *Affect*, *Behaviour*, *Physical*, and *Events*, containing 11 relations $[r_1, r_2, ..., r_{11}]$ in total (See Figure 3). In Affect, we select one

relation: ($[XReact]$). In Behaviour, we select five relations: ($[XIntent]$, $[XNeed]$, $[XWant]$, $[XEffect]$, $[XAttr]$). In Physical, we select three relations: ($[HasProperty]$, $[CapableOf]$, $[Desires]$). In Events, we select two relations: ($[Causes]$, $[XReason]$). For an input sequence $\mathbf{C}$, we use **COMET** (Lewis et al., 2019) to generate five commonsense-inferred entities [$s_1^{r_i}$, $s_2^{r_i}$, $s_3^{r_i}$, $s_4^{r_i}$, $s_5^{r_i}$] for each relation r_i. Then we concatenate all entities generated from relations belonging to the same relation type. Through this way, we obtain four commonsense sequences for each input sequence: S_{Affect}, S_{Behav}, S_{Phys}, and S_{Events}. For example, $S_{Events} = [s_1^{[Causes]}, ..., s_5^{[Causes]}, s_1^{[XReasons]}, ..., s_5^{[XReasons]}]$. We prepend $[CLS]$ to S_{Behav}, S_{Phys}, and S_{Events}. S_{Affect} does not change because the entities for *Affect* are usually independent emotion words (e.g., happy, surprise, sad) rather than semantically coherent sequences. The commonsense sequences are fed to the knowledge encoder:

$$\mathbf{H}_{K_{ABPE}} = \textbf{Knowledge-Encoder}(S_{K_{ABPE}}), \tag{10}$$

where $K_{ABPE} \in \{Affect, Behav, Phys, Events\}$. $\mathbf{H}_{K_{ABPE}} \in \mathbb{R}^{|L_{K_{ABPE}}| \times d}$, with $|L_{K_{ABPE}}|$ being lengths of the commonsense entity sequences.

Next, we use hidden representations of the first position to represent sequences S_{Behav}, S_{Phys}, and S_{Events}, respectively:

$$\mathbf{h}_{K_{BPE}} = \mathbf{H}_{K_{BPE}}[0] \tag{11}$$

where $K_{BPE} \in \{Behav, Phys, Events\}$.

Moreover, we use the mean of hidden representations to represent S_{Affect}:

$$\mathbf{h}_{Affect} = \textbf{Average}(\mathbf{H}_{Affect})|_{axis=0}, \tag{12}$$

Knowledge Refinement In order to refine the emotion causes by knowledge information, we concatenate each commonsense relation representation $\mathbf{h}_{K_{ABPE}}$ to the cause representation $\mathbf{H}_C$ at the token level. In contrast to sequence-level concatenation, token-level concatenation enables us to fuse knowledge within each word in the cause sequence:

$$\mathbf{U}_{K_{ABPE}} = \mathbf{H}_C \oplus \mathbf{h}_{K_{ABPE}}, \tag{13}$$

where $\mathbf{U}_{Affect}, \mathbf{U}_{Behav}, \mathbf{U}_{Phys}, \mathbf{U}_{Events} \in \mathbb{R}^{|L| \times 2d}$.

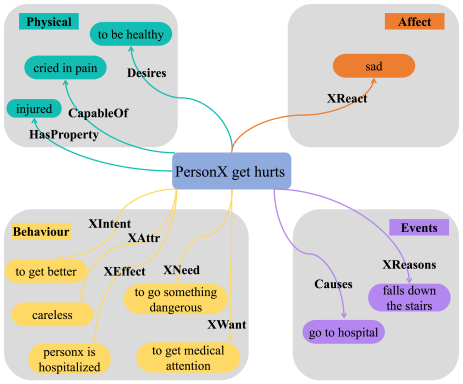

Figure 3: The four modules of the Knowledge Graph.

