APSIPA Transactions on Signal and Information Processing, 2023, 12, e40

Original Paper

Conditional Adversarial Learning for Empathetic Dialogue Response Generation

Ming-Hsiang Su[1], Chung-Hsien Wu[2*] and Chia-Yu Liao[2]

[1] *Department of Data Science, Soochow University, Taipei, Taiwan*
[2] *Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan*

ABSTRACT

This study presents a novel approach for generating empathetic responses in dialogue through conditional adversarial learning. The method involves using a BERT-MLP model to detect the user's emotions and the system's dialogue act, and then utilizing conditional adversarial learning to construct a generator based on the user's emotions, dialogue history, and dialogue act. A sympathy discriminator is trained to distinguish between empathetic and non-empathetic responses, and the corresponding words are filled in the generated template based on the semantic slots. To evaluate the proposed approach, the study collected 1,740 conversations with empathetic responses, which were labeled with the user's emotion, medical history, and system dialogue act. The experimental results based on 5-fold cross-validation showed that the proposed method of applying conditional adversarial learning achieved the best BLEU score (41.3%), the best BERTSCORE (−5.84 for the evaluation on question sentences and generated sentences; −4.15 for the evaluation on answer sentences and generated sentences) and emotion reflection rate (86.4%), which outperformed the Transformer- and conditional Transformer-based methods. This study also conducted subjective evaluations, achieving 77.55%, 79.47%, and 75.87% accuracy

*Corresponding author: Chung-Hsien Wu, chunghsienwu@gmail.com.

Received 02 December 2022; Revised 02 July 2023
ISSN 2048-7703; DOI 10.1561/116.00000081

in the scores of relevance, grammatical correctness, and empathy, respectively. In significance test and Cohen's KAPA score of relevance, grammatical correctness and empathy, the proposed method was better than the Transformer- and conditional Transformer-based methods. In addition, in the consulting performance evaluation, the experimental results showed that the proposed method achieved the best empathy score of 3.8 (average KAPA score was 0.627), which was better than the other methods.

Keywords: Dialogue system, Empathy, Adversarial training.

1 Introduction

In recent decades, spoken dialogue systems have gained popularity among those seeking additional assistance and have been extensively developed in various areas such as ticket booking, hotel reservations, and interview coaching [7, 33–35, 41]. Generally, dialogue systems can be categorized as task-oriented or non-task-oriented based on their intended use. Task-oriented systems are designed to complete specific tasks through conversations with users, such as service reservations or product inquiries. Non-task-oriented systems, on the other hand, provide companionship and entertainment by conversing with users, like chatbots. Dialogue systems use natural language to simulate human conversation, and the system is expected to provide grammatically and semantically consistent responses after receiving the user's input. Despite the success of neural dialogue models in generating responses, they often produce generic and uninteresting replies. While spoken dialogue systems can respond to user queries, they lack the ability to express emotions in their responses.

This study considers the user's health status, including diseases and symptoms, as a means of understanding user behavior. Our approach takes into account the user's personal medical history, emotions, and health-related events to produce empathetic responses. There are two main issues that we address in generating empathetic response sentences. First, current empathetic dialogue systems incorporate user emotional states and personalities into the generative model to generate appropriate responses. Second, to meet the criteria for empathy, this study combines predicted user emotions and personal medical histories as inputs for empathetic response generation, utilizing the adversarial method to accurately reflect user emotions and consider response fluency based on a template-based transformer. To address these issues, this paper employs a Transformer-based generator with conditions on emotion, dialogue act, and personal medical history. To account for the lack of consideration of the user's experience in empathic practice, we use the user's personal medical history

related to health topics as the generation condition for the Transformer-based generation model. Furthermore, to address the problem of dialogue systems generating responses that do not conform to the user's emotions, we use the user's emotion as a condition for the generator and construct a discriminator to guide the generation direction in the conditional adversarial training process. This approach is an improvement over previous models that did not incorporate empathetic techniques rooted in psychology to address the user's emotions and circumstances.

2 Literature Review

2.1 Empathy

In human-human conversations, psychological studies have shown that people experience positive mood [36, 39] or have friendly feelings toward their partner when the partner shows their empathy [12, 38]. In addition, empathy plays an important role in communication with others. It can correctly understand the user's emotions and give an appropriate response after correctly interpreting the user's behavior. Understanding of emotion and semantics are inseparable. Studies have shown that the use of empathic virtual assistants can help improve human-computer interaction [34] and increase user satisfaction and participation [7, 41]. Especially, in all groups, empathy is important in communicating with elderly people. Using empathy and communication skills is helpful to accurately understand and express each other's feelings and meanings, and to listen to their needs. In this study, an empathetic dialogue system is proposed and applied to provide comfort and heart-warming response to the users.

In the concept of empathy, Carkhuff [2] divided empathy into two levels: low and high, which respectively aimed to understand the implicit experience, feelings, and behaviors of the other party. Table 1 is a dialogue response example with and without empathy. An empathetic response allows the recipient to feel that we understand his/her physical and mental health and a chance to talk more about what is truly bothering him/her.

Empathy is the ability to identify other people's experiences, feelings, and behaviors, and express these basic understanding to them [2]. In terms of feeling, we try to understand the user's feelings by detecting the user's emotions and use the detected emotions as one of the conditions for generating empathetic response. In recent years, some dialogue systems have been designed to detect emotion from different signal sources to make the empathetic virtual robots to be more human-like in their interactions [7, 32, 40]. In [7], Zara, a prototype system of an empathetic virtual robot, was designed to recognize user emotions and the most significant step was to make robots to be more

Table 1: An example for comparison of responses w/o empathy.

User	我最近一直咳嗽，真是不舒服！(I've been coughing recently and it's uncomfortable!)
Without Empathy	你今天趕緊去看醫生吧！(Go to the doctor!)
With Empathy	我能理解你的難過，你咳嗽會不會是因為接觸到過敏原呢? (I can understand your pain. Is your cough due to exposure to allergens?)

human-like in their interactions, expecting that future robots will be more compassionate and will not cause harm to humans in their interaction with machines. Siddique *et al.* [32] proposed the enhanced personality module of Zara with improved performance of the recognition based on speech and text using deep learning frameworks. In their framework, empathy analysis which includes emotion recognition, sentiment analysis and personality analysis was used for langue understanding. In [40], empathy analysis on the designed system, Empathetic Psychologist Nora, which considered emotion recognition, sentiment analysis, stress detection and personality analysis was also used for langue understanding. Nora understands, empathizes, and adapts to users using emotional intelligence modules which enable Nora to personalize the content of each conversation. Rashkin *et al.* [28] proposed a new dataset of 25K dialogues grounded in situations prompted by specific emotion labels. Their experiments showed that using this dataset to fine-tune conversation models leads to responses that are more empathetic with evaluation by humans.

