CASE: Aligning Coarse-to-Fine Cognition and Affection for Empathetic Response Generation

Jinfeng Zhou[1,2], Chujie Zheng[1], Bo Wang[2], Zheng Zhang[1], Minlie Huang[1*]

[1]The CoAI Group, DCST, Institute for Artificial Intelligence,
[1]State Key Lab of Intelligent Technology and Systems,
[1]Beijing National Research Center for Information Science and Technology,
[1]Tsinghua University, Beijing 100084, China
[2]College of Intelligence and Computing, Tianjin University, Tianjin, China
jfzhou@tju.edu.cn, chujiezhengchn@gmail.com, aihuang@tsinghua.edu.cn

Abstract

Empathy is a trait that naturally manifests in human conversation. Theoretically, the birth of empathetic responses results from conscious alignment and interaction between cognition and affection of empathy. However, existing works rely solely on a single affective aspect or model cognition and affection independently, limiting the empathetic capabilities of the generated responses. To this end, based on the commonsense cognition graph and emotional concept graph constructed involving commonsense and concept knowledge, we design a two-level strategy to align coarse-grained (between contextual cognition and contextual emotional state) and fine-grained (between each specific cognition and corresponding emotional reaction) **C**ognition and **A**ffection for re**S**ponding **E**mpathetically (**CASE**). Extensive experiments demonstrate that CASE outperforms the state-of-the-art baselines on automatic and human evaluation. **Our code will be released.**

Introduction

Human conversations naturally involve empathetic interactions, which allow both parties to recognize and understand each other's experiences and feelings from the other's shoes (Keskin 2014). As a trait of human conversation, empathy is a crucial factor in establishing seamless relationships (Zech and Rimé 2005). Previous researches (Liu et al. 2021; Wang et al. 2021) attempt to cultivate empathy in dialogue systems confirmed that empathy is also integral to building a warm conversational AI (Huang, Zhu, and Gao 2020).

In social psychology, empathy is commonly known as consisting of two aspects, i.e., cognition and affection (Davis 1983). The cognitive aspect aims to understand the user's current situation (Cuff et al. 2016). The affective aspect involves the emotional simulation in reaction to the observed user's experiences (Elliott et al. 2018). Although both aspects are considered to generate empathetic responses, existing works still remain issues. On the one hand, most works (Rashkin et al. 2019; Lin et al. 2019; Majumder et al. 2020; Li et al. 2020, 2022) rely solely on the affective aspect to detect and utilize contextual emotion for enhancing empathy in response generation. On the other hand, previous researches (Zheng et al. 2021; Sabour, Zheng, and Huang 2022) model cognition and affection as two relatively independent aspects to improve the understanding and expression of empathy.

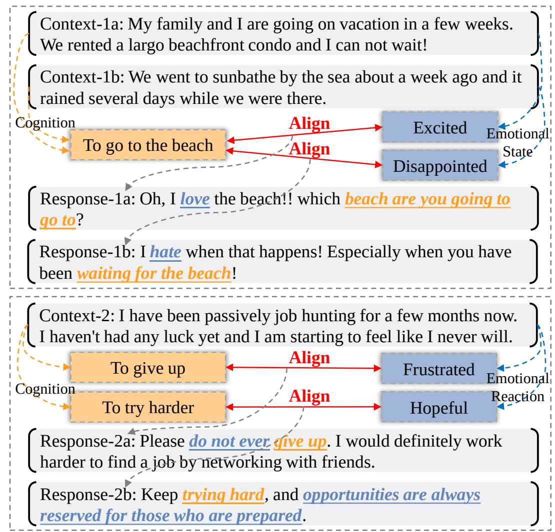

Figure 1: Examples from the EMPATHETICDIALOGUES dataset. The alignment of cognition and affection (i.e., emotional state and emotional reaction) leads to highly empathetic expression in responses.

However, the birth of empathetic responses in human conversation often results from conscious alignment and interaction between cognition and affection of empathy (Westbrook, Kennerley, and Kirk 2011). For one thing, the emotional state manifested by different contexts affects expression about cognitive situations in responses. The alignment of cognition and affection facilitates the response to express empathetic cognition under the appropriate emotional state. As in case-1 of Fig 1, the alignment of cognitive situation, i.e., *intent* "*to go to the beach*", with emotional states, i.e., "*excited*" vs. "*disappointed*", produces empathetic expressions that satisfy different contexts, i.e., "*love*" and "*which beach are you going to go to*" vs. "*hate*" and "*waiting for the beach*". For another, different cognitive situations give rise to different emotional reactions. The alignment of cognition and affection encourages building the association between cognition and emotional reaction to generate highly empathetic responses. As in case-2 of Fig 1, building distinct patterns of association between cognition and emotional reac-

*Corresponding author.

tion, i.e., "*to give up*" and "*frustrated*" vs. "*to try harder*" and "*hopeful*", yields cognitively distinct but both highly empathetic responses, i.e., *response-2a* vs. *response-2b*. The above cases highlight the necessity of aligning cognition and affection for empathy modeling in response generation.

To this end, we propose to align **C**ognition and **A**ffection for re**S**ponding **E**mpathetically (**CASE**), and integrate commonsense knowledge from COMET (Bosselut et al. 2019) and concept knowledge from ConceptNet (Speer, Chin, and Havasi 2017). Commonsense knowledge infers the user's situation as cognition and emotional reaction to the situation, which are implied in the dialogue. Concept knowledge serves to extract the emotional state manifested in the dialogue context. For encoding two knowledge, we first construct two heterogeneous graphs, i.e., commonsense cognition graph and emotional concept graph, where the initial independent representation of cognition and emotional concept is carefully adjusted by dialogue context adopting graph transformers. Then, we design a two-level strategy to align cognition and affection using mutual information maximization (MIM) (Hjelm et al. 2019). The coarse-grained level considers overall cognition and affection manifested in the dialogue context to align contextual cognition and contextual emotional state, which are extracted with a knowledge discernment mechanism. The fine-grained level builds the fine-grained association between cognition and affection implied in the dialogue to align each specific cognition and corresponding emotional reaction. Further, an empathy-aware decoder is devised for generating empathetic expressions.

