IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 23 December 2023, accepted 29 January 2024, date of publication 13 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365533

APPLIED RESEARCH

A Response Generation Framework Based on Empathy Factors, Common Sense, and Persona

WEIJIE LI[1,2], **YONG YANG**[1], **PALIDAN TUERXUN**[1], **XIAOCHAO FAN**[1], **AND YUFENG DIAO**[3]

[1]School of Computer Science and Technology, Xinjiang Normal University, Ürümqi, Xinjiang 830054, China
[2]School of Software, Xinjiang University, Ürümqi, Xinjiang 830091, China
[3]School of Computer Science and Technology, Inner Mongolia Minzu University, Tongliao, Inner Mongolia 028000, China

Corresponding author: Palidan Tuerxun (pldtrs@xjnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 62066044 and Grant 62167008; in part by the Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2022D01A99 and Grant 2021D01B72; in part by the Natural Science Foundation of China under Grant 62006130; in part by the Inner Mongolia Science Foundation under Grant 2022MS06028; and in part by the National and Local Joint Engineering Research Center of Intelligent Information Processing Technology for Mongolian, and Inner Mongolia Directly College and University Scientific Basic in 2022.

ABSTRACT Building a human-like dialogue system is a challenging task that requires effective use of context, common sense and personal information. In a conversation, the responder usually analyzes the emotion, intention, and common sense involved in the speaker's sentence. Based on this analysis, the responder considers both the above-mentioned content and their personal information to formulate a response. Previous work in this area has only focused on one or some aspects, such as emotion, intention, common sense or persona, rather than considering all of them together. To address this issue, we propose a response generation framework called EFCP, which is based on empathy factors, common sense, and persona. This framework simulates a rich dialogue generation process that is rarely seen in previous work. In predicting the type of empathy factors a responder should adopt, we consider both the responder's personal information and the conversation history. Our experiments show that this method effectively improves the accuracy of prediction. EFCP outperforms the baseline on a variety of automatic metrics and manual metrics, showing its potential for building more effective and human-like dialogue systems.

INDEX TERMS Communication systems, emotional responses, frequency response, natural language processing.

I. INTRODUCTION

Large-scale pre-trained language models, as described in [1], have demonstrated outstanding performance across various natural language understanding tasks [2]. In open-domain dialogue systems, the Multi-GPT2 model, which is based on pre-trained language models, can accept multiple inputs to create more personalized and diverse dialogues [3]. Empathy is a crucial capability in open-domain dialog systems, as described in [4]. Conversational models that integrate empathy can improve user satisfaction and receive more positive feedback in various domains [5].

Many factors contribute to the expression of empathy. The communication mechanism plays a critical role in empathy expression. Based on the theoretical definition of empathy, [6] identified three communication mechanisms for text-based empathy expression: emotional reaction (ER), interpretation (IP), and exploration (EX). Moreover, empathy comprises two broad aspects related to *cognition* and *affection* [7], [8]. These two aspects are reflected in the dialog acts (DA) taken [9] and the emotion (EM) expressed in the conversation, respectively. Reference [10] conducted a study on the EmpatheticDialogues dataset [11] to investigate the empathic response intentions associated with listeners in reaction to different emotional situations. Through manual analysis, they identified nine kinds of empathic response

The associate editor coordinating the review of this manuscript and approving it for publication was Mohamed M. A. Moustafa.

intentions (e.g., *agreeing*, *questioning*). The affective aspect of empathy pertains to expressing emotion appropriately in response to the experiences and feelings shared by the interlocutor (e.g., *fear*, *anger*) [12].

Understanding user emotions and expressions can be enhanced by incorporating common sense into the dialogue system. This approach can lead to more empathetic and informative responses [13], [14]. For instance, when PersonA shares their experience of being reunited with their lost fur baby, we can deduce from common sense that PersonA is feeling happy. Additionally, since the fur baby was lost for four years, PersonA will likely be more careful with it. By combining these insights, the listener can craft a more appropriate and emotional response. Moreover, persona is a critical aspect that plays a significant role in generating empathetic responses. According to [15], users tend to adopt different styles when expressing empathy, and personal information can help to personalize conversation content. Therefore, it cannot be ignored. Persona has been shown to be highly correlated with personality [16], and it influences the expression of empathy and response generation. For instance, assuming PersonB is a friendly person, he might reply to PersonA with, "I am really happy for you." Conversely, assuming PersonC has difficulty expressing themselves and is harsh, he might reply to PersonA with, "Do not fall in the same place twice."

To generate emotional responses in dialogue, it is important to consider elements such as communication mechanism (CM), dialogue act (DA), dialogue emotion (EM), common sense, and personal information. To achieve this, we have designed a data processing flow that incorporates all these elements. Here's how the flow works: upon receiving a sentence, the responder uses common sense to comprehend it and then analyzes its intention and emotion. Based on the available information, the responder evaluates their personal information and selects a communication method, intention, and emotional tone to generate a response. This paper introduces a response generation framework called EFCP, which comprises **E**mpathy **F**actors,[1] **C**ommon sense, and **P**ersona. The EFCP concept will be discussed in detail in Section IV. The contributions of this paper are summarized below:

1) We propose a data processing flow that generates responses using empathy factors, common sense, and persona.
2) Common sense enriches the context and predicts empathy factors, whereas persona information improves response quality and predicts empathy factors.
3) Automatic and manual evaluations demonstrate that the proposed model outperforms the strong baseline and produces more sensible empathetic responses.

II. RELATED WORK

Several studies have been conducted to develop models that can enhance empathy factors in the responses generated by chatbots. Reference [17] introduced an Emotional Chatting Machine that uses both internal and external memory. The model takes context x and emotion e as inputs and generates responses that contain the specified emotion. Reference [18] combined modeling of emotion (EM), dialogue acts (DA), and dialogue topics with the dialogue history, achieving a lower perplexity than the baseline. Reference [19] further showed that there is a hierarchical relationship between CM, DA, and EM They found that modeling the hierarchy CM → DA → EM achieves better performance than predicting each factor separately.

TABLE 1. The data set size (size) and average token count (avg) in context of dialogue are reported for different splits.

	#train	#validation	#test
size	86,771	18,991	19,536
avg	41.76	41.70	39.75

Daily conversations often involve commonsense reasoning. It has been demonstrated that integrating commonsense knowledge into dialogue systems is feasible. In a study by [13], the COMET was employed to rewrite the context to obtain commonsense knowledge, thereby improving the quality and empathy of the responses. Reference [20] constructed a large-scale Chinese commonsense knowledge graph, from which they trained a model to suggest relevant commonsense knowledge based on the context, using a Teacher-Student approach [21].

It is important to consider the aspect of emotional expression known as persona, as it is highly correlated with personality [22]. Although the connection between persona and empathy expression is not fully understood, it has been suggested that different speakers may have different styles for expressing empathy, which is natural. To address the limitation of a single input source for transformer-based pre-trained language models, [3] introduced the bidirectional attention module and the attention fusion module into the block of GPT-2.

III. DATA PREPARATION

To train our model, we require a dataset that contains empathy factor labels, persona, and common sense. We choose the PEC [22] dataset that meets these requirements. The dataset is labeled with CM (ER, IP, EX), DA, and EM by [19], who also filtered out some unsentimental data if the response was not labeled with any of ER, IP or EX. To reason about the context on several dimensions, we used the COMET model [23] and took the resulting results as common sense. After processing, the dataset size is shown in Table 1 and a sample of the input data is provided in Table 2. To make it easier to understand the data processing flow, we provide a diagram in Fig. 1. We will now describe in detail the process of preparing the data.