In addition, many scholars have focused on the research on the generation of empathic responses, hoping to generate empathic and fluent response sentences [15–18, 29, 31, 42]. Rashkin *et al.* [29] proposed a novel dataset in emotional situations and a new benchmark for empathetic dialogue generation. Their experiments indicated that dialogue models that used the proposed dataset were perceived to be more empathetic by human evaluators. Li *et al.* [15] proposed the EmpDG model to generate more empathic responses, where the model exploited both coarse-grained dialogue-level and fine-grained token-level emotions. The model also used an interactive adversarial learning framework to identify whether the generated responses evoked emotional perception in the dialogue. Lin *et al.* [17] proposed a novel end-to-end MoEL for modelling empathy in dialogue systems. Human evaluation on the EMPATHETICDIALOGUES dataset [29] confirmed that MoEL outperformed the multi-task training baselines in terms of empathy, relevance, and fluency. Shin *et al.* [31] proposed Sentiment Look-ahead to simulate future user emotional states and showed that their proposed method produced responses that were more empathetic, relevant, and fluent than other competing baselines.

Based on prior studies [7, 15–18, 28, 29, 31, 32, 40, 42], we put forward a conditional adversarial learning framework that takes into account user emotions and experience expectations to produce response sentences that align with Carkhuff's concept of empathy [2].

2.2 Emotion Recognition

At present, some dialogue systems have been designed to detect emotion from facial expression, speech, text, and personality to help generate empathetic responses [7, 32, 40]. Su *et al.* [34] analyzed the human-machine dialogue in specific situations. They set emotional stimulus conditions related to user behavior and determined the emotion expressed by the system in specific situations. Fung *et al.* [8] used sentences in the Twitter database with emojis as emotion tags to train the emotional embedding. They hoped that the dialogue system can feel the user's emotion and make human-computer interaction more empathetic.

Voice-based emotion recognition [13, 20] and image-based emotion recognition [10, 26] have yielded positive results in research. Recently, detecting emotions in textual and spoken dialogues has gained attention as a research topic [3]. Text emotion detection methods can be categorized as supervised and unsupervised. Supervised methods typically use hashtags, emoticons, emotional markers, and other training labels, and utilize machine learning models to classify emotional features [3, 22, 27, 30]. For instance, [3, 22, 27, 30] employed emotion-related tags on Twitter as training labels and utilized a support vector machine (SVM) as a binary classifier for emotion classification. Besides, Hasan *et al.* [11] used an emotion lexicon in addition to hashtags as emotional training labels and compared the performance of various common machine learning algorithms such as Naïve Bayes, SVM, Decision Trees, K-Nearest Neighbor (KNN) for emotion classification. Their experimental results showed that Decision Trees and Naïve Bayes achieved the highest accuracy using all proposed features. However, SVM achieved the highest accuracy by using unigrams, and KNN achieved the highest accuracy by using unigrams and negations.

Unsupervised methods for text emotion detection focus on identifying specific words in sentences, such as nouns, verbs, adjectives, and adverbs, and then estimating the emotion vector of the sentences based on the semantic similarity of these words [1]. On the other hand, supervised methods often use pre-labeled data and machine learning models, such as BERT-MLP, to classify the emotional features of the text [5]. In particular, the use of Transformers [37] in BERT-MLP models can capture the bidirectional features of the text and has been shown to achieve good performance in sentence emotion classification, making it a suitable model for this study.

2.3 Natural Language Generation

Natural language generation is a challenging task due to the ambiguity and polysemy of texts and dialogues. To ensure smoothness, readability, and information content of generated sentences, there are three main methods: template-based, retrieval-based, and neural generative models. Template-based models set rules to regulate response timing and content, resulting in more fluent and high-quality statements but requiring significant manual effort. Retrieval-based models calculate similarity between input and candidate response sentences to generate more grammatical responses, but with low variability. Generative models, the most common approach, include the sequence-to-sequence model, generative adversarial network, and attention mechanism. In a previous study, an adversarial learning framework was proposed to generate conditional responses with improved response quality and controllability by using dialogue acts as features to discriminate generated results that do not follow the given dialogue act. This study integrates the Transformer-based model and adversarial learning method to generate response sentences that balance fluency and user emotion. The Transformer model's multi-head self-attention architecture captures interrelationships between words in a sentence, and parallelization improves computational efficiency. The discriminator considers medical history, user emotion, and dialogue acts as features to distinguish fake generated results that do not reflect user emotion.

3 Database Design and Collection

This study adopts the definition of empathy from [2], which describes it as a complex ability to comprehend and share the emotional states of others, leading to compassionate behavior. Empathetic dialogue systems are the focus of this research, and Table 2 provides examples of empathetic responses given by

Table 2: An Example of the Empathetic Dialogue.

User	我曾找過其他輔導員談話，但根本毫無用處，我也不了解我為何還要再來。但事情實在很糟，我想必須有所行動，所以我又來試試看。(I have talked to other counselors, but it is useless. I don't know why I must come again. But things are really bad, I think I have to act, so I am here again.)
Counselor	你現在相當矛盾，因為你不知道我們再談下去會有什麼結果或用處，但你覺得你必須再試試。(You are quite lost now, because you don't know whether counseling would be helpful, but you think you must try again.)

counselors. The responses show that the counselor's statement contains both emotional and factual aspects. The research collects data on users' medical conditions and symptoms through dialogues. Furthermore, certain rules are established as a reference for compiling an empathy database for the study.

The aim of this study is to develop an empathetic dialogue system that addresses health issues affecting elderly individuals. To collect the necessary corpus, we focused on daily conversations related to health topics commonly discussed by middle-aged and elderly people. We invited 10 participants to engage in character dialogue simulations while following these guidelines:

1. Participants received training in the required labeling skills and the correct labeling method.
2. Participants simulated conversations with senior individuals.
3. Participants labeled the emotion and personal medical history of the user's response.
4. Participants marked the dialogue act of the system's response.
5. The system's response included an expression of the user's received emotions and events, when the user showed a non-neutral emotion and mentioned a health-related event such as a disease or symptom.
6. If the user mentioned health-related issues, the system referred to the user's medical history and healthcare-related websites to respond.

To collect the corpus, the participants were invited to simulate character dialogues, following the specifications we designed. We ensured a balanced and clear representation of different emotions in the corpus design, and made sure that the users' language exhibited a variety of emotions. The dialogue part is collected turn by turn. The user input sentence and the system response sentence are defined as one dialogue turn. A total of 1,740 dialogue turns were collected, and the average collection time of one dialogue round is about three minutes. The total duration of the dialogues is about 83 hours. The average number of sentences in the user turn is 2.01, the average sentence number in the dialog system turn is 2.55, and the vocabulary size is 2487. The database includes labels consisting of 4 fundamental emotion types (happiness, neutrality, anger, and sadness) [23], 4 system dialogue acts (cause query, event response, suggestion, chitchat), 2 dialogue slots (Medical History, Symptom), and 17 medical history categories, which include ailments like myocardial infarction, heart disease, cataract, as well as no disease history. The selection of the numbers for system dialogue acts and dialogue slots was based on the dataset's content, while the number of medical history categories was determined according to the diseases commonly found in the elderly population

provided by KingNet [14]. For the collection of the corpus, each dialogue comprises a conversation between two individuals: one acting as an elderly person, and the other as the system. The participant who portrays the elderly person marks the emotion of their responses, while the participant who plays the system marks the dialogue acts they use. Both participants must agree on the emotion and dialogue labels for corpus annotation. Only dialogues with agreed-upon labels are included in the corpus. Prior to corpus collection, the user's medical history is recorded through questioning. The collected corpus replaces the Slot-Value words with corresponding slot tags, which are manually designed. The final corpus consists of 3480 templates, and an example of the collected empathetic dialogues is shown in Table 3.