Accordingly, we encode the fused representations and obtain knowledge-refined cause representations for each relation type:

$$\mathbf{H}^{ref}_{K_{ABPE}} = \textbf{Refine-Encoder}(\mathbf{U}_{K_{ABPE}}), \tag{14}$$

where $\mathbf{H}^{ref}_{K_{Affect}}, \mathbf{H}^{ref}_{K_{Behav}}, \mathbf{H}^{ref}_{K_{Phys}}, \mathbf{H}^{ref}_{K_{Events}} \in \mathbb{R}^{|L| \times d}$.

We believe that relations of the *Affect* type matter to emotional empathy, meanwhile relations of *Behavior*, *Physical*, and *Events* types matter to cognitive empathy. Hence, we re-represent the knowledge-refined cause representations as below:

$$\tilde{\mathbf{H}}_{K_{BPE}} = \mathbf{H}^{ref}_{K_{BPE}} \oplus \mathbf{H}^{ref}_{Affect}, \tag{15}$$

where $\tilde{\mathbf{H}}_{Behav}, \tilde{\mathbf{H}}_{Phys}, \tilde{\mathbf{H}}_{Events} \in \mathbb{R}^{|L| \times 2d}$.

Next, to highlight important features within the knowledge-refined cause representation, we assign importance scores to $\tilde{\mathbf{H}}_{K_{BPE}}$, followed by a Multi-Layer Perception (MLP) layer with ReLU:

$$\hat{\mathbf{H}}_{K_{BPE}} = \textbf{MLP}(\sigma(\tilde{\mathbf{H}}_{K_{BPE}}) \cdot \tilde{\mathbf{H}}_{K_{BPE}}) \tag{16}$$

where $\hat{\mathbf{H}}_{Behav}, \hat{\mathbf{H}}_{Phys}, \hat{\mathbf{H}}_{Events} \in \mathbb{R}^{|L| \times d}$, and $\cdot$ denotes element-wise multiplication.

Finally, $\hat{\mathbf{H}}_{Behav}, \hat{\mathbf{H}}_{Phys}, \hat{\mathbf{H}}_{Events}$ and $\hat{\mathbf{e}}_{cm}$ (Equation 9), are fed into the decoder:

$$\hat{\mathbf{H}}_C = \hat{\mathbf{H}}_{Behav} \oplus \hat{\mathbf{H}}_{Phys} \oplus \hat{\mathbf{H}}_{Events} \oplus \hat{\mathbf{e}}_{cm} \tag{17}$$

where $\hat{\mathbf{H}}_C \in \mathbb{R}^{|L| \times 4d}$.

3.3 Response Generation

Verbal Expression To acquire emotion dependencies, we concatenate the intermediate emotional signal $\hat{\mathbf{e}}_{emo}$ with word embeddings of the expected response and get [$\mathbf{y}_0^*, \mathbf{y}_1^*, \mathbf{y}_2^*, ..., \mathbf{y}_n^*$]. Here $\mathbf{y}_0^*$ is $\hat{\mathbf{e}}_{emo}$. We then feed the embeddings into the

response decoder. Our decoder is built based on Transformer layers:

$$\mathbf{P}(y_t|y_{<t}, \mathbf{C}) = \mathbf{Decoder}(\mathbf{E}_{y_{<t}}, \hat{\mathbf{H}}_C), \quad (18)$$

where $\mathbf{E}_{y_{<t}}$ denotes embeddings of tokens that have been generated. Note that the cross attention to the encoder outputs is modified to the knowledge-refined cause representation $\hat{\mathbf{H}}_C$, which has fused the information from both the cause and the commonsense-inferred entities.