A. PARTITIONING THE DATASET

The dataset consists of two domains: happy and offmychest. We have noticed that some of the responders appear in

[1] For ease of description, the empathy factors CM, DA, and EM are collectively referred to.

TABLE 2. **An example of data set after data processing.**

Context after a long year of being suicidal, cheated on, and sexually assaulted, i was finally able to enjoy myself again on a family trip to disneyland and make new friends. i didn't think it would, but it gets better. **Context's empathy factors** DA=7, EM=5 **Context's common sense** *xReact* sad; excited; happy; good; relieved *xEffect* have a good time; get a new job; have a better life; gets a new job; get a new friend *xNeed* to be in a relationship; to go on a trip; to go on a vacation; to go to a party; to go to the park *xWant* to have a good time; to make new friends; to go on a vacation; to go on another trip; to go on another vacation *xIntent* to be happy; to have fun; to feel better; to feel happy; to feel good
Response good for you. this video really helped me i wanted to share since i just looked at it again tonight. **Response's empathy factors** ER=1, EX=0, IP=1, DA=7, EM=0
Original replier's persona i do solemnly promise gratitude for any help you might be able to give to us. **Rewritten by COMET** i am grateful. i am thankful. i am helpful. i am generous. i am kind. i feel grateful. i feel thankful. i feel happy. i feel good. i feel relieved. i get a reward. i get a job. i get thanked. i am thanked. i am grateful. i need to ask for help. i need to ask fot it. i need to ask for something. i need to approach someone. i need to make a promise. i want to ask for help. i wang to thank you. i want to thank them. i want to be grateful; i want to be thanked.i want to show gratitude; i want to be helpful. i want to be grateful. i want to show appreciation. i want to be kind.

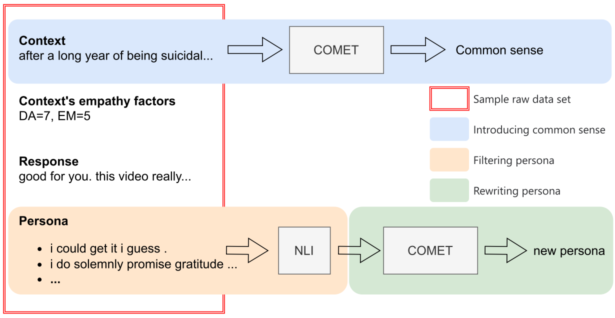

FIGURE 1. **Data processing flow diagram.**

different splits. For instance, a responder named *Evref* appears in Train, Validation, and Test sets of the happy domain. Since our work involves using personal information, the evaluation results of the model on the validation and test sets may be biased if these responders are not divided. To avoid this, we have ensured that the same person only appears in one split.

B. FILTERING PERSONA

It's important to note that just because there are multiple persona sentences in the dataset doesn't mean that all of them will be used when generating responses. To filter the original persona sentences, we use a well-trained RoBERTa model with an accuracy of 90.8% on DialogueNLI dev set as a Natural Language Inference (NLI) model. This model, released by [25], helps us preserve only 530,675 (7.1%) unique persona sentences related to the ground response. We discard data that has no persona associated with the reply. Additionally, in cases where there is more than one persona related to the reply, we only keep the persona with the highest relevance score.

C. PROCESSING CONTEXT

Some studies have indicated that using a transformer encoder-decoder model results in higher attention weights on the last transformer layer for the last utterance as compared to the average of the others. This has been observed in research works such as [25], [26], and [27]. Therefore, for each conversation, we only consider the last sentence of the dialogue history as the context.

D. INTRODUCING COMMON SENSE

It has been shown that introducing common sense to language models can help them better understand the context and generate more emotional responses [13]. To achieve this, we use a BART-based variation of COMET[2] that is trained on the ATOMIC-2020 dataset [28]. By inputting an event and the reasoning dimension into COMET, we can obtain common sense expansions for that 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 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 [29], we neglect *xAttr* and use the remaining five relations.

[2] https://github.com/allenai/comet-atomic-2020

E. REWRITING PERSONA

Having human-written interpretations of a persona sentence by rephrasing can often help provide novel information in persona grounding. Therefore, similar to the approach taken in [31], we use COMET to rewrite persona sentences. We choose to expand the relations *xReact*, *xIntent*, *xWant*, *xNeed*, *xEffect*, and *xAttr*, and generate five expansions per relation for each sentence. Since original persona sentences usually start with "*I*" (e.g. "*I do*","*I feel*"), the extensions generated by COMET are words or phrases (see Table 2). Therefore, for each expansion, we preprocess the generated commonsense inferences to add suitable prefixes to make them similar to the original persona. For example, expansions relating to "*xWant*" and "*xAttr*" are prefixed with "*I want*" and "*I am*", respectively. Then, we concatenate all the expanded sentences to replace the original persona sentences.

IV. METHODOLOGY

In this section, we will describe the response generation process proposed in the introduction. The overall EFCP architecture is depicted in Fig. 2.

A. ENRICHED BY COMMON SENSE

Firstly, the responder analyzes the contextual information related to the conversation. This common sense information is then used to enrich the context. We combine the original context x with the generated common sense sentences to create a new sentence $x' = [t_1, t_2, \ldots, t_N]$, where N is the length of x'.

B. ENHANCED BY EMPATHY FACTORS

In the second step of our response generation process, we infer the speaker's intent and emotion based on contextual clues. To achieve this, we utilized data from [19], where they fine-tuned RoBERTa model [32][3] as DA classifier and EM classifier respectively on EmpatheticIntents [10] and GoEmotions [12]. The context and response in PEC [22] were labeled, and the speaker's intent was derived from the categories embedded during the data pre-processing stage. To integrate empathy factors with context, we describe the process below.

In the first step, we create a special token [USR] to represent the speaker and add it to the vocabulary. Then, we create two embedding layers with sizes $\boldsymbol{M}_A \in \mathbb{R}^{9\times d}$ and $\boldsymbol{M}_E \in \mathbb{R}^{10\times d}$ to convert the numerical class labels for DA and EM into vectors. In the second step, we obtain the embedding $\boldsymbol{e}_{x'}$ of x' by summing up the word embedding, positional embedding, [USR] embedding, DA embedding, and EM embedding. This final result is used for integration, and the formula is expressed as follows:

$$\begin{aligned}\boldsymbol{e}_{t_i} &= \boldsymbol{M}_W[w_{t_i}] + \boldsymbol{M}_W\left[w_{[\mathrm{USR}]}\right] \\ &\quad + \boldsymbol{M}_P\left[p_{t_i}\right] + \boldsymbol{M}_A\left[A_x\right] + \boldsymbol{M}_E\left[E_x\right] \qquad (1)\\ \boldsymbol{e}_{x'} &= \left[\boldsymbol{e}_{t_1}; \boldsymbol{e}_{t_2}, \ldots, \boldsymbol{e}_{t_N}\right], \qquad (2)\end{aligned}$$

where $\boldsymbol{M}_W$ is an embedding matrix that represents words, while $\boldsymbol{M}_P$ is an embedding matrix that represents positions. Here, $\mathcal{V}$ is the vocabulary. The variable p_{t_i} represents the position number of the token t_i, and w_* represents the token id of $*$. We use A_x and E_x to represent the speaker's dialog intention category and emotion category, respectively. $[\cdot]$ indicates an indexing operation, and $[;]$ denotes vector concatenation. We denote the output hidden states after feeding x' into the encoder as $\boldsymbol{H}_x \in \mathbb{R}^{l_x\times d}$, where l_x is the length of x'.