Our contributions are summarized as follows: (1) We devise a unified framework to model the interaction between cognition and affection for integrated empathetic response generation. (2) We construct two heterogeneous graphs involving commonsense and concept knowledge to aid in the modeling of cognition and affection. (3) We propose a two-level strategy to align coarse-grained and fine-grained cognition and affection adopting mutual information maximization. (4) Extensive experiments demonstrate that the superior of CASE in terms of automatic and human evaluation.

Related Work

Emotional & Empathetic Conversation

Emotional conversation gives the manually specified label preset as the emotion generated in the response (Zhou et al. 2018; Wei et al. 2019; Peng et al. 2022). Instead of giving a predefined emotion label, empathetic conversation involves cognitive and affective empathy (Davis 1983; Zheng et al. 2021) and aims to fully understand the interlocutor's situation and feelings and respond empathically (Keskin 2014). On the one hand, most existing works only focus on the affective aspect of empathy and make efforts to detect contextual emotion (Rashkin et al. 2019; Lin et al. 2019; Majumder et al. 2020; Li et al. 2020, 2022) while ignoring the cognitive aspect. On the other hand, some research leverage commonsense as cognition to refine empathetic considerations (Sabour, Zheng, and Huang 2022). However, the relatively independent modeling between the two aspects (i.e., cognition and affection) violates their interrelated characteristics.

Commonsense & Concept Knowledge

As a commonsense knowledge base, ATOMIC (Sap et al. 2019) focuses on inferential knowledge organized as typed *if-then* relations. Six commonsense reasoning relations are defined for the person involved in an event, four of which are used to reason commonsense cognitions of a given event, i.e., PersonX's intent before the event (*xIntent*), what PersonX need to do before the event (*xNeed*), what PersonX want after the event (*xWant*), and the effect of the event on PersonX (*xEffect*). Each commonsense cognition is aligned with user's emotional reaction to the situation implied in dialog inferred by *xReact* (i.e., PersonX's reaction to the event) in our approach. To obtain inferential commonsense knowledge, we use COMET (Bosselut et al. 2019), a pretrained generative model, to generate rich commonsense statements.

As the concept knowledge, ConceptNet (Speer, Chin, and Havasi 2017) provides word-level human knowledge and is widely used in various NLP tasks (Zhang et al. 2020; Zhong et al. 2021). Following Li et al. (2022), we use NRC_VAD (Mohammad 2018) to assign emotion intensity to concepts in ConceptNet (processing details are in Li et al. (2022)) severed to extract the contextual emotional state manifested in the dialogue context, and align it with contextual cognition.

Mutual Information Maximization

Mutual information maximization (MIM) aims to measure the dependence between two random variables X and Y, and the mutual information (MI) between them is defined as: $MI(X,Y) = D_{KL}(P(X,Y)\|P(X)P(Y))$. However, maximizing MI directly is normally intractable. A successful practice to estimate MI with a lower bound is InfoNCE (Kong et al. 2020). Given two different views x and y of an input, InfoNCE is defined by:

$$\mathbb{E}_{P(X,Y)}[f_\theta(x,y) - \mathbb{E}_{Q(\widetilde{Y})}[\log \sum_{\widetilde{y}\in\widetilde{Y}} \exp f_\theta(x,\widetilde{y})]] + \log|\widetilde{Y}|, \tag{1}$$

where f_θ is a learnable function with parameter θ. The set $\widetilde{Y}$ draws samples from a proposal distribution $Q(\widetilde{Y})$, and it comprises $|\widetilde{Y}|-1$ negative samples and a positive sample y. One insight is that when $\widetilde{Y}$ always consists all values of Y and they are uniformly distributed, maximizing InfoNCE is analogous to maximize cross-entropy loss:

$$\mathbb{E}_{P(X,Y)}[f_\theta(x,y) - \log \sum_{\widetilde{y}\in Y} \exp f_\theta(x,\widetilde{y})]. \tag{2}$$

It shows InfoNCE is relevant to maximize $P_\theta(y|x)$ and approximates summation over elements in Y (i.e., partition function) by negative sampling (Zhou et al. 2020). Upon the formula, we replace X and Y with specific cognition and affection to maximize MI between them.

Approach

Architecture Overview

The overall framework of CASE is shown in Figure 2. The dialogue context $X = [x_1, \dots, x_N]$ contains N utterances, where x_i denotes the i-th utterance. CASE contains three stages: (1) The graph encoding stage constructs

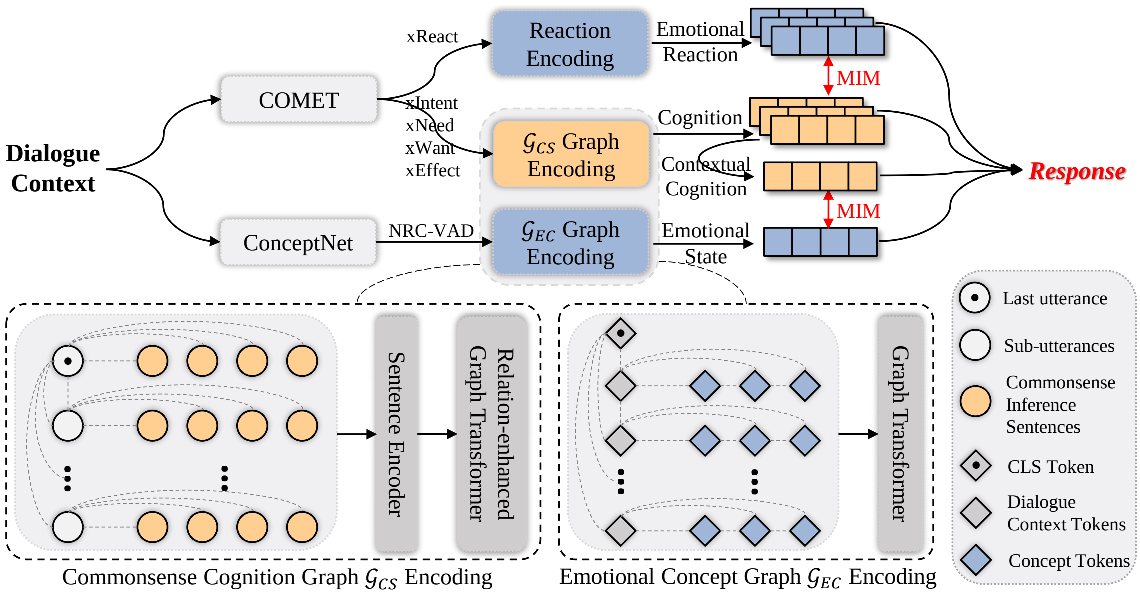

Figure 2: The architecture of the proposed CASE model.