Table 3: An Example of the Collected Empathetic Dialogues.

Speaker	Text	Dialogue Act	Emotion	Personal Medical History
User	講得我都擔心了，會不會是因為黃斑部病變造成我看不清楚啊！(I'm worried because I can't see clearly, which might be caused by macular degeneration!)		Sadness	Macular Degeneration
System	我知道你會擔心，看不清楚可能是因為年紀大，建議多吃綠色蔬果及魚類，補充葉黃素與 omega-3。(I know you're worried. If you can't see clearly, it may be that you are older. I recommend eating more fruits, vegetables, and fish, and taking with lutein and omega-3 supplement.)	Suggestion		

4 System Framework

Figure 1 illustrates the architecture of the proposed empathetic dialogue system, which comprises three parts: empathy analysis, dialogue management, and template generation, during the training phase. The empathy analysis module extracts the user's medical history and emotions, where the medical history is provided by the user in advance. As an example, the user may provide information about their medical history, such as hyperlipidemia (高血脂), which is recorded during the corpus collection. In emotion detection, the user sentences and the corresponding emotion label are used as the training

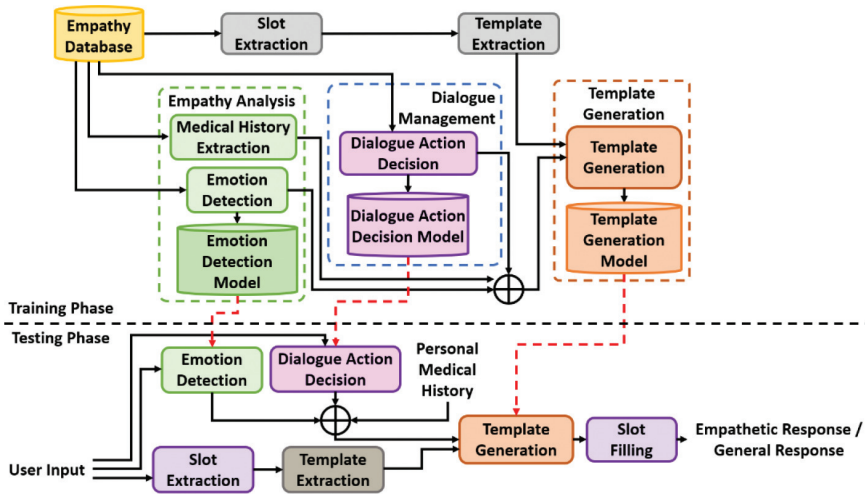

Figure 1: The System Architecture of the Proposed Empathetic Dialogue System.

database of the emotion classification model. For example, the user's response "I'm worried because I can't see clearly, which might be caused by macular degeneration!" is labeled as "sadness", as shown in Table 3. This study extracts dialogue slots from the keywords in the user sentences and then establish a corresponding slot-value table by the maximum matching algorithm; for example, the Slot "Medical History" has Values "Heart disease, diabetes, high blood pressure, hyperlipidemia, etc." The slot tags are predetermined manually, but for domain transfer, a crawler can be employed to extract the top keywords from other domains to obtain relevant slot tags. As for dialogue management, this study utilizes BERT-MLP-based dialogue act decision model to train on user sentences and their corresponding dialogue act labels as inputs. In template generation, this study uses the slot-value table to replace the keywords in user sentences and the corresponding system responses with slot categories to obtain user and system response templates (Figure 2).

For example, the user sentence "我這幾天一直嗜睡，讓我有點煩躁 (I have been sleepy for a few days, which makes me a little annoyed.)" is replaced with "我這幾天一直<symptom>，讓我有點煩躁 (I have been <symptom> for a few days, which makes me a little annoyed.)" and the system response "我能理解你為什麼生氣，但不知道你嗜睡是因為你有熬夜. (I can understand why you are angry, but I don't know that your sleepiness is because you stay up late.)" is replaced with "我能理解你為什麼生氣，但不知道你<symptom> 是因為你有熬夜. (I can understand why you are angry, but I don't know that your <symptom> is because you stay up late.)"

Figure 2: An Example of Template Generation.

The Transformer-based template generation model takes in the user's response sentence template, emotion, medical history, and expected system dialogue act as inputs. The system template is then used as the output for model training. During the test phase, the user's sentence is analyzed by the empathy analysis and dialogue management modules to obtain reference conditions such as user emotions, medical history, and system dialogue act. These conditions are then used to generate the response sentence template using the Transformer model. The response sentence template is filled with the slot value to produce the final empathetic response using the slot-value table. Figure 3 illustrates an example of the response sentence generation process.

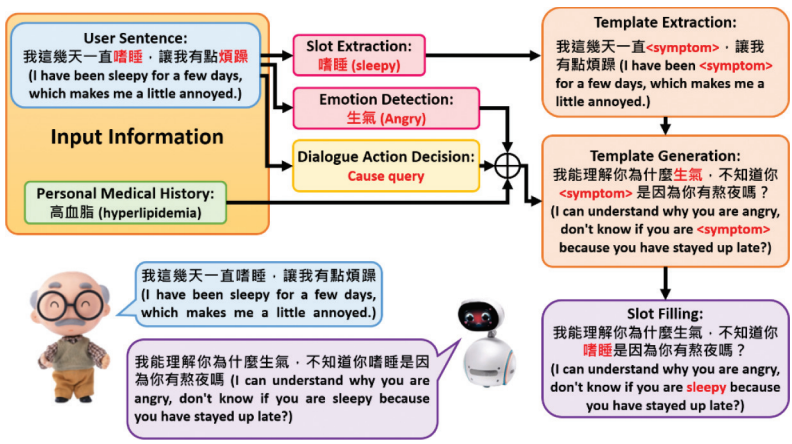

Figure 3: An Example of Response Sentence Generation Process.

