CEM: Commonsense-aware Empathetic Response Generation

Sahand Sabour, Chujie Zheng, Minlie Huang[*]

The CoAI Group, DCST, Institute for Artificial Intelligence,
State Key Lab of Intelligent Technology and Systems,
Beijing National Research Center for Information Science and Technology,
Tsinghua University, Beijing 100084, China
{sahandfer, chujiezhengchn}@gmail.com, aihuang@tsinghua.edu.cn

Abstract

A key trait of daily conversations between individuals is the ability to express empathy towards others, and exploring ways to implement empathy is a crucial step towards human-like dialogue systems. Previous approaches on this topic mainly focus on detecting and utilizing the user's emotion for generating empathetic responses. However, since empathy includes both aspects of affection and cognition, we argue that in addition to identifying the user's emotion, cognitive understanding of the user's situation should also be considered. To this end, we propose a novel approach for empathetic response generation, which leverages commonsense to draw more information about the user's situation and uses this additional information to further enhance the empathy expression in generated responses. We evaluate our approach on EMPATHETICDIALOGUES, which is a widely-used benchmark dataset for empathetic response generation. Empirical results demonstrate that our approach outperforms the baseline models in both automatic and human evaluations and can generate more informative and empathetic responses. Our code is available at https://github.com/Sahandfer/CEM.

1 Introduction

Empathy is a desirable trait of human daily conversations that enables individuals to understand, perceive, and respond appropriately to the situation and feelings of others (Keskin 2014). Previous research has demonstrated that empathetic dialogue systems can improve user experience and satisfaction in multiple domains (Fitzpatrick, Darcy, and Vierhile 2017; Liu et al. 2021; Wang et al. 2021). Hence, it is important to discover ways that allow us to equip open-domain dialogue systems with empathy. Recent work (Rashkin et al. 2019; Lin et al. 2019; Majumder et al. 2020; Li et al. 2020a) has proposed various methods of generating empathetic responses that mainly rely on detecting the user's emotion.

However, empathy is a broad construct that includes aspects of affection and cognition (Davis 1983). The affective aspect is concerned with the emotional simulation in reaction to the user's experiences (Cuff et al. 2016) while the cognitive aspect aims to understand the user's situation and
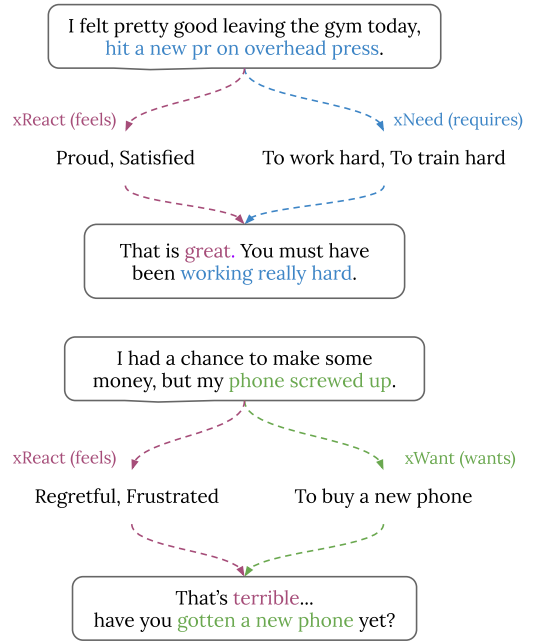
---

[*] Corresponding author.

Figure 1: Examples from the EMPATHETICDIALOGUES dataset in which commonsense is used to gain additional information about the user's emotion and situation before responding empathetically.

the implied feelings (Elliott et al. 2018). Hence, though emotion is one of the important factors of empathy, it is not the only determining factor. This is demonstrated in Figure 1, where both affective and cognitive empathy are used to form empathetic responses. For instance, in the first case, the user shares information about their emotion (*I felt pretty good*) as well as their experience (*hit a new pr on the overhead press*). Accordingly, we can assume that the user is *Proud* of their achievement and must have *worked hard* to reach this level. Since these assumptions are not explicitly mentioned by the user, we as humans tend to rely on our own knowledge and commonsense reasoning to draw these implications. Therefore, we believe that providing dialogue systems with this external knowledge could play a critical role in understanding the user's situation and feelings, which leads to more

informative and empathetic responses.

Towards this end, we propose the **C**ommonsense-aware **E**mpathetic Chatting **M**achine (**CEM**). CEM leverages external commonsense knowledge to obtain more information about the user's situation and feelings (i.e. user's reaction, intention, desire, etc.). Such additional information is used to improve the cognitive understanding and thus, enhance the empathy expression in the generated responses. We evaluate our approach on EMPATHETICDIALOGUES, a widely-used benchmark dataset for empathetic response generation. Both automatic and manual evaluation results demonstrate that compared to previous methods, CEM can generate more informative and empathetic responses.