3.4 Model Training

We use negative log-likelihood of the ground-truth words $\mathbf{y}_t^*$ as the generation loss function:

$$\boldsymbol{L}_{gen} = -\sum_{t=1}^{n} \mathbf{log}\mathbf{P}(\mathbf{y}_t = \mathbf{y}_t^*|\mathbf{y}_0, ..., \mathbf{y}_{t-1}, \mathbf{C}) \quad (19)$$

Dialogue generation models sometimes generate repetitive phrases or generic responses, such as "That is a good idea" and "Oh, it is bad." To solve this problem, we apply the Response Diversity Loss in our model, implementing Frequency-Aware Cross-Entropy (FACE) (Jiang et al., 2019) as an additional loss to penalize high-frequency tokens using a weighting scheme. Hence, during training, prior to receiving a new batch of samples, we derive the frequency-based weight $\mathbf{w}_i$ for each vocabulary token $\mathbf{v}_i$ in the training corpus:

$$\mathbf{w}_i = \mathbf{a} \times \mathbf{FQ}_i + 1, \quad (20)$$

$$\mathbf{FQ}_i = \frac{\mathbf{freq}(\mathbf{v}_i)}{\sum_{j=1}^{V} \mathbf{freq}(\mathbf{v}_j)}, \quad (21)$$

where V denotes the vocabulary size, $\mathbf{a} = -(\mathbf{max}_{0<j<V}(\mathbf{FQ}_j))^{-1}$ is the frequency slope and 1 is added as the bias so that $\mathbf{w}_i$ falls into [0,1]. Lastly, we normalize $\mathbf{w}_i$ to have mean of 1, as done by (Jiang et al., 2019). The diversity loss would then be calculated as below:

$$\boldsymbol{L}_{div} = -\sum_{t}^{n}\sum_{i}^{V} \mathbf{w}_i \delta(\mathbf{v}_i = \mathbf{y}_t^*) \mathbf{log}\mathbf{P}(\mathbf{v}_i|\mathbf{y}_{<t}, \mathbf{C}) \quad (22)$$

where $\mathbf{v}_i$ is a candidate token in the vocabulary and δ is the indicator function, which equals to 1 if and only if $\mathbf{v}_i = \mathbf{y}_t^*$ and 0 otherwise. All parameters of our proposed model are trained and optimized based on the weighted sum of four losses:

$$\boldsymbol{L} = \lambda_1 \boldsymbol{L}_{gen} + \lambda_2 \boldsymbol{L}_{emo} + \lambda_3 \boldsymbol{L}_{cm} + \lambda_4 \boldsymbol{L}_{div}, \quad (23)$$

where λ_1, λ_2, λ_3 and λ_4 are hyper-parameters that we use to control the influence of the four losses. Loss weights λ_1, λ_2, λ_3 and λ_4 are set to 1, 1, 1, and 1.5, respectively.

4 Experimental Settings

4.1 Dataset

We conduct our experiments on the EMPATHETIC-DIALOGUES dataset (Rashkin et al., 2018). It is a large-scale multi-turn empathetic dialogue dataset containing 25k dialogue sessions, each having 3-5 rounds of dialogue. There are 32 different distributions of emotion labels. Following the original dataset definitions, we use the 8:1:1 train/valid/test subset split.

4.2 Comparison Methods

The following models are selected as baselines:
1) Transformer (Vaswani et al., 2017): A Transformer based encoder-decoder model.
2) Multi-TRS (Rashkin et al., 2018): An extension of the Transformer model that has an additional unit for emotion prediction.
3) MoEL (Lin et al., 2019): Another extension of Transformer model which softly combines the response representations from different decoders.
4) MIME (Majumder et al., 2020): Another extension of transformer model which considers emotion clustering and emotional mimicry. Besides, it also introduces sampling stochasticity during training.
5) EMPDG (Li et al., 2019): A multi-resolution empathetic adversarial chatbot which exploits multi-resolution emotions and user feedback.
6) CEM (Sabour et al., 2021): A Transformer encoder-decoder model that integrates affection and cognition into commonsense knowledge.
7) KEMP (Li et al., 2022): A contextual-enhanced empathetic dialogue generator that leverages multi-type external knowledge and emotional signal distilling for response generation.

More implementation details of our **IMAGINE** model is covered in Appendix B.

4.3 Evaluation metrics

Automatic Evaluations Four automatic metrics are applied for evaluation:
1) PPL (Serban et al., 2015): The perplexity (PPL) represents the model's confidence in its set of candidate responses. A low PPL value means high confidence. PPL can be used to evaluate the general quality of the generated responses.