$$\boldsymbol{H}_x = \mathbf{Enc}(\boldsymbol{e}_x) \qquad (3)$$

C. EMPATHY FACTORS PREDICTION

This section describes the third step of our design process. In this step, we rely on context, common sense, and personal information to make predictions about the communication mechanism, response intention, and emotion that the responder should adopt. The context and common sense are merged into a vector $\boldsymbol{H}_x$ to aid in these predictions. To start, we introduce a unique token called [PERSONA]. This token is added to the vocabulary and serves to represent the persona of the responder. Additionally, we create three embedding layers: $\boldsymbol{M}_C^{(\mathrm{ER})}$, $\boldsymbol{M}_C^{(\mathrm{EX})}$, $\boldsymbol{M}_C^{(\mathrm{IP})}$. These layers are of size $\mathbb{R}^{2\times d}$ and are used to convert the numerical class labels for ER, EX, and IP into vectors. We then obtain the embedding $\boldsymbol{e}_p$ of persona p by adding up the word embedding, positional embedding, and [PERSONA] embedding. Next, we feed this embedding into the encoder to obtain $\boldsymbol{H}_p \in \mathbb{R}^{l_p\times d}$, where l_p represents the length of the persona sentence.

$$\boldsymbol{H}_p = \mathbf{Enc}\left(\boldsymbol{e}_p\right) \qquad (4)$$

In the process, a non-linear network is utilized to convert $[\hat{\boldsymbol{h}}_x; \hat{\boldsymbol{h}}_p] \in \mathbb{R}^{2d}$ into $\boldsymbol{H}_t \in \mathbb{R}^d$. Here, $\hat{\boldsymbol{h}}_x$ and $\hat{\boldsymbol{h}}_p$ refer to the hidden state at the final position of $\boldsymbol{H}_x$ and $\boldsymbol{H}_p$, respectively. $\boldsymbol{H}_t$ is now a combination of context, common sense, and personal information. We use it to predict the empathy factors of the response y. It should be noted that in this paper, the ground truth value and predicted value of a variable X are represented by X^* and $\widehat{X}$, respectively.

We predict whether to adopt each CM separately: $\widehat{C}_y^{(i)} \in \{0, 1\}$ for $i \in \{\mathrm{ER,EX,IP}\}$.

$$\boldsymbol{h}_C^{(i)} = \mathbf{F}_C^{(i)}\left(\boldsymbol{H}_t\right) \in \mathbb{R}^d \qquad (5)$$

$$\widehat{C}_y^{(i)} \sim \mathbb{P}\left(C_y^{(i)} \mid x, CS, p\right) = \operatorname{softmax}\left(\boldsymbol{M}_C^{(i)}\boldsymbol{h}_C^{(i)}\right) \qquad (6)$$

$$\widehat{C}_y = \left(\widehat{C}_y^{(\mathrm{ER})}, \widehat{C}_y^{(\mathrm{IP})}, \widehat{C}_y^{(\mathrm{EX})}\right) \qquad (7)$$

$$\boldsymbol{e}_{\widehat{C}_y} = \sum_{i\in\{\mathrm{ER,IP,EX}\}} \boldsymbol{M}_C^{(i)}\left[\widehat{C}_y^{(i)}\right], \qquad (8)$$

where each $\mathbf{F}_C^{(i)}$ is a non-linear layer. Based on the context x and the predicted CMs $\widehat{C}_y$, we next predict DA:

$$\boldsymbol{h}_A = \mathbf{F}_A\left(\left[\boldsymbol{H}_t; \boldsymbol{e}_{\widehat{C}_y}\right]\right) \in \mathbb{R}^d \qquad (9)$$

$$\widehat{A}_y \sim \mathbb{P}\left(A_y \mid x, CS, p, \widehat{C}_y\right) = \operatorname{softmax}\left(\boldsymbol{M}_A\boldsymbol{h}_A\right) \qquad (10)$$

$$\boldsymbol{e}_{\widehat{A}_y} = \boldsymbol{M}_A\left[\widehat{A}_y\right], \qquad (11)$$

[3] https://huggingface.co/roberta-base

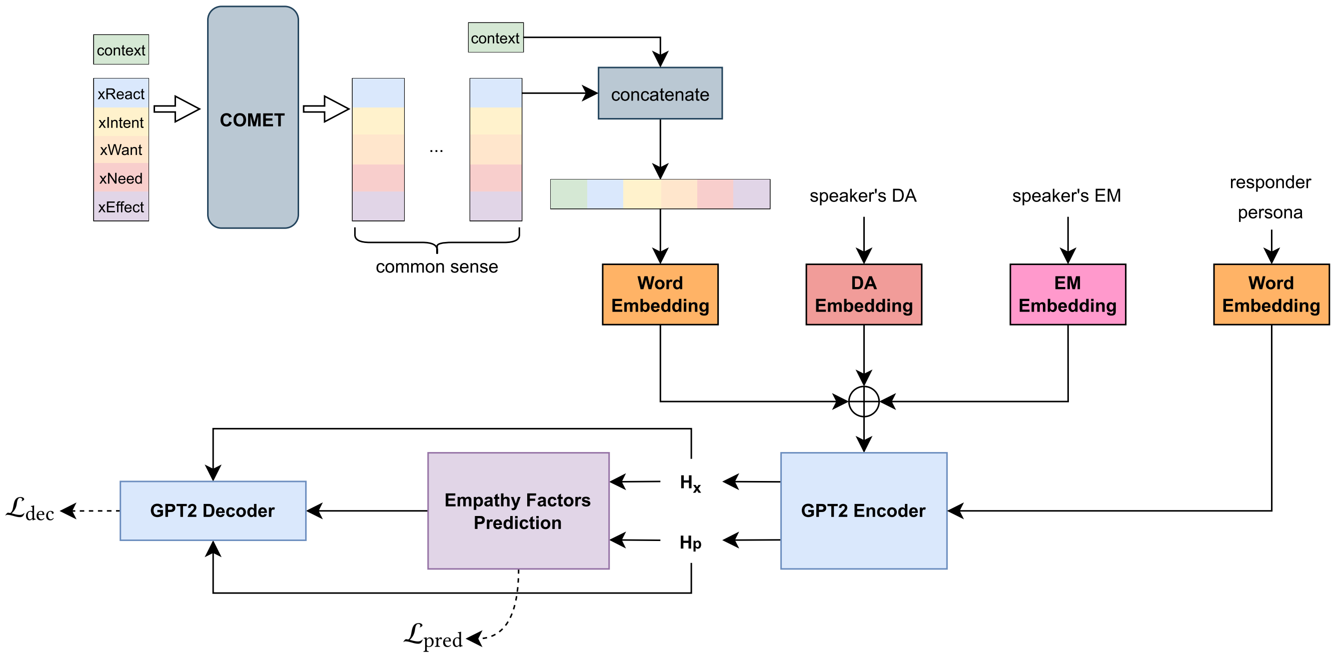

FIGURE 2. **Overview of our model. The dotted line is used only during training.**

where $[\cdot;\cdot]$ denotes vector concatenation and $\mathbf{F}_A$ is a non-linear layer. EM $\widehat{E}_y$ is predicted similarly but conditioned additionally on the predicted DA $\widehat{A}_y$:

$$\boldsymbol{h}_E = \mathbf{F}_E\left(\left[\boldsymbol{H}_t; \boldsymbol{e}_{\widehat{C}_y}; \boldsymbol{e}_{\widehat{A}_y}\right]\right) \in \mathbb{R}^d \tag{12}$$

$$\widehat{E}_y \sim \mathbb{P}\left(E_y \mid x, CS, p, \widehat{C}_y, \widehat{A}_y\right) = \operatorname{softmax}\left(\boldsymbol{M}_E \boldsymbol{h}_E\right), \tag{13}$$

where $\mathbf{F}_E$ is also a non-linear layer.