and encodes heterogeneous commonsense cognition graph $\mathcal{G}_{CS}$ and emotional concept graph $\mathcal{G}_{EC}$ from the dialogue context X. (2) The coarse-to-fine alignment aligns coarse-grained (between contextual cognition and contextual emotional state) and fine-grained (between each specific cognition and corresponding emotional reaction) cognition and affection adopting MIM. (3) The empathy-aware decoder integrates the aligned cognition and affection to generate the response $Y = [y_1, y_2, \ldots, y_M]$ with empathetic and informative expressions.

Graph Encoding

Commonsense Cognition Graph Construction Given the last utterance x_N of the dialogue context X, we segment it into the sub-utterances $U = [u_0, u_1, u_2, \ldots, u_t]$, where we prepend the whole x_N as u_0 for maintaining the global information of x_N. We use COMET to infer l commonsense cognition knowledge $K_i^r = [k_{i,1}^r, k_{i,2}^r, \ldots, k_{i,l}^r]$ for each $u_i \in U$, where r is one of the four commonsense relations $\mathcal{R} = \{\text{xIntent}, \text{xNeed}, \text{xWant}, \text{xEffect}\}$, similar to Sabour, Zheng, and Huang (2022). The idea is that human responses tend to inherit the above and transfer the topic. There are differences in the topic and connotation of different sub-utterances, which also affect the listeners' concerns when making empathetic responses.

For constructing the heterogeneous commonsense cognition graph $\mathcal{G}_{CS}$, we use the utterance set U and the commonsense cognition knowledge set $K_{CS} = \bigcup_{i=0}^{t} \bigcup_{r \in \mathcal{R}} K_i^r$ as vertices, i.e., vertex set $V_{CS} = U \cup K_{CS}$. There are seven relations of undirected edges that connect vertices. (1) The *self-loop* relation for each vertex $v_i^{CS} \in V_{CS}$. (2) The *global* relation between the whole x_N (i.e., u_0) and its sub-utterances $u_i (i \geq 1)$. (3) The *temporary* relation between any two successive sub-utterances u_j and u_{j+1}. (4) The four commonsense relations, i.e., *xIntent*, *xNeed*, *xWant*, *xEffect*, between the utterance $u_i \in U$ and the corresponding K_i^r.

We use a Transformer-based sentence encoder (cognition encoder) to first encode the vertices V_{CS} of the graph $\mathcal{G}_{CS}$. For each $v_i^{CS} \in V_{CS}$, we prepend with a special token $[CLS]$. Following Devlin et al. (2019), we collect the representation derived from $[CLS]$ as the initial embedding matrix for $\mathcal{G}_{CS}$.

Emotional Concept Graph Construction We concatenate the utterances in the dialogue context X to obtain the token set T, i.e., $T = x_1 \oplus \ldots \oplus x_N = [w_1, \ldots, w_n]$, where n is the number of all the tokens in the utterances in X. Following Li et al. (2022), we use ConceptNet to infer the related concepts for each token $w_i \in T$, among which only the the top N' emotional concepts (according to the emotion intensity $\omega(c)$) are used for constructing $\mathcal{G}_{EC}$. Subsequently, the vertices V_{EC} in the heterogeneous emotional concept graph $\mathcal{G}_{EC}$ contains a $[CLS]$ token, the dialogue context tokens T, and the above obtained emotional concepts. There are four relations of undirected edges that connect vertices. (1) The *self-loop* relation for each vertex $v_i^{EC} \in V_{EC}$. (2) The *global* relation between the $[CLS]$ token and other ones. (3) The *temporary* relation between any two successive tokens. (4) The *emotional concept* relation among a token and its related emotional concepts.

We initialize the vertex embedding matrix for $\mathcal{G}_{EC}$ by summing up the token embedding, the positional embedding, and the type embedding for each vertex (signaling whether it is a emotional concept or not).

Graph Encoder Given the commonsense cognition graph $\mathcal{G}_{CS}$, to capture the semantic relationship between vertices, we adopt the Relation-Enhanced Graph Transformer (Li et al. 2021) for graph encoding. It employs a relation-enhanced multi-head attention mechanism (MHA) to encode vertex embedding $\hat{\boldsymbol{v}}_{v_i}$ for vertex v_i (we omit the superscripts

CS for simplicity) as:

$$\hat{\boldsymbol{v}}_{v_i} = MHA_{v_k \in V_{CS}} \left(\boldsymbol{q}_{v_i}, \boldsymbol{k}_{v_k}, \boldsymbol{v}_{v_k} \right), \tag{3}$$

where the semantic relations between vertices are injected into the query and key vectors:

$$\boldsymbol{q}_{v_i} = \boldsymbol{v}_{v_i} + \boldsymbol{l}_{v_i \to v_k}, \; \boldsymbol{k}_{v_k} = \boldsymbol{v}_{v_k} + \boldsymbol{l}_{v_k \to v_i}, \tag{4}$$

where $\boldsymbol{l}_{v_i \to v_k}$ and $\boldsymbol{l}_{v_k \to v_i}$ are learnable relation embeddings between vertices v_i and v_k. The self-attention is subsequently followed by a residual connection and a feed-forward layer, as done in the standard Transformer encoder (Vaswani et al. 2017). Finally, we obtain the commonsense cognition embedding $\boldsymbol{cs}_i$ for each $v_i^{CS} \in V_{CS}$.

To encode the emotional concept graph $\mathcal{G}_{EC}$, we adopt a vanilla Graph Transformer (i.e., omitting the relation enhancement part in the above Relation-Enhanced Graph Transformer). By superimposing the emotion intensity of each token, we obtain the emotional concept embedding $\boldsymbol{ec}_i$ for each $v_i^{EC} \in V_{EC}$.