Our contributions are summarized as follows:

- We propose to leverage commonsense to improve the understanding of interlocutors' situations and feelings, which is an important part of cognitive empathy.
- We introduce CEM, a novel approach that uses various types of commonsense reasoning to enhance empathetic response generation.
- Automatic and manual evaluation demonstrate that with the addition of commonsense, CEM is able to generate more informative and empathetic responses compared with the previous methods.

2 Preliminaries

2.1 Empathetic Dialogue Generation

Empathy is a fairly new term in the literature and therefore, has no specific or widely accepted definition in the fields of social psychology and psychotherapy (Macarov 1978; Elliott et al. 2011). However, empathy is commonly known as a complex multi-dimensional construct that includes broad aspects of affection and cognition (Davis 1983; Zheng et al. 2021). Affective empathy enables us to experience the emotion of others through various emotional stimuli (Cuff et al. 2016), while cognitive empathy enables us to understand the situations and implicit mental states of others, such as intentions, causes, desires, requirements, etc. (Elliott et al. 2018).

In recent years, research on implementing empathy in dialogue systems and generating empathetic responses has gained considerable attention. Initially, Rashkin et al. (2019) demonstrated that detecting the user's emotion is an essential part of generating empathetic responses. Lin et al. (2019) designed a separate decoder for each available emotion and softly combined their outputs. Majumder et al. (2020) proposed that empathetic responses should also mimic the user's emotion to a degree. Li et al. (2020a) leveraged user feedback and proposed a multi-resolution adversarial framework for this task. Recently, (Li et al. 2020b) used commonsense knowledge from ConceptNet (Speer, Chin, and Havasi 2017) to gain a better understanding of the implied emotions within the context. However, these works usually focus on detecting the context emotion and do not pay enough attention to the cognitive aspect of empathy.

2.2 Commonsense and Empathy

As mentioned, a major part of cognitive empathy is understanding the situations and feelings of others. When interacting with a dialogue system, the user is not expected to explicitly share all the information about their situation and how they may feel. As humans, we use our commonsense knowledge to make connections between what is explicitly mentioned and what is implied. Hence, we hypothesize that enabling dialogue systems to leverage commonsense and drive implications from what the user has explicitly shared is highly beneficial for a better understanding of the user's situation and feelings, which leads to more effective cognitive empathy and thus, more empathetic responses.

In this work, we use ATOMIC (Sap et al. 2019) as our commonsense knowledge base. ATOMIC is a collection of commonsense reasoning inferences about everyday if-then events. For each event, ATOMIC infers six commonsense relations for the person involved in the event: the effect of the event on the person (*xEffect*), their reaction to the event (*xReact*), their intent before the event (*xIntent*), what they need in order for the event to happen (*xNeed*), what they would want after the event(*xWant*), and an inferred attribute of the person's characteristics (*xAttr*). Since predicting a person's attributes merely based on a given event would include judging the other person, which is not included in the empathetic process (Peloquin 1995), we neglect *xAttr* in our approach and use the remaining five relations.

In order to generate commonsense inferences for given events, we adopt COMET (Bosselut et al. 2019), which is a pre-trained GPT-2 model (Radford et al. 2018) that is fine-tuned on triplets (e, r, i) from ATOMIC, where e, r, i are the event, the relation type, and the inferred knowledge respectively. More specifically, we use a modified BART-based (Lewis et al. 2020) variation of COMET, which is trained on the ATOMIC-2020 dataset (Hwang et al. 2021). This model is equipped with knowledge that is not readily available to pre-trained language models and is more suitable for inferring knowledge regarding unseen events (Hwang et al. 2021). The latter is necessary for our use-case as many of the events within an empathetic conversation may not occur on a daily basis and therefore, may not exist in the original ATOMIC dataset.

2.3 Task Formulation

We conduct our experiments on the EMPATHETICDIALOGUES (Rashkin et al. 2019), a large-scale multi-turn dataset containing 25k empathetic conversations between crowdsourcing workers. The dataset also provides an emotion label for each conversation from the total 32 available emotions. In this dataset, each conversation is between a speaker and a listener. The task requires a dialogue model to play the role of the listener and generate empathetic responses. Formally, let $D = [u_1, u_2, u_3, ..., u_{k-1}]$ denote a dialogue history of $k - 1$ utterances, where $u_i = [w_1^i, w_2^i, w_3^i, ..., w_{M_i}^i]$ is the i-th utterance that consists of M_i words. Our goal is to generate the listener's next utterance u_k which is coherent to the context, informative, and empathetic to the speaker's situation and feelings.