In the prediction process, we follow the same approach as [19] for predicting empathy factors. However, we incorporate a vector called $\boldsymbol{H}_t$, which combines context, common sense, and personal information, instead of using context alone. We conducted ablation experiments to demonstrate that incorporating common sense or persona information along with context improves the accuracy of predicting empathy factors when compared to using context alone.

D. GENERATE RESPONSE

This section represents the final step in our process. We utilize empathy, common sense, and personas to generate responses. To embed each input token $\widehat{y_t}$ in the response, the following process is followed:

$$\begin{aligned}\boldsymbol{e}_{\widehat{y_t}} = \boldsymbol{M}_W\left[w_{\widehat{y_t}}\right]\boldsymbol{M}_W\left[w_{[\mathrm{SYS}]}\right] \\ + \boldsymbol{M}_P\left[p_{\widehat{y_t}}\right] + \sum_{i\in\{\mathrm{ER,IP,EX}\}} \boldsymbol{M}_C^{(i)}\left[\widehat{C}_y^{(i)}\right] \\ + \boldsymbol{M}_A\left[\widehat{A}_y\right] + \boldsymbol{M}_E\left[\widehat{E}_y\right],\end{aligned} \tag{14}$$

where [SYS] is a special token and represents the responder. Then, we input $\boldsymbol{e}_{\widehat{y_t}}$, $\boldsymbol{H}_x$, and $\boldsymbol{H}_p$ into the decoder to obtain the output hidden state corresponding to $\widehat{y}_t$. This state is denoted as $\boldsymbol{s}_t$. Finally, we predict the next token $\widehat{y}_{t+1}$ through the LM head:

$$\begin{aligned}\boldsymbol{s}_t &= \mathbf{Dec}\left(\boldsymbol{e}_{\widehat{y_t}}, \boldsymbol{H}_x, \boldsymbol{H}_p\right) \\ \widehat{y}_{t+1} &\sim \mathbb{P}\left(y_{t+1} \mid \widehat{y}_{\le t}; x, CS, \widehat{C}_y, \widehat{A}_y, \widehat{E}_y, p\right) && (15) \\ &= \operatorname{softmax}\left(\boldsymbol{M}_W \boldsymbol{s}_t\right), && (16)\end{aligned}$$

where the parameters of the LM head are shared with the word embedding matrix $\boldsymbol{M}_W$.

E. TRAINING

Our overall loss has two components. The first component is the negative log-likelihood loss of decoding, denoted as $\mathcal{L}_{\text{dec}}$. The second component is the prediction losses of empathy factors, denoted as $\mathcal{L}_{\text{pred}}$. The purpose of $\mathcal{L}_{\text{pred}}$ is to optimize the embedding representation of empathy factors, which helps to improve the quality of response generation and the accuracy of predicting the category of empathy factors that should be used in the response. The formula for $\mathcal{L}_{\text{pred}}$ is given below:

$$\mathcal{L}_C = -\sum_{i\in\{\mathrm{ER,IP,EX}\}} \ln\mathbb{P}\left(C_y^{(i)*} \mid x, CS, p\right) \tag{17}$$

$$\mathcal{L}_A = -\ln\mathbb{P}\left(A_y^* \mid x, CS, p, C_y^*\right) \tag{18}$$

$$\mathcal{L}_E = -\ln\mathbb{P}\left(E_y^* \mid x, CS, p, C_y^*, A_y^*\right) \tag{19}$$

$$\mathcal{L}_{\text{pred}} = \mathcal{L}_C + \mathcal{L}_A + \mathcal{L}_E \tag{20}$$

The $\mathcal{L}_{\text{dec}}$, can be obtained by the following formula:

$$\mathcal{L}_{\text{dec}} = -\frac{1}{l_y}\sum_{t=1}^{l_y} \ln\mathbb{P}\left(y_t^* \mid y_{<t}^*; x, CS, p, C_y^*, A_y^*, E_y^*\right) \tag{21}$$

where l_y represents the length of the golden response, the optimization object is the sum of the following losses: $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{dec}}$, where $\lambda_1 = \lambda_2 = 1.0$.

V. EXPERIMENTS

A. COMPARED MODELS

We choose models that are closely related to our work as the baseline.

- **GPT2** [1] without empathy factors, personas and common sense, generates replies based on context.
- **MultiGPT2** [3] accepts multiple input sources. Among the various attention fusion methods, we choose source-level scalar weights (**GPT2-sw**) and linear method (**GPT2-linear**) which had the best performance in their experiment.
- **CoMAE** [19] used empathy factors to enhance the empathetic response generation.
- **CEM** [13] adopted common sense to enhance emotional response, and utilized EM to assist in adjusting the affection-refined encoder.

B. IMPLEMENTATION DETAILS

All the models are implemented with PyTorch and the Transformers library[4] [33]. We use the pre-trained DistilGPT2[5] with the size of 82M parameters (768 hidden sizes, 12 heads, 6 layers). The responses are decoded by Top-p sampling with $p = 0.9$ and the temperature $\tau = 0.7$ [34]. We train the model with AdamW [35] optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. The learning rate is 10^{-4} and is dynamically changed using the linear warmup [36] with 1,300 warmup steps. Our model uses MultiGPT2 as the infrastructure, we choose source-level scalar weights (sw) as the multi-input source fusion mechanism. We combine the two domains' train set as the Train. The Validation is obtained in the same way. All models are fine-turned for 5 epochs with the batch size 32 on one NVIDIA RTX 3090 GPU. The parameter number of EFCP is 200M. We test the model on the validation set every 1,300 steps and reserve the checkpoint with the lowest perplexity score. The training time of our models is about 2 hours for around 13550 iterations. The Code is available at https://github.com/SilverBeats/efcp

C. AUTOMATIC EVALUATION

The automatic evaluation uses the golden responses as a reference to evaluate the responses generated by models. However, when responses are generated based on the predicted empathy factors, it is not appropriate to compare the generated responses with the reference ones [37]. Same as [19], in automatic evaluation, we only considered the setting where the models are fed with the ground truth empathy factors. The automatic metrics we adopt can be divided into three categories: 1) based on word overlap: BLEU-n (**B-**n) [38], ROUGE-L (**R-L**) [39]; 2) based on word embedding: Greedy matching score (**GMS**) [40], Embedding average score (**EAS**) [41], Vector extrema cosine similarity score (**VES**) [42]; and 3) based on PLM: perplexity (**PPL**), F1 score of bertscore (**BERT**) [43].[6] In addition, we report the geometric mean (**GM**) of all automatic measures as an overall performance [44], [45], [46]. In calculating the geometric mean, we use $\frac{1}{\log\text{PPL}}$ rather than the PPL value directly. The PPL takes a different range of values than the other seven metrics (between 0 and 1). By doing logarithmic processing on PPL, the computational bias caused by the different value ranges of the metrics can be reduced. In addition, the lower the PPL result, the better, which is also different from the other seven metrics. Therefore, we take the inverse of PPL. The calculation formula of GM is shown in (22), where $n = 8$ represents the number of automatic evaluation metrics, and V_i represents the result of the i evaluation metric. $i \in \{\text{B-1}, \text{B-2}, \text{R-L}, \text{BERT}, \text{GMS}, \text{ESA}, \text{VES}\}$