Models	PPL	BLEU-2	Dinstinct-1	Distinct-2	ACC	Fluency	Relevance	Empathy
Transformer	37.62	1.32	0.45	2.02	-	3.04	2.49	2.50
Multi-TRS	37.75	1.31	0.41	1.67	33.57	2.99	2.51	2.59
MoEL	36.93	1.32	0.44	2.10	30.62	3.28	2.57	2.63
MIME	37.09	1.34	0.47	1.90	31.36	3.14	2.52	2.59
EmpDG	37.29	1.30	0.46	2.02	30.41	3.07	2.69	2.72
CEM	36.11	1.35	0.66	2.99	39.11	3.40	2.96	2.94
KEMP	36.89	1.34	0.55	2.29	39.31	3.27	2.68	2.68
IMAGINE	**35.10**	**1.37**	**0.76**	**3.40**	**39.60**	**3.58**	**3.09**	**3.09**

Table 1: Results of automatic and human evaluations.

Models	PPL	BLEU-2	Dinstinct-1	Distinct-2	ACC
IMAGINE	35.10	**1.37**	**0.76**	**3.40**	**39.60**
W/O cause	35.43	1.35	0.64	2.57	38.60
W/O cm	35.58	1.34	0.63	2.84	38.88
W/O know	35.0	1.36	0.64	2.92	38.50
W/O DIV	**34.50**	1.37	0.68	2.94	39.10

Table 2: Ablation study.

Models	Win%	Lose%	Tie%
Ours VS Transformer	49.18	16.83	33.99
Ours VS Multi-TRS	42.34	17.66	40.00
Ours VS MOEL	45.49	27.42	27.09
Ours VS MIME	47.34	19.33	33.33
Ours VS EmpDG	47.18	19.60	33.22
Ours VS CEM	42.96	25.80	31.24
Ours VS KEMP	41.90	23.98	34.12

Table 3: Results of human A/B test.

2) BLEU-2 (Papineni et al., 2002): It calculates the co-occurrence frequency of n-grams between candidates and references.

3) Distinct-1 and Distinct-2 (Li et al., 2015): It is the proportion of the distinct unigrams/bigrams in all the generated results to indicate the diversity.

4) ACC: To evaluate the model at the emotional level, we adopt Emotion Accuracy (ACC) as the agreement between the ground truth emotion labels and the predicted emotion labels.

Human Ratings Evaluating open-domain dialogue systems is challenging due to the lack of reliable automatic evaluation metrics (Gao et al., 2021b). Thus, human judgments are necessary. We randomly sample 100 dialogues and generate corresponding responses from different models. Five well-educated native English speakers who work in literary writing, psychology, and teaching are hired to give each response a rating score from three aspects – Fluency, Relevance, and Empathy. Each aspect is on a scale from 1 to 5, where 1, 2, 3, 4, and 5 indicate unacceptable, not good, moderate, good, and excellent performance, respectively. In order to keep the anonymization of compared methods, the order of responses in each dialogue is shuffled.

Human A/B Test In the human A/B test, to make sure fairness, we re-sample another 700 dialogues (100 for each comparison between our model and a baseline model) and form them into A-vs-B types, where A is our model and B is a baseline model. Another three annotators are asked to choose a better response. They can also choose a Tie if they think both are good or bad. All human evaluation tasks are conducted on https://www.fanhantech.com.

5 Experimental Results

5.1 Automatic Evaluation Results

Table 1 reports the evaluation results on automatic metrics. Our model **IMAGINE** achieves the lowest perplexity, indicating that the overall quality of our generated responses is higher than the baselines. Moreover, the results of Distinct-1 and Distinct-2 show that our model generates much more diverse responses than baselines. As for the emotion accuracy, we can see that our model is