3 Methodology

Our proposed model, CEM, is built upon the standard Transformer (Vaswani et al. 2017) and its overview is illustrated
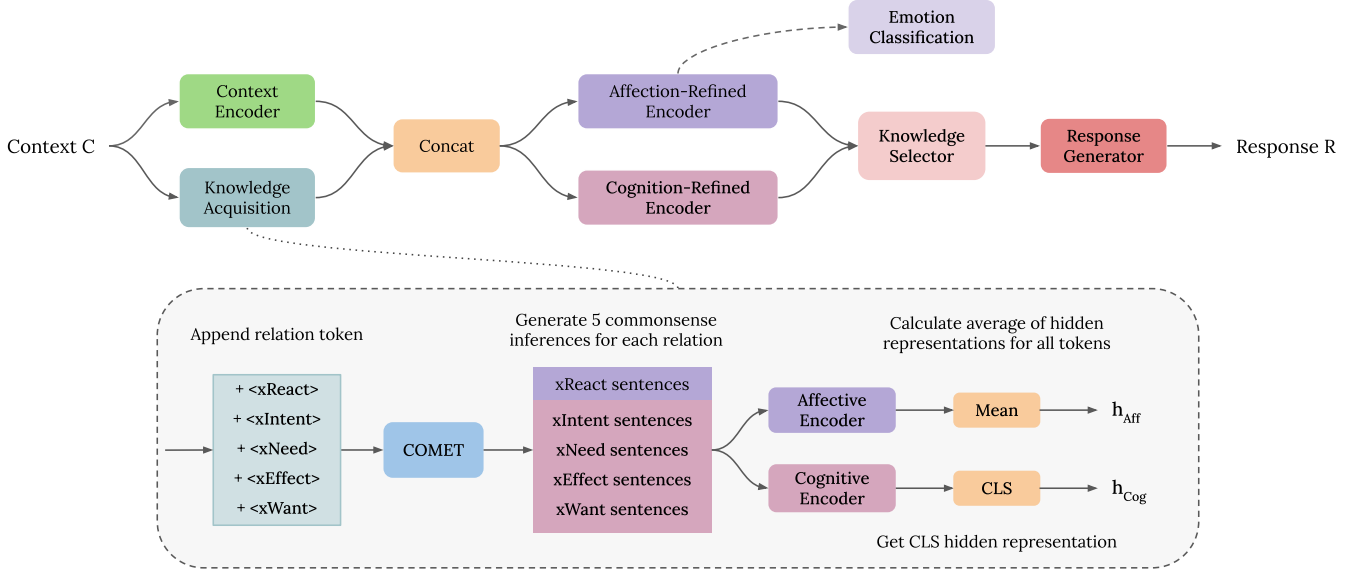
Figure 2: Overview of our model (CEM).

in Figure 2. The process of CEM is mainly divided into five stages: context encoding, knowledge acquisition, context refinement, knowledge selection, and response generation.

3.1 Context Encoding

Following previous work (Lin et al. 2019; Majumder et al. 2020), we concatenate the utterances in the dialogue history and prepend a special token `[CLS]` to obtain the context input $C = \text{[CLS]} \oplus u_1 \oplus u_2 \oplus u_3 \oplus ... \oplus u_{k-1}$, where $\oplus$ is the concatenation operation. Similar to Devlin et al. (2019), we use the final hidden representation of `[CLS]` as the representation of the whole sequence.

We acquire the embedding $\boldsymbol{E}_C$ of the sequence C by summing up the word embedding, positional embedding, and dialogue state embedding. As each utterance in C could be from either the listener or the speaker, we use the dialogue state embedding to distinguish between the two parties. The sequence embedding $\boldsymbol{E}_C$ is then fed to a context encoder to produce the contextual representation:

$$\boldsymbol{H}_{CTX} = \mathbf{Enc}_{CTX}(\boldsymbol{E}_C) \tag{1}$$

where $\boldsymbol{H}_{CTX} \in \mathbb{R}^{L \times d}$, L is the length of the sequence, and d is the hidden size of the context encoder.

3.2 Knowledge Acquisition

For input sequence C, we respectively append five special relation tokens (`[xReact]`, `[xWant]`, `[xNeed]`, `[xIntent]`, `[xEffect]`) to the last utterance in the dialogue history and use COMET to generate five commonsense inferences $[cs_1^r, cs_2^r, ..., cs_5^r]$ per relation r. For each relation, we concatenate the generated commonsense inferences to obtain its commonsense sequence $CS_r = cs_1^r \oplus cs_2^r \oplus ... \oplus cs_5^r$. Given that xReact demonstrates the knowledge regarding the affective state (i.e. user's emotion) while

the other relation represents knowledge regarding the cognitive state (i.e. user's situation), we divide the relations into two groups: affective and cognitive. Accordingly, similar to the previous section, we prepend `[CLS]` to the cognitive sequences. As the inferences for xReact are usually emotion words (e.g. *sad, happy, angry*) rather than sentences, we would simply use the average of its hidden representations to represent these sequences. Based on the mentioned grouping, the resulting sequences are fed to two separate cognitive and affective encoders:

$$\boldsymbol{H}_{xReact} = \mathbf{Enc}_{Aff}(\boldsymbol{E}_{CS_{xReact}}) \tag{2}$$

$$\boldsymbol{H}_r = \mathbf{Enc}_{Cog}(\boldsymbol{E}_{CS_r}) \tag{3}$$

where $\boldsymbol{H}_{xReact} \in \mathbb{R}^{l_{xReact} \times d}$, $\boldsymbol{H}_r \in \mathbb{R}^{l_r \times d}$, with l_{xReact}, l_r being the lengths of the commonsense inference sequences, and $r \in \{xWant, xNeed, xIntent, xEffect\}$.