4.1 Empathy Analysis Module

The empathy analysis module aims to acquire information such as the user's medical history, events or slot values in their sentence, and their emotions to aid in generating empathetic response sentences. To begin, users are asked to input their medical history prior to the conversation, making it a known factor. Slot extraction and user emotion recognition are the two steps involved in obtaining the other necessary information. In the former, dialogue semantic slots are defined as diseases (e.g. Heart Disease, Diabetes, Hypertension, Hyperlipidemia) and symptoms (e.g. Headache, Dizzy, Chest Pain, Tired) related to health issues. User sentence emotion recognition involves capturing sentence characteristics to classify user emotions, which is a classification task. To identify user sentence emotion, the BERT-MLP-based method is used, a pre-training language model developed by Google that performs well on various

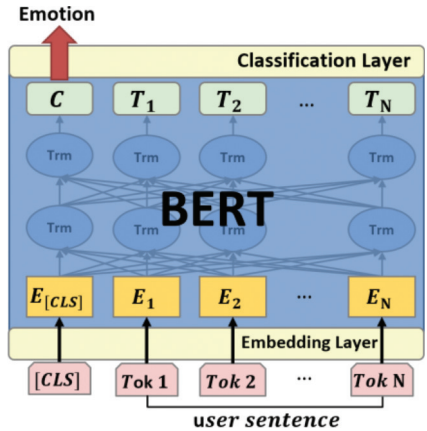

Figure 4: The System Architecture of the BERT-MLP-based Model for Fine-tuning.

natural language classification tasks [20]. To fine-tune the BERT-MLP-based model parameters for the emotion classification task, the pre-trained Chinese BERT-MLP-based model and the collected database are utilized. The input for the emotion classification model consists of the user's sentence and the corresponding emotion label.

For fine-tuning the Chinese pre-trained BERT-MLP-based model for emotion detection, the fine-tuning formula is shown in (1), where $C \in \mathrm{R}^{\mathrm{H}}$ is the word representation of the special classification symbol [CLS], $W \in \mathrm{R}^{\mathrm{B} \times \mathrm{H}}$ is the added linear layer weight, H is the hidden layer size and B is the number of emotion labels. Figure 4 illustrates the process for fine-tuning the BERT-MLP-based model for user sentence emotion classification. The input of the BERT-MLP-based model is the user sentence, and the special classification symbol [CLS] is concatenated to the front of the user sentence. The bidirectional context representation of each word is obtained by the BERT-MLP-based model. Finally, the bidirectional contextual representation of [CLS] of the last hidden layer of the model is fed to the linear classification layer for user emotion recognition.

$$P = softmax(CW^T) \tag{1}$$

4.2 Dialogue Management

The primary objective of dialogue management is to determine the appropriate dialogue act for the system to undertake when presented with the user's input sentence. To achieve the system dialogue act classification task for a single sentence, the Chinese pre-trained BERT-MLP-based model is fine-tuned. The model's input consists of the classification symbol [CLS] concatenated

with the user's sentence, and the system dialogue act associated with the user's sentence serves as the training label. Once the BERT-MLP-based model acquires the bidirectional contextual representation of each word, the contextual representation of [CLS] is fed into the linear layer to determine the suitable dialogue act for the system to respond to the user's sentence input.

4.3 Response Generation

Before the training of the response generation model, we first pre-process the user sentences and response sentences in the collected database. In this study, we selected 17 diseases and 209 symptoms from the database. The Jieba Chinese word segmentation toolkit is employed to segment the user sentence and the corresponding response sentences. Then, we use the maximum matching method to replace the words in the sentence with the corresponding slot tag. It is possible to have multiple slots in a sentence (Figure 5).

For example, the sentence "最近常常打噴嚏、流鼻水，超不舒服的 (I often sneeze and have runny nose recently, it's very uncomfortable)" has two slots "打噴嚏 (sneeze)" and "流鼻水 (runny nose)". After all the user sentences and the corresponding response sentences have been processed, the response templates are obtained and used to train the response template generative model.

Figure 5: An Example of Multiple slots in a sentence.

In this study, a conditional generation approach is employed for response generation. The conditional generative adversarial model takes into account the user's emotions, personal medical history, and the expected system dialogue acts as input for the generation model. The goal is for the system to generate responses that are consistent with the user's emotions and closely resemble natural responses. Generative adversarial networks (GANs) [9] are used to model data distributions and consist of two functions: the generator, which converts a sample from a random uniform distribution to the data distribution, and the discriminator, which assesses the probability of whether a given sample belongs to the data distribution or not. By employing the game-theoretic min-max principle, the generator and discriminator are learned jointly by alternating the training of the generator and the discriminator [43].

To enable the discriminator to classify both emotional categories and the authenticity of the data, we utilize the auxiliary classifier adversarial network (AC-GAN) [24]. In AC-GAN, each generated sample has a corresponding class label, c, in addition to the noise z. Generator uses both inputs to generate data $X_{fake} = G(cz)$. Discriminator gives both inputs a probability distribution over sources and a probability distribution over the class labels. The objective function has two parts: the log-likelihood of the correct source,

L_S, and the log-likelihood of the correct class, L_C, as shown in (2) and (3). The discriminator's purpose is to classify the source of the sample as either the generated sample or the real sample (Xreal). Therefore, the discriminator is trained to maximize $L_C + L_S$. For generators, the goal is that the generated samples can be recognized by the discriminator as real samples and can be efficiently classified. Therefore, the generator is trained to maximize $L_C - L_S$.

$$L_S = E\left[\log P\left(S = real \mid X_{real}\right)\right] + E[\log P(S = fake \mid X_{fake})], \quad (2)$$

$$L_C = E\left[\log P\left(C = c \mid X_{real}\right)\right] + E[\log P(C = c \mid X_{fake})]. \quad (3)$$

In this research, the generator is implemented using the Transformer-based template generative model. The Transformer architecture is utilized to compute attention in three ways, namely the encoder's self-attention, the decoder's self-attention, and the attention between the encoder and the decoder [37]. The attention function maps a query and a set of key-value pairs to an output. It calculates the correlation between each key and the query and assigns weights to the corresponding values based on the correlation, indicating their importance. The weighted sum of the values gives the final attention. Since one attention alone cannot capture the relationship between input words in different spaces, the Transformer model employs multi-head attention to address this issue, as illustrated in Figure 6. Multi-head attention comprises multiple scaled dot-product attentions, where query, key, and value are linearly transformed into different subspaces h times.

The left side of Figure 7 is the encoder. The encoder is composed of a stack of N = 6 identical layers in the Transformer [37]. The sum of word embeddings and positional embeddings are the input of the encoder. Each layer of the encoder has two sub-layers: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network. A residual connection is

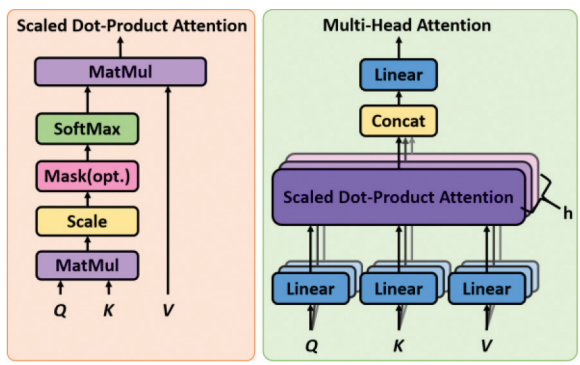

Figure 6: The System Diagram of Scaled Dot-product Attention and Multi-head Attention.

For example, the sentence “最近常常打噴嚏、流鼻水，超不舒服的 (I often sneeze and have runny nose recently, it’s very uncomfortable)” has two slots “打噴嚏 (sneeze)” and “流鼻水 (runny nose)”. After all the user sentences and the corresponding response sentences have been processed, the response templates are obtained and used to train the response template generative model.