Emotion: afraid **Emotion Cause:** [S_1, S_3] **Context:** 【S_1】: i recently had a scary experience . i heard gunshots on my street . 【S_2】: oh my ! you must have been terrified ! did everything turn out ok ? 【S_3】: everything turned out fine . i think somebody in the nearby neighborhoods may have been hurt , though . **Gold:** that is not good at all . that is a very scary situation .	**Emotion:** prepared **Emotion Cause:** [S_1] **Context:** 【S_1】: one time i studied my butt off for my first psychology exam . it was hard and i studied all night. 【S_2】: how did you do 【S_3】: i ended up actually getting a 98 % ! **Gold:** that is great ! my b.s is in psychology .
TRS: that is a pretty big deal ! i hope you can get hurt ! **MULTI-TRS:** i am glad you are okay ! **MOEL:** i would have been too ! **MIME:** i am sorry to hear that . i hope you can find a better place to get that way . **EMPDG:** i would be too ! i would be too ! **CEM:** i am glad you are okay . **KEMP:** i would be too ! i would be terrified !	**TRS:** that is great ! i am sure you will do great ! **MULTI-TRS:** that is great ! i bet you were very proud ! **MOEL:** that is awesome ! i hope you did well . **MIME:** that is great ! **EMPDG:** oh wow ! i bet you were proud of you ! **CEM:** that is great ! **KEMP:** i am sure you will do great !
IMAGINE (ours): oh, no. I would be so scared, too. I'm glad you're safe in the surroundings. (Emotion Reaction, Interpretation) ***Affect*:** terrified, afraid ***Behaviour*:** to be aware of the surroungings, to avoid harm ***Physical*:** frightend, to be scared, feel safe ***Events*:** I was scared, nervous	**IMAGINE (ours):** What is awesome! Did you pass the exam? I bet you were very prepared and made successes. (Emotion Reaction, Exploration) ***Affect*:** proud, happy, prepared ***Behaviour*:** to be prepared, to pass the exam, to get good grades ***Physical*:** study hard, successful ***Events*:** study more, smart

Figure 4: Case study of the generated responses by **IMAGINE** and the baselines.

valid for recognizing emotions.

5.2 Human Evaluation Results

Table 1 illustrates that **IMAGINE** obtains the best performance on Fluency, Relevance, and Empathy scores. It proves that integrating emotion causes, knowledge, and communication mechanisms can generate more informative and empathetic responses. In addition, from the results of the human A/B test in Table 3, we see that responses from **IMAGINE** are more often preferable to humans than the responses from other baseline models, which strongly supports the advantages of our approach.

5.3 Ablation Analysis

We conducted ablation studies to verify the effectiveness of each component in our model. Table 2 reports the results.

1) W/O cause: Looking at Table 2, we can see that removing the emotion cause extraction part leads to a significant performance decrease of both models in terms of response generation and emotion recognition. The original dialogue history may contain emotion-irrelevant information, which results in a shift of focus. The result indicates that emotion cause extraction plays an important role in strengthening the understanding of users' emotions, which improves the generation of empathetic responses.

2) W/O CM: By removing the communication mechanism from the response generation module, as shown in Table 2, we can see that our model is less empathetic and also has a tendency to decline in emotion prediction. The communication mechanism is a state of understanding how people feel; without it, our model will have fewer communication skills.

3) W/O know: When we remove the knowledge module, as shown in Table 2, we can see that the quality and diversity of the model's responses are declined, as a lack of knowledge leads to weaker ability to enrich emotion causes. It also affects the closeness and relevance of the generated responses to the context.

4) W/O DIV: If the diversity loss is removed, we can see from Table 2 that Distinct-1 is reduced from 0.76 to 0.68, and Distinct-2 is reduced from 3.4 to 2.94. It indicates the effectiveness of this loss in generating more diverse responses.

5.4 Case Study

We also present some examples of responses generated by our models and baseline models in Figure 4. Compared with baseline models, our model generates responses closer to the "gold" responses. As shown in the first example, our model can reason deeply about the emotion cause and get a good result in terms of knowledge acquisition. In the second example, from the dialogue context, we learn that the user "studied hard and got good grades." Through the knowledge base, we infer richer in-

formation like "prepared, successful, and pass the exam." Finally, our model congratulates and praises the user and poses an unasked question to him/her.