$$GM = \sqrt[n]{\frac{\prod V_i}{\log\text{PPL}}} \tag{22}$$

D. RESULTS

Our experimental results, as shown in Table 3, demonstrate that the EFCP model outperforms all baseline models. The word overlap metrics (B-n, R-L), reflect the degree of overlap in word usage between generated responses and real responses. On the other hand, word embedding-based metrics (GMS, ESA, VES) and pre-trained language model-based metrics measure the semantic similarity between generated responses and real responses. The EFCP model shows promising results on all three types of automatic evaluation metrics, indicating that it can generate responses that are more relevant to the reality of the responses. To measure the fluency of the generated sentences, we use the PPL metric, where a smaller value of PPL means that the generated sentences are more fluent. Additionally, higher-order BLEU score also reflects sentence fluency to some extent. Our results show that the EFCP model outperforms all baselines in terms of generation on the PEC dataset. It is worth noting that the EFCP model combines multiple aspects related to daily conversations, such as common sense (**cs**), empathy factors (**ef**), and personal information (**per**), in generating responses, while the baselines use only one of them. Our experimental findings suggest that simulating the thought process of human-generated replies for daily conversations and generating replies with a combination of aspects affecting conversation generation can greatly improve the effectiveness of conversation generation.

To further support our argument, we conducted a performance test on different variations of the model in various settings. The results of this experiment can be seen in Table 4. To facilitate a better comparison of the overall performance of the model under different conditions, we extracted the GM metric columns and presented them separately in Table 6.

[4] https://github.com/huggingface/transformers

[5] https://huggingface.co/distilgpt2

[6] Use roberta-large to calculate bertscore.

TABLE 3. Result of automatic evaluation. Best scores are in bold. For comparison purposes, improvements are calculated compared to the best baseline results.

Models	R-L	B-1	B-2	PPL	BERT	GMS	ESA	VES	GM
GPT2	16.47	18.54	7.35	22.74	86.08	69.46	87.73	50.21	21.37
GPT2-sw	18.17	20.61	9.12	21.17	86.53	70.08	87.88	51.17	22.69
GPT2-linear	18.68	21.26	9.51	27.72	86.59	70.49	87.95	51.97	22.80
CoMAE	20.16	23.56	10.65	20.21	87.07	70.89	88.12	52.53	24.02
CEM	16.55	19.37	7.73	23.49	86.25	69.12	87.37	50.41	21.60
EFCP	**21.83** (+1.67)	**25.31** (+1.75)	**12.17** (+1.52)	**19.62** (+0.59)	**87.43** (+0.36)	**71.64** (+0.75)	**88.36** (+0.24)	**53.75** (+1.22)	**25.04** (+1.03)

TABLE 4. Result of ablation experiment. Best scores are in bold.

Models	R-L	B-1	B-2	PPL	BERT	GMS	ESA	VES	GM
EFCP	**21.83**	**25.31**	**12.17**	**19.62**	87.43	**71.64**	**88.36**	**53.75**	**25.04**
w/o cs	21.81 (-0.02)	25.14 (-0.17)	12.13 (-0.04)	19.67 (-0.05)	**87.44** (+0.01)	71.59 (-0.05)	88.32 (-0.04)	53.73 (-0.02)	25.00 (-0.04)
w/o per	20.45 (-1.38)	23.17 (-2.14)	10.62 (-1.55)	20.52 (-0.90)	87.19 (-0.24)	70.70 (-0.94)	87.82 (-0.54)	52.77 (-0.98)	23.99 (-1.06)
w/o ef	18.50 (-3.33)	21.74 (-3.57)	9.77 (-2.40)	21.13 (-1.51)	86.72 (-0.71)	69.99 (-1.65)	87.54 (-0.82)	51.65 (-2.10)	23.11 (-1.93)
w/o ef & per	16.50 (-5.33)	20.06 (-5.25)	8.01 (-4.16)	22.43 (-2.81)	86.34 (-1.09)	68.87 (-2.77)	86.97 (-1.39)	50.49 (-3.26)	21.81 (-3.23)
w/o ef & cs	18.33 (-3.50)	21.51 (-3.80)	9.55 (-2.62)	20.86 (-1.24)	86.67 (-0.76)	69.91 (-1.73)	87.49 (-0.87)	51.56 (-2.19)	22.99 (-2.05)
w/o cs & per	20.34 (-1.49)	23.10 (-2.21)	10.55 (-1.62)	20.54 (-0.92)	87.19 (-0.24)	70.66 (-0.98)	87.73 (-0.63)	52.76 (-0.99)	23.93 (-1.11)
w/o all	16.65 (-5.18)	20.28 (-5.03)	8.25 (-3.92)	22.55 (-2.93)	86.39 (-1.04)	68.94 (-2.70)	87.05 (-1.31)	50.74 (-3.01)	21.96 (-3.08)

TABLE 5. Empathy factors prediction accuracy results (%). Best scores are in bold. EFCP uses CoMAE as the basis to calculate the lifting range. w/o * use EFCP as the benchmark to calculate the index change amplitude.

Models	ER	EX	IP	DA		EM	
	Hit@1	Hit@1	Hit@1	Hit@1	Hit@3	Hit@1	Hit@3
CoMAE	78.41	87.86	78.19	49.94	82.22	58.45	86.62
EFCP	**81.19** (+2.78)	**88.95** (+1.09)	**80.66** (+2.47)	53.13 (+3.19)	**84.73** (+2.51)	60.74 (+2.29)	**88.84** (+2.22)
w/o cs	81.18 (-0.01)	88.85 (-0.10)	80.47 (-0.19)	**53.40** (+0.27)	84.57 (-0.16)	**60.85** (+0.11)	88.54 (-0.30)
w/o pwp	78.40 (-2.79)	88.24 (-0.71)	78.41 (-2.25)	49.91 (-3.22)	82.38 (-2.35)	58.75 (-1.99)	86.98 (-1.86)

TABLE 6. Composite score for each model in the ablation experiment. ✓indicates that the model used the information when generating the reply.

Models	GM	cs	ef	per
EFCP	25.04	✓	✓	✓
w/o cs	25.00		✓	✓
w/o per	23.99	✓	✓	
w/o cs & per	23.93		✓	
w/o ef	23.11	✓		✓
w/o ef & cs	22.99			✓
w/o all	21.96			
w/o ef & per	21.81	✓		

The table shows the differences between the model variations. Based on the results of the ablation experiment, we can draw the following conclusions:

1) Common sense and empathy are two important factors that work better together. Humans often unconsciously use common sense in daily conversations, while omitting common sense knowledge while speaking. Therefore, we use COMET to generate common sense content that refines the complementary context. However, we have observed that COMET tends to produce duplicate inference results which was also mentioned in [47]. If incorrect inferences are generated and repeated multiple times, the encoded results could be out of the original context's meaning. Our observation of two groups of comparative experiments, namely *w/o all* and *w/o ef & per (only use cs)*, *w/o ef & cs (only use per)* and *w/o ef (use cs and per)* found that the performance of *w/o all* and *w/o ef & cs* decreased after combining common sense, thus supporting this statement. Despite the problems with COMET, we cannot deny that the addition of common sense can enrich the context's content. As long as we can do some special processing when encoding the context that combines common sense, common sense can play a positive role. By observing two groups of comparison experiments of *w/o cs & per (only use ef)* and *w/o per (use cs and ef)*, *w/o cs (use ef and per)* and *EFCP*, this special treatment is to add empathy factors information when encoding.