Then, we use the average hidden representation for affective relations and the hidden representation of `[CLS]` for cognitive relations to represent these sequences respectively:

$$\boldsymbol{h}_{xReact} = \mathrm{Average}(\boldsymbol{H}_{xReact}) \tag{4}$$

$$\boldsymbol{h}_r = \boldsymbol{H}_r[0] \tag{5}$$

where $h_{xReact}, h_r \in \mathbb{R}^d$.

3.3 Context Refinement

Similar to Majumder et al. (2020), in order to refine the context by additional information, we first respectively concatenate each of the commonsense relation representations (Equations 4 & 5) to the context representation $\boldsymbol{H}_{CTX}$ at the token level (i.e. $\boldsymbol{U}_r \in \mathbb{R}^{L \times 2d}$):

$$\boldsymbol{U}_{xReact}[i] = \boldsymbol{H}_{CTX}[i] \oplus \boldsymbol{h}_{xReact} \tag{6}$$

$$\boldsymbol{U}_r[i] = \boldsymbol{H}_{CTX}[i] \oplus \boldsymbol{h}_r \tag{7}$$

In contrast to concatenating the representations at a sequence level (i.e. adding additional information to the end of the context representation), token-level concatenation enables us to fuse the additional knowledge within each word in the sequence.

Accordingly, we use two separate encoders (affection-refined and cognition-refined), corresponding to the two groups of relations, to encode the fused representations and obtain commonsense-refined context representations for each relation respectively:

$$\boldsymbol{H}_{Aff} = \mathbf{Enc}_{CTX-Aff}(\boldsymbol{U}_{xReact}) \qquad (8)$$

$$\boldsymbol{H}_{Cog,r} = \mathbf{Enc}_{CTX-Cog}(\boldsymbol{U}_r) \qquad (9)$$

where $\boldsymbol{H}_{Aff}, \boldsymbol{H}_{Cog,r} \in \mathbb{R}^{L \times d}$.

Emotion Classification In order to acquire a more accurate prediction of the user's affective state, given that we are provided with an emotion label e^* for each conversation, we use the hidden representation of the [CLS] token from the affection-refined context representation ($\boldsymbol{h}_{Aff}$) to perform emotion classification:

$$\boldsymbol{h}_{Aff} = \boldsymbol{H}_{Aff}[0] \qquad (10)$$

where $\boldsymbol{h}_{Aff} \in \mathbb{R}^d$. Hence, we pass $\boldsymbol{h}_{Aff}$ through a linear layer followed by a Softmax operation to produce the emotion category distribution $P_{emo} \in \mathbb{R}^q$, where q is the number of available emotion categories:

$$P_{emo} = \mathrm{Softmax}(\boldsymbol{W}_e \boldsymbol{h}_{Aff}) \qquad (11)$$

where $\boldsymbol{W}_e \in \mathbb{R}^{d \times q}$ is the weight vector for the linear layer. During training, we optimize these weights by minimizing the Cross-Entropy (CE) loss between the emotion category distribution P and the ground truth label e^*:

$$\mathcal{L}_{emo} = -\log(P_{emo}(e^*)) \qquad (12)$$

3.4 Knowledge Selection

Merely using one of the commonsense representations to produce an empathetic response is not ideal. For instance, if we only rely on the affection-refined context, the generated responses would likely be about how the user's emotions (e.g. *You must be proud.*), whereas using the cognition-refined contexts may lead to responses that focus more on the situation (e.g. *You must have worked really hard.*). Hence, we want to enable our model to generate responses based on the mixture of both affective and cognitive information. To this end, we first concatenate all the five relation-refined contexts at the token level:

$$\boldsymbol{H}_{Cog}[i] = \bigoplus_{r \in \{xWant, xNeed, xIntent, xEffect\}} \boldsymbol{H}_{Cog,r}[i] \qquad (13)$$

$$\boldsymbol{H}_{Refine}[i] = \boldsymbol{H}_{Aff}[i] \oplus \boldsymbol{H}_{Cog}[i] \qquad (14)$$

where $\boldsymbol{H}_{Refine} \in \mathbb{R}^{L \times 5d}$. To highlight the more important features within the refined context representation, we apply the Sigmoid function on $\boldsymbol{H}_{Refine}$ to measure the importance of each relation-refined context for response generation. Then, we multiply $\boldsymbol{H}_{Refine}$ by the consequent importance scores, as done by Majumder et al. (2020). Finally, the

obtained representation is passed through a Multi-Layer Perceptron (MLP) with ReLU activation, which learns how to mix the commonsense knowledge of different relations into a combined contextualized representation:

$$\widetilde{\boldsymbol{H}}_{CTX} = \mathbf{MLP}(\sigma(\boldsymbol{H}_{Refine}) \odot \boldsymbol{H}_{Refine}) \qquad (15)$$

where $\widetilde{\boldsymbol{H}}_{CTX} \in \mathbb{R}^{L \times d}$ and $\odot$ denotes element-wise multiplication.