6 Conclusion

This paper presents a novel framework that integrates emotion causes, knowledge graphs, and communication mechanisms for empathetic response generation. The emotion cause detection allows us to determine what events stimulate a user's emotion. We can understand the events with the knowledge graph, enriching the contextual information. Furthermore, the communication mechanisms enhance our ability to let users feel that we are tryining to feel what they feel. Automatic and human evaluations show that our proposed approach can generate more informative and empathetic responses.

Limitations

The first challenge is a common problem current chatbots face, e.g., traceability of models and reasoning ability. Second, for mental health support chatbots, each person is analyzed on a case-by-case basis. Each person with a mental health impairment needs a personalized approach to communication, which is not overly generalized. Finally, the shortcomings of the knowledge graph - size, breadth, diversity, and rationality - directly determine the quality of the causes' associative expansion and also affect the closeness and relevance of the generated responses to the context.

Ethics Statement

The empathetic-dialogues dataset (Rashkin et al., 2018) used in our paper protects the privacy of real users. Furthermore, we make sure anonymization in the human evaluation process. We believe our research work meets the ethics of EMNLP.

References

Arthur C Bohart and Leslie S Greenberg. 1997. *Empathy reconsidered: New directions in psychotherapy.* American Psychological Association.

Mark H Davis et al. 1980. A multidimensional approach to individual differences in empathy.

Jean Decety and Meghan Meyer. 2008. From emotion resonance to empathic understanding: A social developmental neuroscience account. *Development and psychopathology*, 20(4):1053–1080.

Phoebe C Ellsworth and Klaus R Scherer. 2003. *Appraisal processes in emotion.* Oxford University Press.

Kathleen Kara Fitzpatrick, Alison Darcy, and Molly Vierhile. 2017. Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (woebot): a randomized controlled trial. *JMIR mental health*, 4(2):e7785.

Jun Gao, Wei Bi, Ruifeng Xu, and Shuming Shi. 2021b. Ream: An enhancement approach to reference-based evaluation metrics for open-domain dialog generation: An enhancement approach to reference-based evaluation metrics for open-domain dialog generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2487–2500.

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.

Cecilia Heyes. 2018. Empathy is not in our genes. *Neuroscience & Biobehavioral Reviews*, 95:499–507.

Martin L. Hoffman. 2000. Empathy and moral development : implications for caring and justice. *Contemporary Sociology*, 30:487.

Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2020. Comet-atomic 2020: On symbolic and neural commonsense knowledge graphs. *arXiv preprint arXiv:2010.05953*.

Shaojie Jiang, Pengjie Ren, Christof Monz, and Maarten de Rijke. 2019. Improving neural response diversity with frequency-aware cross-entropy loss. In *The World Wide Web Conference*, pages 2879–2885.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. 2019. Empdg: Multiresolution interactive empathetic dialogue generation. *arXiv preprint arXiv:1911.08698*.

Qintong Li, Piji Li, Zhaochun Ren, Pengjie Ren, and Zhumin Chen. 2022. Knowledge bridging for empathetic dialogue generation.

Yanran Li, Ke Li, Hongke Ning, Xiaoqiang Xia, Yalong Guo, Chen Wei, Jianwei Cui, and Bin Wang. 2021. Towards an online empathetic chatbot with emotion causes. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2041–2045.

Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. 2019. Moel: Mixture of empathetic listeners. *arXiv preprint arXiv:1908.07687*.

Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. 2020. Caire: An end-to-end empathetic chatbot. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13622–13623.

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. *arXiv preprint arXiv:2010.01454*.

Stewart W Mercer and William J Reynolds. 2002. Empathy and quality of care. *British Journal of General Practice*, 52(Suppl):S9–12.

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.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13(5):1317–1332.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Sahand Sabour, Chujie Zheng, and Minlie Huang. 2021. Cem: Commonsense-aware empathetic response generation. *arXiv preprint arXiv:2109.05739*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8):434–441.

Ashish Sharma, Adam S Miner, David C Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. *arXiv preprint arXiv:2009.08441*.