2) Having common sense can be an added benefit. After conducting comparative experiments between *w/o cs* and *EFCP*, as well as *w/o cs & per* and *w/o per*, it was observed that introducing common sense led to an improvement in the GM score. However, the rate of improvement was not very significant.

TABLE 7. Results of manual evaluation. Ties are not shown. The metrics significant gaps area marked with * (sign test, *p*-value < 0.01). κ denotes Fleiss' Kappa, whose values indicate fair agreement ($0.2 < \kappa < 0.4$) or moderate agreement ($0.4 < \kappa < 0.6$).

Comparisons	Metrics	Win	Lose	κ
EFCP	Flu	31	**53**	0.420
vs.	Coh*	**47**	21	0.338
GPT2-linear	Emp*	**48**	37	0.238
EFCP	Flu	39	**46**	0.466
vs.	Coh*	**46**	26	0.262
CoMAE	Emp*	**51**	39	0.324
EFCP	Flu	35	**56**	0.394
vs.	Coh*	**49**	22	0.218
CEM	Emp*	**47**	35	0.238

3) The benefits of using empathy factors and persona are significant in improving GM. Using either persona or empathy factors alone can result in significant improvement, and combining them can further enhance the model's performance. This is because EFCP includes a component that predicts the categories of empathy factors that should be adopted by respondents. By integrating the personal information of the respondent, the prediction accuracy of empathy factors can be improved during training. We present experimental results in Table 5 to support this claim. When empathy factors are predicted without considering persona (*w/o pwp*), the accuracy drops significantly. The process of predicting empathy factors in EFCP is similar to [19], but EFCP combines common sense and personal information to improve prediction accuracy. Table 5 demonstrates that introducing common sense or persona can significantly enhance the prediction accuracy of empathy factors.

E. MANUAL EVALUATION

For manual evaluation, the responses generated by different models are compared in pairs based on various metrics. These metrics include **Fluency** (which response has better fluency and readability), **Coherence** (which response has better coherence and higher relevance to the context), and **Empathy** (which response shows better understanding of the partner's experiences and feelings, and which response expresses empathy in the way that the annotators prefer) [13], [48]. We randomly selected 100 samples from the test set. and each model generated responses for these samples. We compared the responses generated by the EFCP model with those generated by GPT2-linear, CEM, and CoMAE models. Finally, we ended up with 300 binary tuples to be evaluated.

$$\begin{aligned}&\{(\mathrm{EFCP}_1, \mathrm{CEM}_1)\ldots(\mathrm{EFCP}_{100}, \mathrm{CEM}_{100}),\\&\quad(\mathrm{EFCP}_1, \mathrm{CoMAE}_1)\ldots(\mathrm{EFCP}_{100}, \mathrm{CoMAE}_{100}),\\&\quad(\mathrm{EFCP}_1, \text{GPT2-linear}_1)\ldots(\mathrm{EFCP}_{100}, \text{GPT2-linear}_{100})\}.\end{aligned}$$

We recruited three master's degree students majoring in English as volunteers to manually evaluate 300 binary tuples. We stayed in touch with them throughout the evaluation process to resolve any issues that arose as soon as possible. Each binary tuple was evaluated by three people, resulting in three evaluations. We then voted to determine which model performed better on each binary tuple. For instance, if the fluency of $(\mathrm{EFCP}_1, \mathrm{CEM}_1)$ was rated 1, 2, 1 by the three evaluators, with a score of 1 indicating better EFCP performance and a score of 2 indicating better CEM performance, the winner was EFCP for fluency in that binary tuple. In case the winner could not be determined by vote, it was considered a tie. The results of the manual assessment are presented in Table 7.

We reviewed 300 sets of samples and found that the EFCP model was weaker in fluency compared to the baseline model. According to the reviewers' feedback, they preferred the model that generated longer sentences, provided that the responses were free of grammatical errors and internal logical contradictions. The average sentence length generated by EFCP, GPT2-linear, CoMAE, and CEM were 15.03, 17.39, 16.49, and 17.82, respectively. EFCP generated the shortest responses, which made it less fluent than the baseline models. However, reviewers reported that all three baseline models had problems generating generic responses. The question of how to allow diversity in the generated responses is a problem that needs to be addressed in the field of dialog systems. We expect the model to express emotions based on understanding the speaker and to generate smooth and coherent responses that are relevant to the context. Compared to the baseline model, EFCP is far ahead in terms of relevance and empathy.

F. CASE STUDY

The comparison between the generated responses of our models and the baselines is shown in Table 9. It is clear from the examples that the content produced by the EFCP model is more specific compared to other models. While other models generate content that aligns with the current context, they lack unique vocabulary that closely relates to that context, such as the word "sobriety". Even if the current topic is changed to "celebration", the responses generated by models other than EFCP will still be applicable. On the other hand, the response generated by EFCP conforms to the setting of the empathy factor. It shows "admiration" in terms of emotion and "wish" in terms of dialog action, and generates content that elicits a strong emotional reaction. The EFCP model generates content that is more empathetic and emotionally connected to the user.

G. APPLICATION TO LARGE LANGUAGE MODELS

We conducted two experiments to test the effectiveness of the EFCP approach on advanced language models such as GPT-3.5 and GPT-4 [49]. In the first experiment, we inputted the dialog context into the model without providing any additional information and let the model generate the reply. In the second experiment, We designed a prompt template that encompasses the principles of the EFCP methodology.

TABLE 8. Results of automated metrics for the Large language model equipped EFCP methodology.

Models	R-L	B-1	B-2	BERT	GMS	ESA	VES	GM
GPT-3.5	12.49	10.75	3.30	84.12	65.90	84.87	40.88	26.22
GPT-3.5+EFCP	**12.89** (+0.40)	**10.80** (+0.05)	**3.63** (+0.33)	**84.19** (+0.07)	**67.12** (+1.22)	**85.76** (+0.89)	**42.03** (+1.15)	**26.94** (+0.72)
GPT-4	11.87	10.00	2.77	83.78	64.67	84.32	39.02	24.86
GPT-4+EFCP	**12.69** (+0.82)	**10.57** (+0.57)	**3.12** (+0.35)	**84.09** (+0.31)	**65.56** (+0.89)	**85.27** (+0.95)	**40.87** (+1.85)	**26.00** (+1.14)

TABLE 9. Case study of the generated responses by models.

Speaker Post	i've been clean and sober for three years today! i'm a good dad today and i couldn't be happier with this life i live!
Common sense	get a job; to be happy; get married; happy; to be a good parent; to be a good person; to be a good father; to have a good life; to live a good life; to be a good parent; to live a good life; to be a good person; to be a good father; to have a good life; happy; proud; satisfied; good; content; to be a good parent; to be a good father; to have a good life; to be a good dad; to be a good husband
Responder Original Persona	i love this man .
Rewritten Persona	i am loving . i am affectionate . i am romantic . i am caring . i am loving . i have a partner . i am loved back . i have a relationship . i get married . i gets married . i want to be loved . i want to be happy . i want to be with . i want to love . i want to get to know him . i want to get to know them . i want to be in a relationship . i want to have a relationship . i want to like him . i feel happy . i feel loved . i feel loving . i feel romantic . i feel good . i want to have a family . i want to marry him . i want to have a relationship . i want to get married . i want to spend time together .
empathy factors	emotional reaction: yes exploration: no interpretation: no dialog act: wishing emotion: admiration
GPT2	congrats man! i'm happy for you. i hope that life gets better soon for your family and friends.
CoMAE	congratulations!! may you have many happy years ahead of them!
CEM	congratulations! i'm so happy for you and your family.
GPT2-sw	congratulations, man! i'm proud of you. your life is a great one!
GPT2-linear	congratulations man, i'm so happy for you!
EFCP	congrats brother! that's a great achievement and i wish you many years of sobriety to come.
Gold	congrats man! keep it up!