3.5 Response Generation

Lastly, the target response $Y = [y_1, \ldots, y_T]$ with length T, which has the same meaning of u_k in Section 2.3, is generated by the decoder token by token:

$$P(y_t|y_{<t}, C) = \mathbf{Dec}(\boldsymbol{E}_{y_{<t}}, \widetilde{\boldsymbol{H}}_{CTX}) \qquad (16)$$

where $\boldsymbol{E}_{y_{<t}}$ denotes the embeddings of the tokens that have been generated. Note that the cross attention to the encoder outputs is modified to the commonsense-refined contextual representation $\widetilde{\boldsymbol{H}}_{CTX}$, which has fused the information from both the context and the commonsense inferences.

3.6 Training Objectives

We adopt the standard negative log-likelihood (NLL) loss on the target response Y:

$$\mathcal{L}_{nll} = -\sum_{t=1}^{T} \log P(y_t|C, y_{<t}) \qquad (17)$$

Response Diversity In our preliminary experiments, we noticed that the models trained on our studied dataset tend to generate similarly generic empathetic responses. As shown in Table 1, there are phrases that are extensively repeated within the model responses in this dataset. Hence, similar to the problem raised by Li et al. (2016) for generic response generation in Seq2Seq models, we believe that models trained on this dataset tend to assign a higher probability to responses that include these phrases and thus, generate safe empathetic responses (e.g. *I am sorry to hear that*, *That is good to hear*, and *Oh no that is awful*). We consider these responses safe and generic as they do not necessarily rely on nor give much information about the user's context and can be employed in many different situations.

Phrases	Prop. (%)
That is a / Oh no that / To hear that	≥ 67
I am so / Sorry to hear / Wow that is	≥ 50
That is really / I am sure / That is great	≥ 40

Table 1: Most common trigrams in the training set of EMPATHETICDIALOGUES. Proportion represents the number of responses that include the trigram divided by the total number of responses (e.g. more than 50% of the responses include the trigram *Sorry to hear*).

To tackle this issue, we implement 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 and prior to receiving a new batch of samples, we first calculate the relative frequency RF_i for each vocabulary token c_i in the training corpus:

$$RF_i = \frac{\text{freq}(c_i)}{\sum_{j=1}^{V} \text{freq}(c_i)} \qquad (18)$$

where V denotes the vocabulary size. Accordingly, we derive the frequency-based weight w_i as follows:

$$w_i = a \times RF_i + 1 \qquad (19)$$

where $a = -(\max_{1 \leq j \leq V}(RF_j))^{-1}$ is the frequency slope and 1 is added as the bias so that w_i falls into $[0, 1]$. Since more frequent tokens would have a higher relative frequency, the obtained weights ensure that these tokens have lower weights. Lastly, we normalize w_i to have a mean of 1, as done by Jiang et al. (2019). The diversity loss would then be calculated as below:

$$\mathcal{L}_{div} = -\sum_{t=1}^{T}\sum_{i=1}^{V} w_i \delta_t(c_i) \log P(c_i | y_{<t}, C) \qquad (20)$$

where c_i is a candidate token in the vocabulary and $\delta_t(c_i)$ is the indicator function, which equals to 1 if and only if $c_i = y_t$ and 0 otherwise. All the parameters for our proposed model are trained and optimized based on the weighted sum of the three mentioned losses.

$$\mathcal{L} = \gamma_1 \mathcal{L}_{nll} + \gamma_2 \mathcal{L}_{emo} + \gamma_3 \mathcal{L}_{div} \qquad (21)$$

where γ_1, γ_2, and γ_3 are hyper-parameters that we use to control the influence of the three losses. In our experiments, we set $\gamma_1 = 1$, $\gamma_2 = 1$, and $\gamma_3 = 1.5$. During our analysis, we found that setting the same coefficients for all losses did not produce sufficient penalties for the generic responses. Hence, we assigned a slightly higher value to γ_3.

4 Experiments

4.1 Baselines

We selected the following baseline models for comparison:

- **Transformer** (Vaswani et al. 2017): The original Transformer, which is trained to optimize the NLL loss ($\mathcal{L}_{nll}$).
- **Multi-Task Transformer (Multi-TRS)** (Rashkin et al. 2019): A variation of the Transformer that has an additional unit for predicting the emotion. This model is trained to jointly optimize the NLL loss ($\mathcal{L}_{nll}$) and the cross-entropy loss for emotion classification ($\mathcal{L}_{emo}$).
- **MoEL** (Lin et al. 2019): A Transformer-based model that uses a decoder for each possible user emotion, referred to as *listener*, and softly combines the representations from these decoders to generate a response. Therefore, each decoder is optimized to learn how to respond to one type of emotion while a meta decoder is optimized to combine their representations and generate a response.
- **MIME** (Majumder et al. 2020): Another Transformer-based model that mimics the detected user emotion to a degree. In this approach, the emotions are separated into

negative and positive emotions. The model initially generates mimicking and non-mimicking representations for the response based on the emotion groups and is optimized to effectively blend these representations and generate an empathetic response.
- **EmpDG** (Li et al. 2020a): A multi-resolution adversarial framework that consists of an empathetic generator and an interactive discriminator. The generator produces empathetic responses based on the detected emotion while the discriminator ensures that the generated responses are consistent with the context and are also empathetic. To provide a fair comparison with our model and the other baselines, we only implement the empathetic generator in our experiments as the discriminator requires information from the future turns within the conversation.