Coarse-to-Fine Alignment

Context Encoding Following previous works (Majumder et al. 2020; Sabour, Zheng, and Huang 2022), we first concatenate all the utterances in the dialogue context X and prepend with a $[CLS]$ token: $[CLS] \oplus x_1 \oplus \ldots \oplus x_N$. This sequence is fed into a standard Transformer encoder (context encoder) to obtain the representation $\boldsymbol{S}_X$ of the dialogue context. We denote the representation of $[CLS]$ as $\boldsymbol{s}_X$.

Coarse-grained Alignment To reproduce the interaction of cognition and affection manifested in the dialogue context, we align contextual cognition and contextual emotional state at an overall level. They are separately acquired by cognitive and emotional knowledge discernment mechanisms, which select golden-like knowledge guided by the response.

To obtain the contextual cognitive representation $\boldsymbol{r}_{cog}$, the knowledge discernment calculates the prior cognitive distribution $P_{CS}\left(\boldsymbol{cs}_i \mid X\right)$ over the commonsense cognition knowledge (that is, only K_{CS} rather than all the vertices V_{CS} in $\mathcal{G}_{CS}$, and we thus use $1 \leq i \leq |K_{CS}|$ for simplicity):

$$\boldsymbol{r}_{cog} = \sum_{i=1}^{|K_{CS}|} P_{CS}\left(\boldsymbol{cs}_i \mid X\right) \cdot \boldsymbol{cs}_i, \tag{5}$$

$$P_{CS}\left(\boldsymbol{cs}_i \mid X\right) = \mathrm{softmax}_i \left(\boldsymbol{cs}_i^T \varphi_{CS}\left(\boldsymbol{s}_X\right) \right), \tag{6}$$

where $\varphi_{CS}(\cdot)$ is a MLP layer activated by $\tanh$. Similarly, we calculate the prior emotional distribution $P_{EC}\left(\boldsymbol{ec}_i \mid X\right)$ ($1 \leq i \leq |V_{EC}|$) and obtain the contextual emotional representation $\boldsymbol{r}_{emo}$.

During training, we use the ground truth response Y to guide the learning of knowledge discernment mechanisms. We feed Y into the cognition encoder (used for initializing the embeddings of $\mathcal{G}_{CS}$ above) and the context encoder to get the hidden states $\boldsymbol{S}_Y^{cog}$ and $\boldsymbol{S}_Y^{ctx}$, where the $[CLS]$ representations are $\boldsymbol{s}_Y^{ctx}$ and $\boldsymbol{s}_Y^{cog}$ respectively. The posterior cognitive distribution $P_{CS}\left(\boldsymbol{cs}_i \mid Y\right)$ and the emotional one $P_{EC}\left(\boldsymbol{ec}_i \mid Y\right)$ are calculated as follows:

$$P_{CS}\left(\boldsymbol{cs}_i \mid Y\right) = \mathrm{softmax}_i \left(\boldsymbol{cs}_i^T \boldsymbol{s}_Y^{cog} \right), \tag{7}$$

$$P_{EC}\left(\boldsymbol{ec}_i \mid Y\right) = \mathrm{softmax}_i \left(\boldsymbol{ec}_i^T \boldsymbol{s}_Y^{ctx} \right). \tag{8}$$

We then optimize the KL divergence between the prior and posterior distributions during training:

$$L_{KL} = L_{KL}^{CS} + L_{KL}^{EC}, \tag{9}$$

$$L_{KL}^{CS} = \sum_{i=1}^{|K_{CS}|} P_{CS}\left(\boldsymbol{cs}_i \mid Y\right) \cdot \log \frac{P_{CS}\left(\boldsymbol{cs}_i \mid Y\right)}{P_{CS}\left(\boldsymbol{cs}_i \mid X\right)}, \tag{10}$$

$$L_{KL}^{EC} = \sum_{i=1}^{|V_{EC}|} P_{CS}\left(\boldsymbol{ec}_i \mid Y\right) \cdot \log \frac{P_{EC}\left(\boldsymbol{ec}_i \mid Y\right)}{P_{EC}\left(\boldsymbol{ec}_i \mid X\right)}. \tag{11}$$

To further ensure the accuracy of discerned knowledge, similar to Bai et al. (2021), we employ the BOW loss to force the relevancy between cognitive / emotional knowledge and the target response. The BOW loss L_{BOW} is defined as:

$$L_{BOW} = -\frac{1}{|B|} \sum_{y_t \in B} \log \eta \left(y_t \mid \boldsymbol{r}'_{cog}, \boldsymbol{r}'_{emo} \right), \tag{12}$$

where $\eta(\cdot)$ is a MLP layer followed by $\mathrm{softmax}$ and the output dimension is the vocabulary size, B denotes the word bags of Y, $\boldsymbol{r}'_{cog} = \sum_{i=1}^{|K_{CS}|} P_{CS}\left(\boldsymbol{cs}_i \mid Y\right) \cdot \boldsymbol{cs}_i$, and $\boldsymbol{r}'_{emo} = \sum_{i=1}^{|V_{EC}|} P_{EC}\left(\boldsymbol{ec}_i \mid Y\right) \cdot \boldsymbol{ec}_i$.

Finally, we align the coarse-grained representations of the contextual cognition $\boldsymbol{r}_{cog}$ and the contextual emotional state $\boldsymbol{r}_{emo}$ using mutual information maximization (MIM). Specifically, we adopt the binary cross-entropy (BCE) loss L_{coarse} as the mutual information estimator that maximizes the mutual information between $\boldsymbol{r}_{cog}$ and $\boldsymbol{r}_{emo}$:

$$\begin{aligned} L_{coarse} &= 2 f_{coarse}\left(\boldsymbol{r}_{cog}, \boldsymbol{r}_{emo}\right) \\ &\quad - \log \sum_{\widetilde{\boldsymbol{r}}_{emo}} \exp\left(f_{coarse}\left(\boldsymbol{r}_{cog}, \widetilde{\boldsymbol{r}}_{emo}\right)\right) \\ &\quad - \log \sum_{\widetilde{\boldsymbol{r}}_{cog}} \exp\left(f_{coarse}\left(\widetilde{\boldsymbol{r}}_{cog}, \boldsymbol{r}_{emo}\right)\right), \end{aligned} \tag{13}$$

where $\widetilde{\boldsymbol{r}}_{emo}$ and $\widetilde{\boldsymbol{r}}_{cog}$ are the encoded negative samples. $f_{coarse}(\cdot, \cdot)$ is a scoring function implemented with a bilinear layer activated by the sigmoid function:

$$f_{coarse}\left(\boldsymbol{a}, \boldsymbol{b}\right) = \sigma\left(\boldsymbol{a}^T \boldsymbol{W}_{coarse} \boldsymbol{b}\right). \tag{14}$$

Fine-grained Alignment To simulate the interaction of fine-grained cognition and affection implied in the dialogue during human express empathy, the fine-grained alignment builds the fine-grained association between each inferred specific cognition and corresponding emotional reaction.