TABLE 10. Prompt templates.

Common prompt template for LLMs Please reply to the following sentence. Sentence: { }
EFCP prompt template for LLMs Speaker: { } Additional information related to the above sentence: 1. The intent of the speaker's statement is: { } 2. The emotion of the speaker's statement is: { } 3. The common sense involved in the sentence is: { } 4. Your personal information is: { } 5. Do you want to express emotions: { } 6. Do you want to speculate about the speaker's feelings and experiences based on his words: { } 7. Do you want to speculate about the feelings and experiences that the speaker doesn't mention in his words: { } 8. Dialog intent that your response should have: { } 9: Emotion that your response should have: { } The task you need to complete: 1. First, you must combine the first three points of additional information to understand the speaker. 2. Then, you need to generate a response by combining the last six points of additional information. 3. Finally, please give me a direct response.

The prompt templates we used are shown in Table 10. The results presented in Table 8 demonstrate the effectiveness of the EFCP method in enhancing the automatic evaluation metrics of Large language models.

However, when compared to the models in Table 3, the automatic metric results for the Large language models are considerably lower. This is because the large models have not been fine-tuned on the PEC dataset. The computational and financial resources required for fine-tuning these Large language models with hundreds of billions (GPT-3.5) or even trillions (GPT-4) of parameters are unaffordable. To address this issue, certain prompt templates can be designed to guide large language models to perform a specific task [50], [51], [52], [53], which is the approach adopted in our experiments. Another reason for the low word overlap metrics observed in the Large language models is that humans often use informal language in their responses (e.g. *congrats*), while these models tend to generate more formal language (e.g. *congratulations*). This results in a lower word overlap metrics.

VI. CONCLUSION

Our proposed model, EFCP, generates replies by combining common sense, empathy factors, and the personal information of the responder. Our experiments have shown that EFCP outperforms the baseline, as it better understands the context and produces responses that conform to the responder's role characteristics. And EFCP method also has a positive effect on the existing large language models. We suggest that the responder's personal information be incorporated when predicting the empathy factors they should adopt. Our ablation experiments have provided evidence that this approach is reasonable. By optimizing one part of the process or combining it with other dialogue information, we believe that even better results can be achieved using this architecture.

REFERENCES

[1] A. Radford, J. Wu, and R. Child. *Language Models Are Unsupervised Multitask Learners*. Accessed: Apr. 17, 2019. [Online]. Available: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

[2] T. Young, D. Hazarika, S. Poria, and E. Cambria, "Recent trends in deep learning based natural language processing," *IEEE Comput. Intell. Mag.*, vol. 13, no. 3, pp. 55–75, Aug. 2018, doi: 10.1109/MCI.2018.2840738.

[3] Y. Cao, "Pretrained language models for dialogue generation with multiple input sources," in *Proc. Findings EMNLP*, 2020, pp. 909–917, doi: 10.18653/v1/2020.findings-emnlp.81.

[4] L. Zhou, "The design and implementation of Xiaoice, an empathetic social chatbot," *Comput. Linguistics.*, vol. 46, no. 1, pp. 53–93, Mar. 2020, doi: 10.1162/coli_a_00368.

[5] S. Liu, "Towards emotional support dialog systems," in *Proc. ACL*, 2021, pp. 3469–3483, doi: 10.18653/v1/2021.acl-long.269.

[6] A. Sharma, "A computational approach to understanding empathy expressed in text-based mental health support," in *Proc. EMNLP*, 2020, pp. 5263–5276, doi: 10.18653/v1/2020.emnlp-main.425.

[7] B. L. Omdahl, *Cognitive Appraisal, Emotion, and Empathy*. London, U.K.: Psychology Press, Mar. 2014. [Online]. Available: https://www.routledge.com/Cognitive-Appraisal-Emotion-and-Empathy/Omdahl/p/book/9781138970984

[8] A. Paiva, I. Leite, H. Boukricha, and I. Wachsmuth, "Empathy in virtual agents and robots: A survey," *ACM Trans. Interact. Intell. Syst.*, vol. 7, no. 3, pp. 1–40, Sep. 2017, doi: 10.1145/2912150.

[9] F. de Vignemont and T. Singer, "The empathic brain: How, when and why?" *Trends Cognit. Sci.*, vol. 10, no. 10, pp. 435–441, Oct. 2006, doi: 10.1016/j.tics.2006.08.008.

[10] A. Welivita and P. Pu, "A taxonomy of empathetic response intents in human social conversations," in *Proc. 28th Int. Conf. Comput. Linguistics*, Barcelona, Spain, 2020, pp. 4886–4899, doi: 10.18653/v1/2020.coling-main.429.

[11] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 5370–5381, doi: 10.18653/v1/p19-1534.

[12] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, and S. Ravi, "GoEmotions: A dataset of fine-grained emotions," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, Jul. 2020, pp. 4040–4054, doi: 10.18653/v1/2020.acl-main.372.

[13] S. Sabour, C. Zheng, and M. Huang, "CEM: Commonsense-aware empathetic response generation," in *Proc. Conf. Artif. Intell. (AAAI)*, Feb. 2022, vol. 36, no. 10, pp. 11229–11237, doi: 10.1609/aaai.v36i10.21373.

[14] L. Wang, J. Li, Z. Lin, F. Meng, C. Yang, W. Wang, and J. Zhou, "Empathetic dialogue generation via sensitive emotion recognition and sensible knowledge selection," in *Proc. Findings Assoc. Comput. Linguistics: EMNLP*, Abu Dhabi, United Arab Emirates, 2022, pp. 4634–4645, doi: 10.18653/v1/2022.findings-emnlp.340.

[15] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 2204–2213, doi: 10.18653/v1/p18-1205.

[16] M. R. Leary and A. B. Allen, "Personality and persona: Personality processes in self-presentation," *Wiley Online Library.*, vol. 79, no. 6, pp. 1191–1218, Dec. 2011, doi: 10.1111/j.1467-6494.2010.00704.x.

[17] H. Zhou, "Emotional chatting machine: Emotional conversation generation with internal and external memory," in *Proc. AAAI*, New Orleans, LO, USA, 2018, pp. 730–739, doi: 10.1609/AAAI.V32I1.11325.

[18] R. Zandie and M. H. Mahoor, "EmpTransfo: A multi-head transformer architecture for creating empathetic dialog systems," 2020, *arXiv:2003.02958*.

[19] C. Zheng, "CoMAE: A multi-factor hierarchical framework for empathetic response generation," in *Proc. Findings ACL*, 2021, pp. 813–824, doi: 10.18653/v1/2021.findings-acl.72.

[20] S. Wu, Y. Li, D. Zhang, Y. Zhou, and Z. Wu, "TopicKA: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact," in *Proc. Twenty-Ninth Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 3766–3772, doi: 10.24963/ijcai.2020/521.

[21] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[22] P. Zhong, C. Zhang, H. Wang, Y. Liu, and C. Miao, "Towards persona-based empathetic conversational models," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 6556–6566, doi: 10.18653/v1/2020.emnlp-main.531.