4.2 Implementation Details

We implemented all the models using PyTorch[1] and used 300-dimensional pre-trained GloVE vectors (Pennington, Socher, and Manning 2014) to initialize the word embeddings, which were shared between the encoders and the decoders. The hidden dimension for all corresponding components were set to 300. Adam (Kingma and Ba 2017) optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ was used for training. The initial learning rate was set to 0.0001 and we varied this value during training according to Vaswani et al. (2017). All the models were trained on one single TITAN Xp GPU using a batch size of 16 and early stopping. We used a batch size of 1 and a maximum of 30 decoding steps during testing and inference. We used the same 8:1:1 train/valid/test split as provided by Rashkin et al. (2019).

4.3 Automatic Evaluation

We employed Perplexity (**PPL**) and Distinct-n (**Dist-n**) (Li et al. 2016) as our main automatic metrics. PPL represents the model's confidence in its set of candidate responses, with higher confidence resulting in a lower PPL. This can be used to evaluate the general quality of the generated responses. Dist-n measures the proportion of unique n-grams in the generated responses and is commonly used to evaluate generation diversity. In addition, since our proposed model and the baselines models (except Transformer) all perform emotion classification as a part of their training process, we also report the prediction accuracy (**Acc**). As Liu et al. (2016) had found that word overlap-based automatic metrics such as BLEU (Papineni et al. 2002) are not appropriate for evaluating dialogue systems, we do not report such metrics.

Table 2 shows the automatic evaluation results. CEM achieves the lowest perplexity, which suggests the overall quality of our generated responses is higher than the baselines. In addition, our model also considerably outperforms the baselines in terms of Dist-n, which highlights the importance of the diversity loss. In terms of emotion classification, CEM had a much higher accuracy compared to the baselines, which suggests the addition of commonsense knowledge is also beneficial for detecting the user's emotion.

---

[1] https://pytorch.org/

Models	PPL	Dist-1	Dist-2	Acc (%)
Transformer	37.62	0.45	2.02	-
Multi-TRS	37.75	0.41	1.67	33.57
MoEL	36.93	0.44	2.10	30.62
MIME	37.09	0.47	1.90	31.36
EmpDG	37.29	0.46	2.02	30.41
CEM	36.11	**0.66**	**2.99**	**39.11**
w/o Aff	36.49	0.56	2.52	33.76
w/o Cog	36.63	0.56	2.47	36.42
w/o Div	**35.60**	0.48	1.96	38.82

Table 2: Results of automatic evaluation. The best results among all models are highlighted in **bold**.

4.4 Human Evaluation

In previous work, human evaluation was conducted via two tasks: first, crowdsourcing workers were asked to assign a score from 1 to 5 to the generated responses based on the aspects of fluency, relevancy, and empathy; second, they were required to choose the better response between two models within the same context. However, the criteria for giving a score from 1 to 5 is highly likely to vary between different individuals, which results in low inter-annotator agreement and is not a suitable indicator of a model's performance. In addition, asking workers to choose the better responses without any guidelines and solely relying on their own preference is not satisfactory. This is due to the fact that each person may consider aspects that are different from what is being investigated when making their choices, which is also not a reliable indicator of user preference.

Comparisons	Aspects	Win	Lose	κ
CEM vs. MoEL	Coh.	53.6‡	37.6	0.57
	Emp.	52.0‡	38.0	0.57
	Inf.	61.0‡	30.6	0.51
CEM vs. MIME	Coh.	52.0‡	42.3	0.44
	Emp.	50.3‡	41.6	0.57
	Inf.	48.6	45.0	0.51
CEM vs. EmpDG	Coh.	46.3†	42.6	0.52
	Emp.	54.3‡	33.3	0.51
	Inf.	47.6†	43.3	0.41

Table 3: Human evaluation results (%). Ties are not shown. κ denotes the inter-annotator agreement measured by Fleiss's kappa, where $0.4 < \kappa < 0.6$ indicates moderate agreement. $\dagger,\ddagger$ represent significant improvement with p-value $< 0.1/0.05$ respectively (sign test).

To address these issues, we conducted an aspect-based pairwise preference test. That is, for a given context, we paired our model's response with a response from the baselines and asked annotators to choose the better response based on the context and the following three aspects: 1) *Coherence* (**Coh.**): which response is more coherent in con-

tent and relevant to the context; 2) *Empathy* (**Emp.**): which response shows more understanding of the user's situation and presents a more appropriate emotion; 3) *Informativeness* (**Inf.**): which response conveys more information about the context. Then, we randomly sampled 100 response pairs and assigned three crowdsourcing workers to annotate each pair. Ties were allowed but the annotators were encouraged to choose one of the responses.