For each $u_i \in U$, we infer the commonsense knowledge about emotional reaction $K_i^{\text{xReact}} = \left[k_{i,1}^{\text{xReact}}, \ldots, k_{i,l}^{\text{xReact}}\right]$ using COMET, which is regarded as the user's possible emotional reaction to the current cognitive situation. Since $k_{i,j}^{\text{xReact}} \in K_i^{\text{xReact}}$ is usually an emotion word (e.g., happy, sad), we concatenate K_i^{xReact} and feed it into the Transformer-based encoder (reaction encoder) to get the representation of the emotional reaction $\boldsymbol{H}_i^{er}$. Similar to (Majumder et al. 2020) and (Sabour, Zheng, and Huang 2022), we use the average pooling to represent the reaction sequence, i.e., $\boldsymbol{h}_i^{er} = \mathrm{Average}\left(\boldsymbol{H}_i^{er}\right)$. To avoid over-alignment of out-of-context emotional reaction with cognition, we inject contextual information into the representation of reaction. We first connect $\boldsymbol{h}_i^{er}$ with the context representation $\boldsymbol{S}_X$ at the token level, i.e., $\boldsymbol{S}_i^{er}[j] = \boldsymbol{S}_X[j] \oplus \boldsymbol{h}_i^{er}$. Then another Transformer-based encoder takes $\boldsymbol{S}_i^{er}$ as input and output the fused information $\boldsymbol{S}_i^{er\prime}$. We take the hidden representation of $[CLS]$ in $\boldsymbol{S}_i^{er\prime}$ as the emotional reaction representation $\boldsymbol{er}_i$ of u_i.

Finally, we build the association between the inferred specific cognition $\left\{\bigcup_{j=1}^{l} \boldsymbol{cs}_{i,j}^{r}\right\}$ from u_i for $r \in \mathcal{R} = \{\text{xIntent, xNeed, xWant, xEffect}\}$ and the emotional reaction $\boldsymbol{er}_i$ using MIM. Recall that $\left\{\bigcup_{i=0}^{t} \bigcup_{r\in\mathcal{R}} \bigcup_{j=1}^{l} \boldsymbol{cs}_{i,j}^{r}\right\}$ exactly correspond to the commonsense cognition knowlege set K_{CS}. The fine-grained BCE Loss L_{fine} is defined as:

$$\begin{aligned} L_{fine} = & \sum_{i=0}^{t} \sum_{r\in\mathcal{R}} \sum_{j=1}^{l} \Big[2 f_{fine}\left(\boldsymbol{cs}_{i,j}^{r}, \boldsymbol{er}_i\right) \\ & - \log \sum_{\widetilde{\boldsymbol{er}}_i} \exp\left(f_{fine}\left(\boldsymbol{cs}_{i,j}^{r}, \widetilde{\boldsymbol{er}}_i\right)\right) \\ & - \log \sum_{\widetilde{\boldsymbol{cs}}_{i,j}^{r}} \exp\left(f_{fine}\left(\widetilde{\boldsymbol{cs}}_{i,j}^{r}, \boldsymbol{er}_i\right)\right) \Big], \end{aligned} \tag{15}$$

where $\widetilde{\boldsymbol{er}}_i$ and $\widetilde{\boldsymbol{cs}}_{i,j}^{r}$ are the encoded negative samples. $f_{fine}(\cdot,\cdot)$ is implemented as:

$$f_{fine}(\boldsymbol{a}, \boldsymbol{b}) = \sigma\left(\boldsymbol{a}^T \boldsymbol{W}_{fine} \boldsymbol{b}\right). \tag{16}$$

Altogether, the coarse-to-fine alignment module can be jointly optimized by L_{align} loss:

$$L_{align} = L_{BOW} + L_{KL} + L_{coarse} + \alpha L_{fine}, \tag{17}$$

where α is a hyper-parameter.

Emotion Prediction We fuse the contextual emotional state and emotional reaction to distill the affective representation, where we use $\boldsymbol{er}_0$ as the distillation signal of emotional reaction. This is because $\boldsymbol{er}_0$ is derived from the speaker's last utterance and represents the overall emotional reaction. A gating mechanism is designed to capture affective representation $\boldsymbol{r}_{aff}$:

$$\boldsymbol{r}_{aff} = \mu \cdot \boldsymbol{r}_{emo} + (1-\mu) \cdot \boldsymbol{er}_0, \tag{18}$$

$$\mu = \sigma\left(\boldsymbol{w}_{aff}^T \left[\boldsymbol{r}_{emo}; \boldsymbol{er}_0\right]\right). \tag{19}$$

We then project $\boldsymbol{r}_{aff}$ to predict the user's emotion e:

$$P_{emo}(e) = \operatorname{softmax}\left(\boldsymbol{W}_{emo} \boldsymbol{r}_{aff}\right), \tag{20}$$

which is supervised by the ground truth emotion label e^* using the cross-entropy loss:

$$L_{emo} = -\log P_{emo}(e^*). \tag{21}$$

Empathy-aware Response Generation

We employ a Transformer-based decoder to generate the response. To improve empathy perception in response generation, we devise two strategies to fuse the two parts of empathy (i.e., cognition and affection). First, we concatenate the cognitive and affective signals $\boldsymbol{r}_{cog}$ and $\boldsymbol{r}_{aff}$ with the dialogue context representation $\boldsymbol{S}_X$ at the token level, which is then processed by a MLP layer activated by $ReLU$ to integrate cognition and affection into the dialogue context:

$$\boldsymbol{S}'_X[i] = MLP\left(\boldsymbol{S}_X[i] \oplus \boldsymbol{r}_{cog} \oplus \boldsymbol{r}_{aff}\right). \tag{22}$$

Second, we modify the original Transformer decoder layer by adding two new cross-attention to integrate commonsense cognition knowledge $\boldsymbol{K}_{CS} = \{\boldsymbol{cs}_i\}_{i=1}^{|K_{CS}|}$ and emotional concept knowledge $\boldsymbol{K}_{EC} = \{\boldsymbol{ec}_i\}_{i=1}^{|V_{EC}|}$, which are inserted between the self-attention and cross-attention for $\boldsymbol{S}'_X$. The decoder then predicts the next token y_m given the previously decoded tokens $y_{<m}$, as done in the standard Transformer decoder.