[23] A. Bosselut, H. Rashkin, M. Sap, C. Malaviya, A. Celikyilmaz, and Y. Choi, "COMET: Commonsense transformers for automatic knowledge graph construction," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 4762–4779, doi: 10.18653/v1/P19-1470.

[24] S. Welleck, J. Weston, A. Szlam, and K. Cho, "Dialogue natural language inference," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 3731–3741, doi: 10.18653/v1/p19-1363.

[25] Y. Cao, W. Bi, M. Fang, S. Shi, and D. Tao, "A model-agnostic data manipulation method for persona-based dialogue generation," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, Dublin, Ireland, 2022, pp. 7984–8002, doi: 10.18653/v1/2022.acl-long.550.

[26] C. Sankar, S. Subramanian, C. Pal, S. Chandar, and Y. Bengio, "Do neural dialog systems use the conversation history effectively? An empirical study," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, Florence, Italy, 2019, pp. 32–37, doi: 10.18653/v1/p19-1004.

[27] U. Khandelwal, H. He, P. Qi, and D. Jurafsky, "Sharp nearby, fuzzy far away: How neural language models use context," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, Melbourne, VIC, Australia, 2018, pp. 284–294, doi: 10.18653/v1/p18-1027.

[28] J. D. Hwang, C. Bhagavatula, R. Le Bras, J. Da, K. Sakaguchi, A. Bosselut, and Y. Choi, "(Comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, 2020, pp. 6384–6392, doi: 10.1609/AAAI.V35I7.16792.

[29] S. M. Peloquin, "The fullness of empathy: Reflections and illustrations," *Amer. J. Occupational Therapy*, vol. 49, no. 1, pp. 24–31, Jan. 1995, doi: 10.5014/ajot.49.1.24.

[30] M. Sap, "Atomic: An atlas of machine commonsense for if-then reasoning," in *Proc. AAAI Conf. Artif. Intell.*, Honolulu, HI, USA, 2019, vol. 33, no. 1, pp. 3027–3035, doi: 10.1609/AAAI.V33I01.33013027.

[31] B. P. Majumder, H. Jhamtani, T. Berg-Kirkpatrick, and J. McAuley, "Like hiking? You probably enjoy nature: Persona-grounded dialog with commonsense expansions," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 9194–9206, doi: 10.18653/v1/2020.emnlp-main.739.

[32] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[33] T. Wolf, "Transformers: State-of-the-art natural language processing," in *Proc. EMNLP*, 2020, pp. 38–45, doi: 10.18653/v1/2020.emnlp-demos.6.

[34] A. Holtzman, J. Buys, L. Du, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," 2019, *arXiv:1904.09751*.

[35] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.

[36] M. Popel and O. Bojar, "Training tips for the transformer model," *Prague Bull. Math. Linguistics*, vol. 110, no. 1, pp. 43–70, Apr. 2018, doi: 10.2478/pralin-2018-0002.

[37] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Austin, TX, USA, 2016, pp. 2122–2132, doi: 10.18653/v1/d16-1230.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics ACL*, Philadelphia, PA, USA, 2001, pp. 311–318, doi: 10.3115/1073083.1073135.

[39] C.-Y. Lin, "ROUGE: A package for automatic evaluation of summaries," in *Proc. Text Summarization Branches Out*, Barcelona, Spain, 2004, pp. 74–81.

[40] V. Rus and M. Lintean, "A comparison of greedy and optimal assessment of natural language student input using word-to-word similarity metrics," in *Proc. 7th Workshop Building Educ. Appl. Using NLP*, Montréal, QC, Canada, 2012, pp. 157–162.

[41] J. Wieting, M. Bansal, K. Gimpel, and K. Livescu, "Towards universal paraphrastic sentence embeddings," 2015, *arXiv:1511.08198*.

[42] G. Forgues, "Bootstrapping dialog systems with word embeddings," in *Proc. NIPS Modern ML+NLP*, Montreal, QC, Canada, 2014, pp. 168–172.

[43] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," 2019, *arXiv:1904.09675*.

[44] F. Xiao, L. Pang, Y. Lan, Y. Wang, H. Shen, and X. Cheng, "Transductive learning for unsupervised text style transfer," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Punta Cana, Dominican Republic, 2021, pp. 2510–2521, doi: 10.18653/v1/2021.emnlp-main.195.

[45] J. Lee, "Stable style transformer: Delete and generate approach with encoder–decoder for text style transfer," in *Proc. INLG*, Dublin, Ireland, 2020, pp. 195–204, doi: 10.18653/v1/2020.inlg-1.25.

[46] R. Liu, C. Gao, C. Jia, G. Xu, and S. Vosoughi, "Non-parallel text style transfer with self-parallel supervision," 2022, *arXiv:2204.08123*.

[47] H. Zhang, "TransOMCS: From linguistic graphs to commonsense knowledge," in *Proc. IJCAI*, Yokohama, Japan, 2020, pp. 4004–4010, doi: 10.24963/IJCAI.2020/554.

[48] W. Peng, Y. Hu, L. Xing, Y. Xie, Y. Sun, and Y. Li, "Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation," in *Proc. 31st Int. Joint Conf. Artif. Intell.*, Vienna, Austria, Jul. 2022, pp. 4324–4330, doi: 10.24963/ijcai.2022/600.

[49] OpenAI et al., "GPT-4 technical report," 2023, *arXiv:2303.08774*.

[50] X. Yang, Y. Li, X. Zhang, H. Chen, and W. Cheng, "Exploring the limits of ChatGPT for query or aspect-based text summarization," 2023, *arXiv:2302.08081*.

[51] C. Qin, A. Zhang, Z. Zhang, J. Chen, M. Yasunaga, and D. Yang, "Is ChatGPT a general-purpose natural language processing task solver?" 2023, *arXiv:2302.06476*.

[52] S. Frieder, L. Pinchetti, A. Chevalier, R.-R. Griffiths, T. Salvatori, T. Lukasiewicz, P. C. Petersen, and J. Berner, "Mathematical capabilities of ChatGPT," 2023, *arXiv:2301.13867*.

[53] W. Jiao, W. Wang, J.-T. Huang, X. Wang, S. Shi, and Z. Tu, "Is ChatGPT a good translator? Yes with GPT-4 as the engine," 2023, *arXiv:2301.08745*.

WEIJIE LI received the bachelor's degree in resource exploration from the China University of Mining and Technology, in 2020. He is currently pursuing the master's degree in software engineering with the Software School, Xinjiang University. His research interests include text generation and sentiment analysis.

YONG YANG received the Ph.D. degree in computer applications from the University of Chinese Academy of Sciences, in 2013. He is currently a Professor with the College of Computer Science and Technology, Xinjiang Normal University. His research interests include natural language processing and intelligent education.

PALIDAN TUERXUN received the Ph.D. degree in computer software and theory from Northwestern University. She is currently an Associate Professor with Xinjiang Normal University. Her research interest includes multimodal information processing.

XIAOCHAO FAN received the Ph.D. degree in computer application technology from the Dalian University of Technology, in 2021. He is currently an Associate Professor with the College of Computer Science and Technology, Xinjiang Normal University. His research interests include sentiment analysis, text mining, and text generation.

YUFENG DIAO received the Ph.D. degree in computer application technology from the Dalian University of Technology, in 2020. She is currently an Associate Professor with the College of Computer Science and Technology, Inner Mongolia Minzu University. Her research interests include text mining and sentiment analysis.