Jamin Shin, Peng Xu, Andrea Madotto, and Pascale Fung. 2019. Happybot: Generating empathetic dialogue responses by improving user experience lookahead. *arXiv preprint arXiv:1906.08487*.

Adam Smith. 2006. Cognitive empathy and emotional empathy in human behavior and evolution. *The Psychological Record*, 56(1):3–21.

Charles B Truax and Robert Carkhuff. 1967. *Toward effective counseling and psychotherapy: Training and practice*. Transaction Publishers.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. Comae: a multi-factor hierarchical framework for empathetic response generation. *arXiv preprint arXiv:2105.08316*.

A knowledge Graph

In this work, we use the ATOMIC-2020 dataset (Hwang et al., 2020) as our commonsense knowledge base, which is a collection of commonsense reasoning inferences about everyday if-then contexts. They fall into three natural categories based on their meaning: physical-entity, social-interaction, and event-centered commonsense, which are 22 relationships under three categories (e.g., XReact, XWant, XReason, CapableOf) (See Fig 5). Based on the given contexts, we select those that are speaker-centered and contribute positively to the speaker. We neglect (oReact,oEffect,oWant, Etc.) in our work. Finally, We have extracted 11 important relationships from ATOMIC. These relationships are divided into four modules, which are Physical (CapableOf, HasProperty, Desires), Affect (XReact), Behaviour (XEffect, XNeed, XWant, XIntent, XAttr), Events (Causes, XReason). As shown in Fig 6 .

B Implementation Details

Our models are implemented using Pytorch, a modularized, versatile, and extensible toolkit for machine learning and text generation tasks. We used 300-dimensional word embedding and 300-dimensional hidden size everywhere in our experiments. The word embedding is initialized using pre-trained Glove vectors. We initialize the transformer encoder with one layer and two attention heads for the task. We train our models using Adam optimization with a learning rate of 0.0001. Early

Categories	Head	Relations	COMET (BART)
Physical-entity	Common sense	ObjectUse	Make a good decision
	bird lover	CapableOf	Look at birds
	ice cream	MadeUpOf	milk and warter
	mouse	HasProPerty	Long tail
	doctors	Desires	cure patient
	doctors	NotDesires	malpractice suit
	gambler	AtLocation	casino
Social-Interaction	X accepts Y's apology	XIntent	To be forgiving
	X gives Y gifts	XReact	good about [one]self
	X gives Y gifts	OReact	appreciated
	X steals a car	XAttr	evil
	X get hurts	XEffect	x is hospitalized
	X gives Y gifts	XNeed	buy the presents.
	X gives Y gifts	XWant	to hug [Y]
	X gives Y gifts	OEffect	blush
	X gives Y gifts	OWant	open the gift
event-centered commonsense.	accident	Causes	hurt
	X accepts Y's apology	HinderedBy	X is too angry
	why one has to "walk"	XReason	car has broken down
	X does yard work	isAfter	X gets a job as a gardener
	X does yard work	isBefore	X takes a shower
	Move car	HasSubevent	Get out of car

Figure 5: Example generations of models on relations from ATOMIC-2020 dataset (Hwang et al., 2020).

Categories	Head	Relations	COMET (BART)
Physical	bird lover	CapableOf	Look at birds
	mouse	HasProPerty	Long tail
	doctors	Desires	cure patient
Affect	X gives Y gifts	XReact	good about [one]self
Behaviour	X get hurts	XEffect	x is hospitalized
	X gives Y gifts	XNeed	buy the presents.
	X gives Y gifts	XWant	to hug [Y]
	X accepts Y's apology	XIntent	To be forgiving
	X steals a car	XAttr	evil
Events	accident	Causes	hurt
	why one has to "walk"	XReason	car has broken down

Figure 6: Our knowledge graph, which uses 11 relationships and is inspired by psychology, is divided into four modules: Physical, Affect, Behaviour, Events

stopping is applied during training. We use a batch size of 1 and a maximum of 30 decoding steps during testing and inference.