We adopt the standard negative log-likelihood loss L_{gen} to optimize the response generation module:

$$L_{gen} = -\sum_{m=1}^{M} \log P\left(y_m \mid X, \mathcal{G}_{CS}, \mathcal{G}_{EC}, y_{<m}\right). \tag{23}$$

Finally, we jointly optimize the coarse-to-fine alignment loss, emotion prediction loss, generation loss, and diversity loss proposed by (Sabour, Zheng, and Huang 2022) as:

$$L = \gamma_1 L_{align} + \gamma_2 L_{emo} + \gamma_3 L_{gen} + \gamma_4 L_{div}, \tag{24}$$

where γ_1, γ_2, γ_3 and γ_4 are hyper-parameters.

Experiments

Experimental Setup

Dataset We conduct our experiments on the widely used EMPATHETICDIALOGUES (Rashkin et al. 2019), which is a crowdsourced multi-turn empathetic conversation dataset, comprising 25k open domain conversations between speakers and listeners. This dataset considers 32 emotional situations, and each conversation is highly related to a single emotional situation. In a conversation, the speaker confides personal experiences and feelings, and the listener infers the current situation and emotion of the speaker and responds empathetically. Following Rashkin et al. (2019), we split the train/valid/test set by 8:1:1.

Baselines We compare our *CASE* with several competitive baselines as follows: (1) *Transformer* (Vaswani et al. 2017): A vanilla Transformer-based response generation model. (2) *Multi-TRS* (Rashkin et al. 2019): A Multi-Task Transformer model that jointly optimizes a response generation task and an emotion prediction task. (3) *MoEL* (Lin et al. 2019): The Mixture of Empathetic Listeners softly combines the output representation of different listeners. (4) *MIME* (Majumder et al. 2020): A MIMic Emotion-based model considers the polarity-based emotion clusters and emotional mimicry. (5) *EmpDG* (Li et al. 2020): An Empathetic adversarial Dialogue Generator that utilizes multi-resolution user emotions and interactive user feedback. (6) *KEMP* (Li et al. 2022): A Knowledge-aware EMPathetic dialogue generation model that only uses concept knowledge. (7) *CEM* (Sabour, Zheng, and Huang 2022): A Commonsense-aware Empathetic chatting Machine that only exploits commonsense knowledge.

Models	PPL	Dist-1	Dist-2	Acc
Transformer	37.65	0.47	2.05	-
Multi-TRS	37.45	0.51	2.12	0.347
MoEL	38.35	0.44	2.10	0.322
MIME	37.33	0.41	1.62	0.296
EmpDG	37.77	0.53	2.26	0.314
KEMP	37.32	0.55	2.31	0.341
CEM	36.86	0.64	2.84	0.373
CASE	**35.37**	**0.74**	**4.01**	**0.402**

Table 1: Results of automatic evaluation.

Comparisons	Aspects	Win	Lose	κ
CASE vs. EmpDG	Coh.	**48.1**‡	39.2	0.54
	Emp.	**51.9**‡	32.9	0.55
	Inf.	**58.9**‡	31.6	0.50
CASE vs. KEMP	Coh.	**44.4**†	41.8	0.45
	Emp.	**50.0**‡	34.4	0.53
	Inf.	**51.1**‡	34.0	0.53
CASE vs. CEM	Coh.	**45.9**‡	42.2	0.51
	Emp.	**53.2**‡	34.6	0.47
	Inf.	**57.8**‡	29.8	0.56

Table 2: Human evaluation results (%). The agreement ratio kappa $\kappa \in [0.41, 0.6]$ denotes the moderate agreement. †,‡ represent significant improvement with p-value $< 0.1/0.05$, respectively. Ties are not shown.

Models	PPL	Dist-1	Dist-2	Acc
CASE	**35.37**	**0.74**	**4.01**	**0.402**
w/o Graph	36.10	0.68	3.50	0.280
w/o Align	35.75	0.65	3.34	0.369
w/o CSGraph	35.51	0.64	3.18	0.375
w/o ECGraph	36.24	0.72	3.94	0.329
w/o CGAlign	35.67	0.68	3.60	0.378
w/o FGAlign	35.55	0.67	3.43	0.370

Table 3: Results of overall-to-part ablation study.

Implementation Details We implemented all models with Pytorch. We initialize the word embeddings with pretrained GloVE word vectors (Pennington, Socher, and Manning 2014). The dimensionality of embeddings is set to 300 for all corresponding modules. We set hyper-parameters $l = 5$, $N' = 10$, $\alpha = 0.2$, $\gamma_1 = \gamma_2 = \gamma_3 = 1$ and $\gamma_4 = 1.5$. We use Adam optimizer (Kingma and Ba 2015) with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The batch size is 16 and early stopping is adopted. The initial learning rate is set to 0.0001 and we varied it during training following Vaswani et al. (2017). The maximum decoding step is set to 30 during inference. All models are trained on a GPU-P100 machine. The training process of CASE is split into two phases. We first minimize L_{BOW} for pretraining knowledge discernment mechanisms, and then minimize L for training overall model.

Automatic Evaluation

To evaluate the generative performance of the model, we adopt the widely used Perplexity (**PPL**) and Distinct-1/2 (**Dist-1/2**) (Li et al. 2016). Perplexity evaluates the general generation quality of a model. Distinct-1/2 evaluates the generated diversity by measuring the ratio of unique unigrams/bigrams in the response. To evaluate the emotion classification performance of the model, we measure the accuracy (**Acc**) of emotion prediction. Following KEMP and CEM, we do not report word overlap-based automatic metrics (Liu et al. 2016), e.g., BLEU (Papineni et al. 2002).