Figure 7: The Block Diagram of Transformer-based Template Generative Model.

around each of the two sub-layers, followed by normalization. The right side of Figure 7 is the decoder. The decoder is composed of a stack of N = 6 identical layers in the Transformer. Each layer of the decoder has three sub-layers: a multi-head self-attention mechanism, a position-wise fully connected feed-forward network and a third sub-layer which performs multi-head attention over the output of the encoder stack. The mechanisms of self-attention, residual connection and normalization are the same as the encoder. We modify the self-attention sub-layer in the decoder stack to prevent positions from attending to subsequent positions. The masking technique, along with the shift of the output embeddings by one position, guarantees that the predictions for a certain position i only rely on the known outputs at positions smaller than i.

In this study, the input of the Transformer-based template generative model is the user’s template sentences and the output of the Transformer-based template generative model is the system’s template response. In order to consider the information of user’s personal medical history for detecting user’s emotion and dialogue act, we use one-hot encoding method to embed them into 17-dimensional personal medical history (per_info), 4-dimensional emotion (emo) and 4-dimensional dialogue act vectors (act), and concatenate them into a 25-dimensional vector as the condition vector Con. The architecture of the Transformer-based template generative model is shown in Figure 7, and the formulas are shown in (4)–(8).

$$Con = Concat(emo, per_{info}, act), \tag{4}$$

$$K' = Concat(K, Con), \tag{5}$$

$$V' = Concat(V, Con), \tag{6}$$

$$head_i = Attention(QW_i^Q K' W_i^{K'} V' W_i^{V'}), \tag{7}$$

$$MultiHead(QK'V') = Concat(head_1 \dots head_h)W^O, \tag{8}$$

where Q is query vector, K is key vector, d is value vector, $W_i^Q \in R^{d_{model} \times d_k}$, $W_i^{K'} \in R^{(d_{model}+d_{con}) \times d_k}$, $W_i^{V'} \in R^{(d_{model}+d_{con}) \times d_v}$ and $W_i^O \in R^{hd_v \times d_{model}}$.

In this study, we use the Chinese pre-trained BERT-MLP-based model as the discriminator. It uses Transformer-based encoder to capture the long-distance dependency between texts and deep bidirectional features of texts.

The output layer of the model classifies the authenticity and emotion of the system response template. The adversarial training in this study involves using the Chinese pre-trained BERT-MLP-based model as the discriminator and the conditional Transformer for template generation as the generator. The two networks are trained in alternating turns, with one network being frozen while the other is trained.

Figure 8 shows the training process of the generator and discriminator. In the training process of the generator, the inputs of the generator are the conditions (user emotion, personal medical history and the expected system dialogue act) and the user's template. Next, we extract the output vector of the generator as the vector V_{fake} $\mathrm{WV}_{\mathrm{fake}}$ of the generated response template. Simultaneously, we utilize a pre-trained Word2Vec-based embedding method to transform the system template that corresponds to the user template into the vector V_{real}. We feed the vector V_{fake} and vector V_{real} into the discriminator to obtain the loss of authenticity L_S and the loss of emotion classification L_C. The loss of authenticity L_S and the loss of emotion classification L_C are used to update the weights of the generator, as shown in (9) and (10). In order to make the generated response template similar to the system response template of the collected corpus, the output vector of the generator is linearly converted into word probability, as shown in (11). The word probability and the one-hot encoding of the target word are used to calculate the loss L_{tgt} to update the generator's weight, as shown in (12), so that the system response template is expected to be similar to the collected corpus and can appropriately reflect the user's emotion.

$$L_S = E\left[\log D(V_{real})\right] + E[\log(1 - D(V_{fake}))], \tag{9}$$

$$L_C = E\left[\log P(C = c \mid V_{real})\right] + E[\log P(C = c \mid V_{fake})], \tag{10}$$

$$\hat{Z} = softmax(WV_{fake}), \tag{11}$$

$$L_{tgt} = -\frac{1}{N}\sum_{n=1}^{N}[Z_n \log\left(\hat{Z}_n\right) + (1 - Z_n)\log(1 - \hat{Z}_n)], \tag{12}$$

where V_{fake} is the output of the generator, W is the weight matrix, $\hat{Z}$ is the vector of linear transformation result, Z_n is the vector of one-hot encoding of the target sentence, and N is the number of words in the target sentence. Once the response template is generated, the next step is to fill the words into it using the slot-value table.

During the training process of the discriminator, the system template that corresponds to the user template is transformed into a vector V_{real} through the learned embedding, serving as a real sample. On the other hand, the output vector generated by the generator is considered as a fake sample. They are sent to the discriminator to obtain the loss L_S and the loss L_C to update the weight of the discriminator. It is expected to improve the ability of the

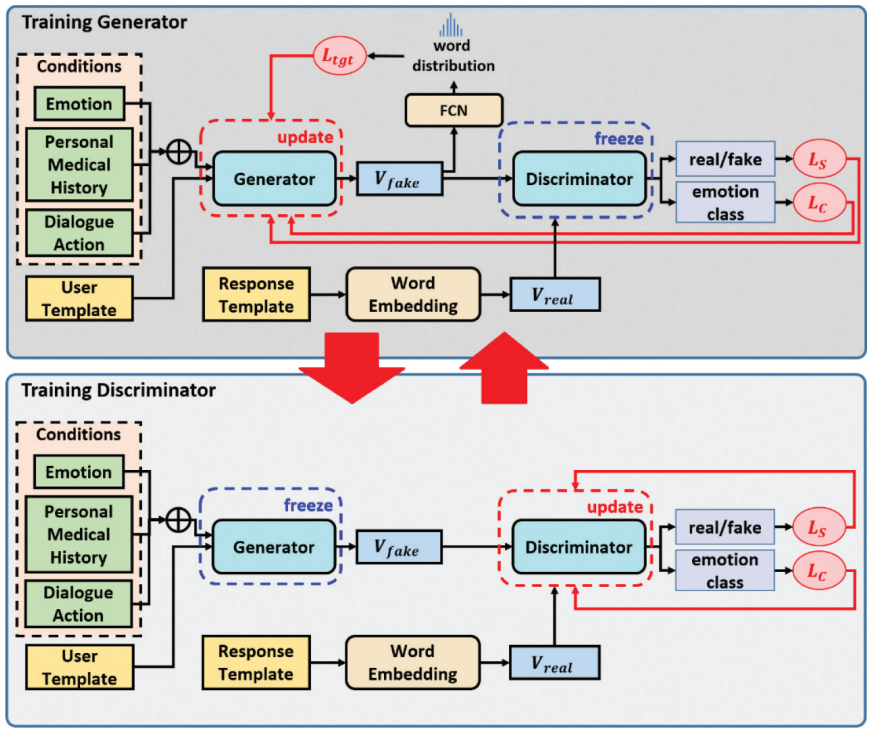

Figure 8: The Block Diagram of Conditional Adversarial Training.

discriminator to distinguish between authenticity and emotional classification to enable the generator to generate the conditional response template. Once the template is generated, the next step is to fill in the slots with words based on the slot-value table. The training process then alternates between training the Transformer-based generator and the BERT-MLP-based discriminator until convergence.