As shown in Table 3, CEM outperforms the baselines in all of the three aspects. Particularly, with the enhancement of commonsense knowledge, our model was able to produce responses that conveyed more specific and informative content and thus were more empathetic. We also note CEM did not significantly outperform MIME in *informativeness*. Upon further investigation, we realized that on average, MIME tends to generate longer responses (12.8 words / response) compared to CEM (9.6 words / response). It is possibly due to some annotators considering these responses as more informative since they included more words. However, as shown by the results of the automatic evaluation (Table 2), we can observe that MIME has the second-lowest Dist-2 score, which suggests that its generated responses may follow similar patterns and have less diversity.

4.5 Ablation Studies

We conducted ablation studies to verify the effectiveness of each of the components in our model. Specifically, we designed three variants of CEM: **1) w/o Aff:** the affective and affection-refined encoders are removed (Equations 2 & 8), the affect representation is neglected in the commonsense-refined representation (Equation 14), and the hidden representation of the [CLS] token from the encoded context (Equation 1) is used for emotion classification; **2) w/o Cog:** the cognitive and cognition-refined encoders are removed (Equations 3 & 9), the cognition representation is neglected in the commonsense-refined representation (Equation 14), and the MLP is replaced with a linear layer (Equation 15); **3) w/o Div:** the diversity loss is removed from the training objectives (Equation 21).

The obtained results are shown in Table 2. Both affective and cognitive information have a considerable impact on the emotion classification accuracy, which suggests that information about both the user's emotion and their situation are necessary for correctly identifying their feelings. In addition, we also observed that removing the diversity loss results in considerably lower Dist-n scores, which indicates the effectiveness of this loss in generating more diverse responses.

4.6 Case Study

Table 4 shows comparisons between the generated responses of CEM and the three main baselines. In the first case, the baselines fail to realize the meaning behind *ready for a puppy*, which implies that the user wants to buy or adopt a puppy. It can be observed that MoEL dismisses this implication while the other two baselines mistake the meaning behind the phrase for *being ready for an event or exam*, which may cause the user to be *proud* of themselves. By accessing external knowledge, CEM better acknowledges the

Emotion	Excited
Context	I decided a few months I was ready for a puppy, so I put the word out.
MoEL	That is good, I am glad you did not get a dog.
MIME	That is so sweet! I bet you are so proud!
EmpDG	That is awesome! I bet you are very proud of you!
Affect	Happy, **Excited**, Hopeful
Cognition	To have a pet, **To get a dog**, To have a companion
CEM	That is **great**! Did you **get a good dog**?
Gold	Put what word? You mean published the info?
Emotion	Sentimental
Context	Every time I look through older photos of my children, I am taken right back to the time they were that size, and what we were doing ! It is a bittersweet feeling, for sure.
MoEL	I am sorry to hear that, I am sorry to hear that.
MIME	That is so sweet. I am sure you miss them.
EmpDG	I am so happy for you.
Affect	**Nostalgic**, Happy, Sad
Cognition	**To reminisce memories**, Look at old photos, To remember the past
CEM	Yeah, it is always **nice to have those memories.**
Gold	Oh, it sure is, I feel the same way. They grow up so fast.
Emotion	Guilty
Context	*Speaker:* Yeah about 10 years ago I had a horrifying experience. It was 100% their fault but they hit the water barrels and survived. They had no injuries but they almost ran me off the road. *Listener:* Did you suffer any injuries? *Speaker:* No I was not hit. It turned out they were drunk. I felt guilty but realized it was his fault.
MoEL	That is so scary! I hope you are okay!
MIME	I am sorry to hear that. I hope you can find someone else to do that.
EmpDG	Oh no! I am sorry to hear that. I hope you were able to get it fixed.
Affect	Remorseful, **Guilty**, Sorry
Cognition	To be safe, **No harm done**, **No one hit**
CEM	I am glad you **did not get hurt**.
Gold	Why did you feel guilty? People really shouldn't drive drunk.

Table 4: Case study of the generated responses by CEM and the baselines.

user's situation and implied feelings and generates an empathetic response that covers both aspects of empathy. That is, by detecting that the user might be *excited* and may want to *get a dog*, it responds with both affective (*that is great*) and cognitive (*did you get a good dog?*) statements.

Similarly, in the second case, unlike the baselines, CEM successfully detects that the user is being nostalgic, happy and sad, where the latter two emotions are likely to be implied in the word *bittersweet*. In addition, CEM realizes that the user's intent behind looking through photos of their children was to *reminisce memories*, which suggests that the user enjoys having those memories.

The final case demonstrates CEM's ability to express both affective and cognitive empathy in multi-turn dialogue. As shown, all the baselines dismiss the user's statement *I was not hit*, which implies that they are fine and no harm was done. In contrast, CEM correctly recognizes that there is no

harm done to the user and regardless of detecting that the user might have remorse and guilt, it chooses to focus more on the important part of this situation, which is the user's health and safety.