The automatic evaluation results are shown in Table 1. Our model outperforms all compared models and achieves a significant improvement on all metrics. **First**, our model achieves about 4.0% reduction on PPL compared to the best baseline, which indicates that CASE is more likely to generate ground truth responses. **Second**, our model achieves about 15.6% and 41.2% improvement on Dist-1 and Dist-2 compared to CEM, which indicates the superiority of CASE in terms of generating informative responses at the unigrams and bigrams level. This is attributed to the coarse-to-fine alignment that allows CASE to inject more informative commonsense cognition on the premise of ensuring the perplexity of the generated response. **Third**, our model achieves about 17.9% and 7.8% improvement in prediction accuracy compared to KEMP and CEM, respectively. This verifies that CASE considers both aspects of affection (i.e., contextual emotional state and emotional reaction) more effectively than focusing only on a single aspect as KEMP and CEM.

Human Evaluation

In human evaluation, 200 contexts are randomly sampled and each context is associated with two responses generated from our model CASE and baseline. Following Sabour, Zheng, and Huang (2022), three crowdsourcing workers are asked to choose the better one (**Win**) from two responses by considering three aspects, respectively, i.e., (1) Coherence (**Coh.**): which model's response is more fluent and context-related? (2) Empathy (**Emp.**): which model's response expresses a better understanding of the user's situation and feelings? (3) Informativeness (**Inf.**): which model's response incorporates more information related to the context? We use the Fleiss' **kappa** (κ) (Fleiss 1971) to measure the inter-annotator agreement. As in Table 2, the results show that CASE outperforms three more competitive baselines on all three aspects. Especially, CASE outperforms baselines significantly in terms of empathy and informativeness, which demonstrates the superior of modeling the interaction between cognition and affection of empathy, and supports the observations from automatic evaluation.

Overall-to-Part Ablation Study

We conduct an overall-to-part ablation study in Table 3 by removing key components of CASE for further dissection. In the overall ablation, **first**, we remove the commonsense cognition graph and emotional concept graph, called "w/o Graph". The emotion prediction accuracy decreases significantly, which indicates that the two heterogeneous graphs make remarkable contribution to detecting emotion. **Second**, we remove the coarse-to-fine alignment, called "w/o

Emotion	Terrified
Context	*Speaker*: As a kid, I was so **frightened** any time there was a tornado warning. Not sure if it was the wizard of oz effect or what i just remember being frozen with **fear**. *Listener*: Oh, wow! My son was **terrified** of tornadoes because of that darn movie, ... *Speaker*: My mom said the **best** thing once when I was babysitting and a tornado warning went off and I did not want my own **fear** to affect the little boy I was with. She was on the phone and just said you have big plans for the future, right? Nothing is going to get in the way of that. For some reason, It worked a **charm**.
MoEL	I am sorry to hear that. I hope you can get better soon!
MIME	I am sure it was a great feeling.
EmpDG	I am glad you are ok! I am sure you will have a great time.
KEMP	I would be too! I would have been terrified!
CEM	I am sure you will be fine!
Cognition	To make a plan, **To be safe**, To be helpful, **Have a good life**
Reaction	Scared, Hopeful, Nervous, **Good**, **Happy**
Concept	frighten, surprising, terrify, aghast, **good**, charming, **protect**
CASE	That is **good**, I am **glad** you are able **to stay safe** and **get a little better**!
Ground-Truth	Your mom sounds awesome! Wow, what a cool thing to say!
Emotion	**Embarrassed**
Context	*Speaker*: I burned my hair with my hair drier I am so embarrased to go out in public. *Listener*: I am **sorry** to hear that, my wife has **done** the same thing. She wore a hat for almost a month when she left the house. *Speaker*: Yeah it is the **worst**, I look so **weird** with my hair like this it is so embarassing.
MoEL	That is a pretty funny story.
MIME	I am sure it was a good idea.
EmpDG	Oh, that is a shame. I hope you are not too!
KEMP	I am sure you will not have to be able to do it again.
CEM	I am sorry to hear that.
Cognition	**To look bad**, To get a haircut, **To fix it**, **To fix the hair**, To get a wig
Reaction	Embarrassed, **Bad**, Ashamed, Sad
Concept	**bad**, ill, evil, freak, supernatural, exhausted, **hope**, sorrow
CASE	I **hope** you are able to **get it fixed**, or just let her know **it is not too bad**.
Ground-Truth	Well, like I told her, it will grow out, it will just take time.

Table 4: Case Study. Generated responses from baselines and CASE. **Commonsense cognition**, **emotional reactions**, reasoned **emotional concepts** by **contextual words**, and corresponding **informative expressions** in responses are highlighted.

Align". The diversity of generation decreases significantly and emotion prediction accuracy drops distinctly. It supports our motivation that the alignment of cognition and affection leads to informative and highly empathetic expression.

In the part ablation, **first**, we remove two heterogeneous graphs, called "w/o CSGraph" and "w/o ECGraph", respectively. From the results, we find that the contribution of the commonsense cognition graph is mainly to improve the diversity of generation (i.e., Dist-1/2), while the role of the emotional concept graph is mainly located in the recognition of emotion (i.e., Acc). This also supports our constructed motivation. **Second**, we remove coarse-grained and fine-grained alignments, called "w/o CGAlign" and "w/o FGAlign", respectively. We observe that the alignment at the fine-grained level is more significant than the coarse-grained level in terms of overall contribution. This also matches our intuition that building the fine-grained association between cognition and affection is closer to the conscious interaction process during human express empathy.

Case Study

Two cases generated from six models are selected in Table 4. Compared to the baselines, CASE has two main advantages:

(1) CASE is more likely to accurately identify the conversational emotion, being consistent to "Acc". This is due to the complementary effect of considering both the emotional concept and the emotional reaction. For instance, in the first case, only "*scared*" in the emotional reaction is insufficient to identify the current conversational emotion, i.e., "*Terrified*", while "*frighten, terrify, etc.*" in the emotional concept can complement the former. The opposite example is shown in the second case, i.e., for the conversational emotion "*Embarrassed*", "*embarrassed, ashamed, etc.*" in the emotional reaction vs. only "*bad*" in the emotional concept.