5 Experimental Results and Discussion

In order to create an empathetic dialogue system, this study utilizes three modules: Empathy Analysis, Dialogue Management, and Template Generation. During the training phase, these modules are trained in sequence, and the experiments in this section assess the effectiveness of each module in achieving its intended goals. To evaluate the overall system performance of the proposed methods, we used a five-fold cross-validation.

5.1 Evaluation on Emotion Recognition and Dialogue Management

The BERT-MLP-based, BERT-SVM-based, and Word2VecSVM-based classifiers in this study were trained using 1740 user dialogue sentences and their corresponding emotion labels. To assess the performance of our proposed methods, we utilized a five-fold cross-validation approach. Specifically, for the emotion recognition and dialogue management experiment, we split the corpus into five equal folds based on the categories of emotion and dialogue act. We used four of these parts for training, while the remaining part was employed for testing. In Word2Vec-SVM-based, BERT-SVM-based and BERT-MLP-based classifier, the BERT model and Word2Vec model were used to convert the sentence into embedding representation and the SVM and MLP model were used for emotion recognition. For the BERT-MLP-based model, the following parameters were configured with a batch size of 32, a learning rate of 3e-5, 18 epochs, with the Bert Pre-Trained Model being "Bert-Base-UNCASED", and a 12-layer MLP hidden layer. Regarding the BERT-SVM-based model, the parameters included a batch size of 32, a learning rate of 3e-5, 18 epochs, with the Bert Pre-Trained Model being "Bert-Base-UNCASED," and an SVM kernel of rbf. Lastly, for the Word2Vec-SVM-based model, the parameters comprised a batch size of 32, a learning rate of 3e-5, 18 epochs, a word embedding dimension of 300, and an SVM kernel of rbf.

The performance of different models was evaluated by using five-fold cross-validation method. Table 4 shows that the accuracies of the BERT-MLP-based, the BERT-SVM-based, and the Word2Vec-SVM-based emotion recognition models achieved 95.5±0.7%, 84.1±2.6% and 65.2±1.6%, respectively. Figure 9 shows the normalized confusion matrix of the BERT-MLP-based emotion recognition results for the input sentences. In addition, the 1740 user dialogue sentences and corresponding dialogue act labels were also used to train the BERT-MLP-based dialogue act decision model, BERT-SVM-based dialogue act decision model and the Word2Vec-SVM-based dialogue act decision model. Table 5 shows that the accuracies of the BERT-MLP-based, the BERT-SVM-based and the Word2Vec-SVM-based dialogue act decision models were 95.1 ± 0.5%, 84.1±2.6% and 64.8±2.3%, respectively. Figure 10 shows the normalized confusion matrix of the BERT-MLP-based dialogue act recognition results. Since the performance of the Bert-MLP-based model was better than that of the Bert-SVM-based and Word2Vec-SVM models, we thus used the Bert-MLP-based model as a discriminator in our proposed method.

Furthermore, we assessed whether the emotion recognition model correctly identified the user's emotion in the model-generated response. We used an independent emotion recognition model based on BERT-MLP, which performed a single sentence emotion classification task on the response sentence. We evaluated three different emotion recognition models: BERT-MLP-based, BERT-SVM-based, and Word2Vec-SVM-based models. Table 6 shows that the

Table 4: Evaluation of Emotion Recognition Using Dialogue Sentences.

Model	Accuracy
BERT-MLP-based	$95.5 \pm 0.7\%$
BERT-SVM-based	$84.1 \pm 2.6\%$
Word2Vec-SVM-based	$65.2 \pm 1.6\%$

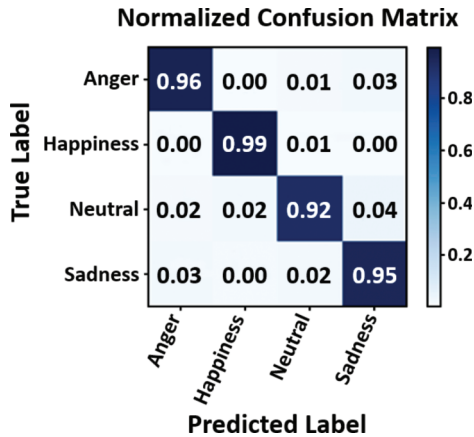

Figure 9: The Confusion Matrix of Emotion Recognition Results for User Sentence.

Table 5: Evaluation of Emotion Recognition Using the Response Generated by the Model.

Model	Accuracy
BERT-MLP-based	$98.1 \pm 0.9\%$
BERT-SVM-based	$91.3 \pm 1.8\%$
Word2Vec-SVM-based	$74.0 \pm 1.5\%$

accuracies of the BERT-MLP-based, the BERT-SVM-based and the Word2Vec-SVM-based emotion reflection by the response sentence were $98.1 \pm 0.9\%$, $91.3 \pm 1.8\%$ and $74.0 \pm 1.5\%$, respectively. Figure 11 shows the normalized confusion matrix of the BERT-MLP-based emotion reflection results for the system response sentences. Because the BERT-MLP-based emotion recognition model had high accuracy for the system response to reflect the user emotion, we used the BERT-MLP-based emotion recognition model to determine the emotion reflected in the generated response.

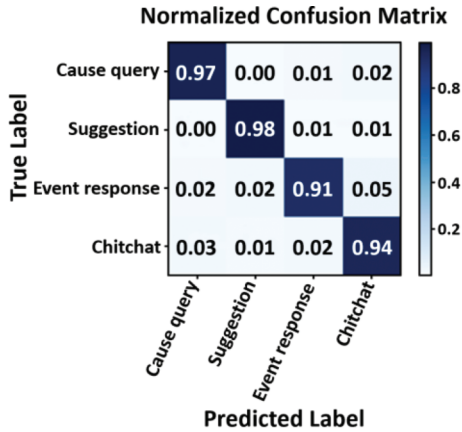

Figure 10: The Confusion Matrix of Dialogue Act Decision Results.

Table 6: Evaluation of Dialogue Act Prediction Using Dialogue Sentences.

Model	Accuracy
BERT-MLP-based	$95.1 \pm 0.5\%$
BERT-SVM-based	$84.1 \pm 2.6\%$
Word2Vec-SVM-based	$64.8 \pm 2.3\%$

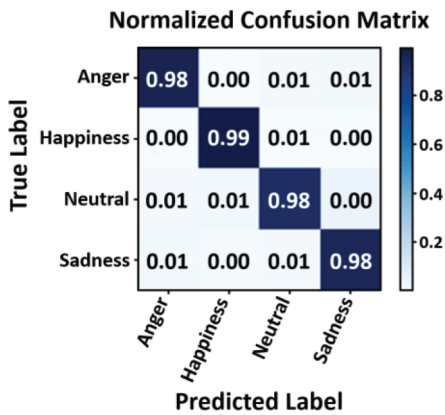

Figure 11: The Confusion Matrix of Emotion Recognition Results for System Sentences.