5 Conclusions and Future Work

In this paper, we proposed the Commonsense-aware Empathetic Chatting Machine (CEM) to demonstrate how leveraging commonsense knowledge could benefit the understanding of the user's situation and feelings, which leads to more informative and empathetic responses. Our empirical automatic and manual evaluation indicated that the effectiveness of our approach in empathetic response generation.

In the future, our work can inspire other approaches to leverage commonsense knowledge for empathetic response generation and similarly promising tasks (e.g. providing emotional support (Liu et al. 2021)).

Acknowledgements

References

Bosselut, A.; Rashkin, H.; Sap, M.; Malaviya, C.; Celikyilmaz, A.; and Choi, Y. 2019. COMET: Commonsense Transformers for Automatic Knowledge Graph Construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 4762–4779. Florence, Italy: Association for Computational Linguistics.

Cuff, B. M.; Brown, S. J.; Taylor, L.; and Howat, D. J. 2016. Empathy: A Review of the Concept. *Emotion Review*, 8(2): 144–153.

Davis, M. H. 1983. Measuring individual differences in empathy: Evidence for a multidimensional approach. *Journal of Personality and Social Psychology*, 44(1): 113–126.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Elliott, R.; Bohart, A.; Watson, J.; and Greenberg, L. 2011. Empathy. *Psychotherapy relationships that work (2nd ed.)*, 132–152.

Elliott, R.; Bohart, A. C.; Watson, J. C.; and Murphy, D. 2018. Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*, 55(4): 399–410.

Fitzpatrick, K. K.; Darcy, A.; and Vierhile, M. 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).

Hwang, J. D.; Bhagavatula, C.; Le Bras, R.; Da, J.; Sakaguchi, K.; Bosselut, A.; and Choi, Y. 2021. (Comet-) Atomic 2020: On Symbolic and Neural Commonsense Knowledge Graphs. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7): 6384–6392.

Jiang, S.; Ren, P.; Monz, C.; and de Rijke, M. 2019. Improving Neural Response Diversity with Frequency-Aware Cross-Entropy Loss. In *The World Wide Web Conference*, WWW '19, 2879–2885. New York, NY, USA: Association for Computing Machinery. ISBN 9781450366748.

Keskin, S. C. 2014. From what isn't Empathy to Empathic Learning Process. *Procedia - Social and Behavioral Sciences*, 116: 4932–4938. 5th World Conference on Educational Sciences.

Kingma, D. P.; and Ba, J. 2017. Adam: A Method for Stochastic Optimization. arXiv:1412.6980.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Online: Association for Computational Linguistics.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 110–119. San Diego, California: Association for Computational Linguistics.

Li, Q.; Chen, H.; Ren, Z.; Ren, P.; Tu, Z.; and Chen, Z. 2020a. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4454–4466. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Li, Q.; Li, P.; Chen, Z.; and Ren, Z. 2020b. Towards Empathetic Dialogue Generation over Multi-type Knowledge. arXiv:2009.09708.

Lin, Z.; Madotto, A.; Shin, J.; Xu, P.; and Fung, P. 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)*, 121–132. Hong Kong, China: Association for Computational Linguistics.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Liu, S.; Zheng, C.; Demasi, O.; Sabour, S.; Li, Y.; Yu, Z.; Jiang, Y.; and Huang, M. 2021. Towards Emotional Support Dialog Systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 3469–3483. Online: Association for Computational Linguistics.

Macarov, D. 1978. Empathy: The charismatic chimera. *Journal of Education for Social Work*, 14(3): 86–92.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. MIME: MIMicking Emotions for Empathetic Response Generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8968–8979. Online: Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, 311–318. USA: Association for Computational Linguistics.

Peloquin, S. M. 1995. The Fullness of Empathy: Reflections and Illustrations. *American Journal of Occupational Therapy*, 49(1): 24–31.

Pennington, J.; Socher, R.; and Manning, C. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. Doha, Qatar: Association for Computational Linguistics.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving Language Understanding by Generative Pre-Training.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 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*, 5370–5381. Florence, Italy: Association for Computational Linguistics.

Sap, M.; Le Bras, R.; Allaway, E.; Bhagavatula, C.; Lourie, N.; Rashkin, H.; Roof, B.; Smith, N. A.; and Choi, Y. 2019. ATOMIC: An Atlas of Machine Commonsense for If-Then Reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 3027–3035.

Speer, R.; Chin, J.; and Havasi, C. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 4444–4451. AAAI Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L. u.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wang, L.; Wang, D.; Tian, F.; Peng, Z.; Fan, X.; Zhang, Z.; Ma, S.; Yu, M.; Ma, X.; and Wang, H. 2021. CASS: Towards Building a Social-Support Chatbot for Online Health Community. arXiv:2101.01583.

Zheng, C.; Liu, Y.; Chen, W.; Leng, Y.; and Huang, M. 2021. CoMAE: A Multi-factor Hierarchical Framework for Empathetic Response Generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 813–824. Online: Association for Computational Linguistics.