(2) CASE is more likely to express informative cognition in a highly empathetic tone. For example, in the first case, CASE uses the established association between cognition, e.g., "*to be safe*", and affection, e.g., "*good*", to appease the user's "*Terrified*" experience, i.e., "*to stay safe*" and "*get a little better*" in response. In the second case, in the identified user's "*Embarrassed*" emotional state, CASE expresses empathetic affection with "*it is not too bad*" and makes informative cognitive statement, i.e., "*get it fixed*", in response.

Conclusion and Future Work

In this paper, we propose CASE to align cognition and affection for responding empathetically by simulating the conscious alignment and interaction of the two in human conversation, which is inspired by CBT (Westbrook, Kennerley, and Kirk 2011). Extensive experiments verify the superiority of CASE on automatic and human evaluation. In the future, our work will inspire subsequent works to borrow psychological theories for empathetic dialogue and other promising tasks (e.g., emotional support conversation (Liu et al. 2021)).

References

Bai, J.; Yang, Z.; Liang, X.; Wang, W.; and Li, Z. 2021. Learning to Copy Coherent Knowledge for Response Generation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 12535–12543. AAAI Press.

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 4762–4779. 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.

Devlin, J.; Chang, M.; 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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

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.

Fleiss, J. L. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5): 378.

Hjelm, R. D.; Fedorov, A.; Lavoie-Marchildon, S.; Grewal, K.; Bachman, P.; Trischler, A.; and Bengio, Y. 2019. Learning deep representations by mutual information estimation and maximization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Huang, M.; Zhu, X.; and Gao, J. 2020. Challenges in Building Intelligent Open-domain Dialog Systems. *ACM Trans. Inf. Syst.*, 38(3): 21:1–21:32.

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

Kingma, D. P.; and Ba, J. 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*.

Kong, L.; de Masson d'Autume, C.; Yu, L.; Ling, W.; Dai, Z.; and Yogatama, D. 2020. A Mutual Information Maximization Perspective of Language Representation Learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL HLT 2016, The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, San Diego California, USA, June 12-17, 2016*, 110–119. The Association for Computational Linguistics.

Li, J.; Zhao, W. X.; Wei, Z.; Yuan, N. J.; and Wen, J. 2021. Knowledge-based Review Generation by Coherence Enhanced Text Planning. In *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*, 183–192. ACM.

Li, Q.; Chen, H.; Ren, Z.; Ren, P.; Tu, Z.; and Chen, Z. 2020. EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 4454–4466. International Committee on Computational Linguistics.

Li, Q.; Li, P.; Ren, Z.; Ren, P.; and Chen, Z. 2022. Knowledge bridging for empathetic dialogue generation. In *AAAI*.

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 2019, Hong Kong, China, November 3-7, 2019*, 121–132. Association for Computational Linguistics.

Liu, C.; 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, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, 2122–2132. The 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, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, 3469–3483. Association for Computational Linguistics.

Majumder, N.; Hong, P.; Peng, S.; Lu, J.; Ghosal, D.; Gelbukh, A. F.; 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 2020, Online, November 16-20, 2020*, 8968–8979. Association for Computational Linguistics.

Mohammad, S. M. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20, 000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, 174–184. Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, 311–318. ACL.

Peng, W.; Hu, Y.; Xing, L.; Xie, Y.; Zhang, X.; and Sun, Y. 2022. Modeling Intention, Emotion and External World in Dialogue Systems. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, 7042–7046. IEEE.

Pennington, J.; Socher, R.; and Manning, C. D. 2014. Glove: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, 1532–1543. ACL.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y. 2019. Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, 5370–5381. Association for Computational Linguistics.

Sabour, S.; Zheng, C.; and Huang, M. 2022. CEM: Commonsense-Aware Empathetic Response Generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, 11229–11237. AAAI Press.

Sap, M.; Bras, R. L.; 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. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, 3027–3035. AAAI Press.

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, February 4-9, 2017, San Francisco, California, USA*, 4444–4451. AAAI Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 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*, 5998–6008.

Wang, L.; Wang, D.; Tian, F.; Peng, Z.; Fan, X.; Zhang, Z.; Yu, M.; Ma, X.; and Wang, H. 2021. Cass: Towards building a social-support chatbot for online health community. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1): 1–31.

Wei, W.; Liu, J.; Mao, X.; Guo, G.; Zhu, F.; Zhou, P.; and Hu, Y. 2019. Emotion-aware Chat Machine: Automatic Emotional Response Generation for Human-like Emotional Interaction. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*, 1401–1410. ACM.

Westbrook, D.; Kennerley, H.; and Kirk, J. 2011. *An introduction to cognitive behaviour therapy: Skills and applications*. Sage.

Zech, E.; and Rimé, B. 2005. Is talking about an emotional experience helpful? Effects on emotional recovery and perceived benefits. *Clinical Psychology & Psychotherapy: An International Journal of Theory & Practice*, 12(4): 270–287.

Zhang, H.; Liu, Z.; Xiong, C.; and Liu, Z. 2020. Grounded Conversation Generation as Guided Traverses in Commonsense Knowledge Graphs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 2031–2043. Association for Computational Linguistics.

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, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, 813–824. Association for Computational Linguistics.

Zhong, P.; Wang, D.; Li, P.; Zhang, C.; Wang, H.; and Miao, C. 2021. CARE: Commonsense-Aware Emotional Response Generation with Latent Concepts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 14577–14585. AAAI Press.

Zhou, H.; Huang, M.; Zhang, T.; Zhu, X.; and Liu, B. 2018. Emotional Chatting Machine: Emotional Conversation Generation with Internal and External Memory. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, 730–739. AAAI Press.

Zhou, K.; Wang, H.; Zhao, W. X.; Zhu, Y.; Wang, S.; Zhang, F.; Wang, Z.; and Wen, J. 2020. S3-Rec: Self-Supervised Learning for Sequential Recommendation with Mutual Information Maximization. In *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, 1893–1902. ACM.