5.2 Evaluation on Response Generation

Several evaluation criteria were used to evaluate the response generation model. First, for evaluating the quality of the generated response, we used BLEU

(Bilingual Evaluation Understudy) [19, 25] to calculate the similarity between the generated response and the golden response. We used smoothing techniques [4] that worked better for sentence-level evaluation. Carkhuff [2] defines the initial level of empathy in the context of empathy as the capacity to comprehend and articulate the other person's observable or implicit experiences, emotions, and actions. As such, we employed this as our second evaluation criterion to gauge whether the responses generated by our generation model could effectively mirror the user's emotions and events, thereby fostering empathy.

Originally designed to assess the precision of machine-translated text [6], the BLEU (BiLingual Evaluation Understudy) algorithm is employed in this study to evaluate the quality of the generated responses. BLEU was one of the earliest evaluation metrics to claim a high correlation with human quality judgments. It utilizes an n-gram modeling approach to compare the generated response text with the reference text in the ground truth test data. In this study, we used 2-gram (BLUE-2), 3-gram (BLUE-3), and 4-gram (BLUE-4) to gauge the quality of the responses. For comparing the performance of the models considering different situations, conditions consisting of emotion, dialogue act and personal medical history are integrated into the Transformer model for comparison, as shown in Table 7. First of all, we compared the difference in BLEU scores of Transformer model considering emotion, personal medical history and dialogue act, and we found that considering dialogue act can bring a slight benefit to Transformer model. Then we compared the difference in BLEU scores of any two conditions added to the Transformer model, and we found that considering the condition of emotion combined with personal medical history can bring a slight benefit to the Transformer model. Finally, we considered three conditions to join the Transformer model, and we found that considering three conditions at the same time can obtain the best BLEU score for the Transformer model. Therefore, we decided to use Transformer, which considers three conditions at the same time, as the generator architecture of this study.

In terms of response generation evaluation, we compared several models, including the **Transformer**, the **conditional Transformer**, the model of adversarial training based on **conditional Transformer with R/F loss**, the model of adversarial training based on **conditional Transformer with Emo class loss**, and the model of adversarial training based on **conditional Transformer with R/F + Emo class loss**. The R/F loss means the loss of real and fake discrimination and Emo class loss means the loss of emotion classification. The generation model parameters were configured with a batch size of 32, a number of epochs of 20, an embedding dimension of 768, a number of hidden layers of 3, and a hidden size of 3072. For the proposed conditional adversarial training model, in order to ensure that the loss of emotion classification and the loss of real and fake discrimination were included to help model learning, we trained two models: **conditional**

Table 7: BLEU Scores for Comparison Models.

Input Method	Model	BLEU-2	BLEU-3	BLEU-4
Template-based	**Transformer (TF)**	42.4 ± 1.3%	40.0 ± 1.3%	37.8 ± 1.4%
	TF conditioned on Emo	41.8 ± 1.4%	39.5 ± 1.5%	37.2 ± 1.6%
	TF conditioned on DA	**42.8 ± 1.6%**	**40.4 ± 1.6%**	**38.1 ± 1.6%**
	TF conditioned on Pmh	42.3 ± 1.7%	40.1 ± 1.6%	37.8 ± 1.6%
	TF conditioned on Emo + DA	41.6 ± 1.5%	39.3 ± 1.5%	37.0 ± 1.5%
	TF conditioned on Emo + Pmh	**42.7 ± 0.7%**	**40.3 ± 0.7%**	**38.0 ± 0.8%**
	TF conditioned on Pmh + DA	41.3 ± 1.4%	38.9 ± 1.3%	37.5 ± 1.2%
	CT (TF conditioned on Emo + DA + Pmh)	**42.8 ± 1.9%**	**40.5 ± 1.8%**	**38.2 ± 1.8%**

Note: TF conditioned on Emo: Transformer conditioned on emotion
TF conditioned on DA: Transformer conditioned on dialogue act
TF conditioned on Pmh: Transformer conditioned on personal medical history
TF conditioned on Emo + DA: Transformer conditioned on emotion and dialogue act
TF conditioned on Emo + Pmh: Transformer conditioned on emotion and personal medical history
TF conditioned on Pmh + DA: Transformer conditioned on dialogue act and personal medical history
CT: Transformer conditioned on emotion, dialogue act and personal medical history

Transformer with R/F loss and **conditional transformer with Emo class loss**. In addition, we also evaluated the impact of different input methods on the performance of response generation. Table 8 shows the experimental results of the BLEU scores of each model with different input methods.

Compared with the template-based **Transformer** model, **conditional Transformer** slightly improved the generation quality by 0.4% ∼ 0.5%. The experimental results showed that additional consideration of user emotions, personal medical history and dialogue act were helpful for response generation. Compared with the Transformer model, the BLEU score of **conditional Transformer with R/F loss** was improved by 4.2% ∼ 4.8% and the BLEU score of **conditional Transformer with Emo loss** was improved by 4.1% ∼ 4.7%. The experimental results illustrated that these two losses both contributed to the response performance.

Finally, we considered two loss functions at the same time and considered the methods with/without updating the discriminator. Compared with the **Transformer** model and the **conditional Transformer with R/F loss and Emo loss** without updating the discriminator (**CTwithR/F + Emo-UD**), the proposed approach improved the generation quality by 1.8% ∼ 3.6%. Compared with the **Transformer** model and the **conditional Transformer with R/F loss and Emo loss** with updating the discriminator (**CTwithR/F + Emo + UD**), the generation quality was improved by 4.4% ∼ 4.9%. The experimental results showed that considering these two losses at the same time has a better contribution to the generation model and the discriminator can effectively guide the generator for empathetic response generation. The experimental results also showed that the performance of

Table 8: BLEU Scores for Comparison Models.

Input Method	Model	BLEU-2	BLEU-3	BLEU-4
End-to-end-based	**Transformer**	25.3 ± 1.6%	22.9 ± 1.7%	27.0 ± 1.6%
	CT	27.1 ± 1.3%	24.7 ± 1.3%	28.6 ± 1.1%
	CTwithR/F	**31.4 ± 1.2%**	**29.2 ± 1.2%**	**32.9 ± 1.4%**
	CTwithEmo	31.1 ± 1.0%	28.8 ± 1.0%	32.2 ± 1.0%
	CTwithR/F + Emo	31.3 ± 0.6%	29.0 ± 0.6%	32.5 ± 0.6%
Template-based	**Transformer**	42.4 ± 1.3%	40.0 ± 1.3%	37.8 ± 1.4%
	CT	42.8 ± 1.9%	40.5 ± 1.8%	38.2 ± 1.8%
	CTwithR/F	46.6 ± 1.3%	44.6 ± 1.3%	42.6 ± 1.2%
	CTwithEmo	46.5 ± 0.6%	44.5 ± 0.6%	42.5 ± 0.7%
	CT + R/F + Emo-UD	45.6 ± 0.7%	41.8 ± 1.8%	41.1 ± 1.8%
	CT + R/F + Emo + UD	**46.8 ± 1.1%**	**44.8 ± 1.2%**	**42.7 ± 1.2%**

Note: CT: Conditional Transformer
CTwithR/F: Conditional Transformer with R/F loss
CTwithEmo: Conditional Transformer with Emo class loss
CT + R/F + Emo-UD: Conditional Transformer with R/F + Emo class loss and without updating discriminator
CT + R/F + Emo + UD: Conditional Transformer with R/F + Emo class loss and with updating discriminator

the template-based method was better than that of the end-to-end-based method. Finally, the **conditional Transformer** with template-based input method considering both R/F loss, emotion classification loss and updating the discriminator achieved the highest BLEU score, and we thus used the template-based method for the subsequent experiments.

For the proposed conditional adversarial training model, we also compared the BLEU scores of the proposed method with updating the discriminator (**CTwithR/F + Emo + UD**) and the proposed method without updating discriminator (**CTwithR/F + Emo-UD**) in the template generation component. The experimental results are shown in Table 8. In Table 8, using only the pre-trained BERT-MLP-based model without updating the discriminator (**CTwithR/F + Emo-UD**) achieved a slight performance improvement. In contrast, the proposed method with updating the discriminator (**CTwithR/F + Emo + UD**) could achieve a significant improvement through subjective and objective experimental analysis. The results somewhat showed that the revenue came from the adversarial method more, instead of the pre-trained BERT-MLP-based model.

BERTSCORE is an automatic evaluation metric for text generation that computes a similarity score for each token in the candidate sentence and each token in the reference sentence. Zhang et al. used an adversarial paraphrase detection task to show that BERTSCORE is more robust to challenging examples when compared to existing metrics [44]. Since BERTSCORE correlates better with human judgment and provides stronger model selection perfor-

Table 9: Performance Evaluation of Question Sentences and Generative Sentences.

	BERTSCORE
Transformer	−6.05
CT	−5.88
CTwithR/F	−5.88
CTwithEmo	−5.90
CTwithR/F + Emo	**−5.84**

Table 10: Performance Evaluation of Answer Sentences and Generative Sentences.

	BERTSCORE
Transformer	−4.59
CT	−4.33
CTwithR/F	−4.23
CTwithEmo	−4.28
CTwithR/F + Emo	**−4.15**

mance than existing metrics, we used BERTSCORE to evaluate the system performance of the proposed method and traditional systems. As shown in Tables 9 and 10, we compared the generated sentence and question similarity scores and the generated sentence and correct answer similarity scores, respectively. In BERTScore evaluation, our proposed method outperformed **Transformer** and **conditional Transformer** methods, which proved that the sentences generated by our proposed method were similar to the golden response sentences and question sentences.

Table 11 shows the results of each model reflecting the correct rate of overall emotion recognition. Compared with the Transformer model, the reflection rate of the conditional Transformer model in the overall emotions was improved. The experimental result showed that the system considering user emotions, personal medical history, and dialogue act could help improve the model's emotional response rate. The model that considered both losses had higher BLEU scores and emotional response rates than when only a single loss was considered. We believed that this was because the classification of real/fake and the emotions could strengthen the generative model, so that the generative model could generate the responses to reflect the correct emotion close to the real data distribution.

Table 12 shows the ratio of each emotion being correctly reflected in each model. Compared with the Transformer and conditional Transformer, conditional Transformer with adversarial learning, the proposed approach had

Table 11: Overall Emotion Reflection Rate for Comparison Models.

Method	Emotion Refection Rate
Transformer	$83.3 \pm 1.5\%$
CT	$85.0 \pm 1.6\%$
CTiwthR/F	$85.7 \pm 2.2\%$
CTwithEmo	$86.2 \pm 2.0\%$
CT + R/F + Emo + UD	$\mathbf{86.4 \pm 2.1\%}$

Note: CT: Conditional Transformer
CTwithR/F: Conditional Transformer with R/F loss
CTwithEmo: Conditional Transformer with Emo class loss
CT + R/F + Emo + UD: Conditional Transformer with R/F + Emo class loss and with updating discriminator

Table 12: The Emotion Reflection Rate of Each Emotion in Each Model.

Method	Anger	Happiness	Neutral	Sadness
Transformer	$83.2 \pm 6.5\%$	$81.8 \pm 7.3\%$	$80.6 \pm 4.2\%$	$87.5 \pm 4.7\%$
CT	$89.1 \pm 3.4\%$	$81.4 \pm 4.0\%$	$78.1 \pm 4.3\%$	$91.4 \pm 3.7\%$
CTwithR/F	$\mathbf{89.9 \pm 5.2\%}$	$83.7 \pm 1.8\%$	$77.2 \pm 4.4\%$	$92.0 \pm 2.6\%$
CTwithEmo	$89.7 \pm 3.9\%$	$\mathbf{84.6 \pm 3.0\%}$	$79.3 \pm 4.2\%$	$91.3 \pm 3.7\%$
CT + R/F + Emo + UD	$89.1 \pm 6.1\%$	$82.7 \pm 7.8\%$	$\mathbf{81.3 \pm 4.0\%}$	$\mathbf{92.4 \pm 2.5\%}$

Note: CT: Conditional Transformer
CTwithR/F: Conditional Transformer with R/F loss
CTwithEmo: Conditional Transformer with Emo class loss
CT + R/F + Emo + UD: Conditional Transformer with R/F + Emo class loss and with updating discriminator

a small increase in the overall emotional response rate, but the response rate for each emotion was improved significantly.

5.3 Subjective Evaluation on Response Generation

The subjective evaluation focused on three aspects: the relevance between the generated sentence and the user's input, the grammatical correctness of the generated sentence, and whether it demonstrated empathy. The generated sentence was generated by the **Transformer**, the **conditional Transformer**, and the **conditional Transformer with R/F + Emo class loss + updating the discriminator** method. A total of seven master's students aged between 22 and 25 participated in the subjective evaluation. Each participant was asked to rate the appropriateness of 341 generated sentences based on three criteria: relevance to the user sentence, grammatical correctness, and empathy of the generated sentences. The rating scale ranged from 0 to 2, as presented in Tables 13, 14, and 15. A rating of 2 indicated agreement with the relevance, grammatical correctness, and empathy of the generated sentence, while a rating of 0 indicated disagreement. For relevance, a score of 2 indicated that

Table 13: Subjective Evaluation of Relevance.

Method	2	1	0
Transformer	68.96%	13.28%	**17.76%**
CT	71.05%	12.23%	16.72%
CT + R/F + Emo + UD	**77.55%**	**16.72%**	12.74%

Table 14: Subjective Evaluation of Grammatical Correctness.

Method	2	1	0
Transformer	71.01%	10.35%	**18.65%**
CT	70.50%	**11.07%**	18.43%
CT + R/F + Emo + UD	**79.47%**	9.18%	11.35%

Table 15: Subjective Evaluation of Empathy.

Method	**2**	**1**	**0**
Transformer	69.08%	**15.88%**	**15.04%**
CT	71.64%	15.79%	12.57%
CT + R/F + Emo + UD	**75.87%**	12.20%	11.93%

the response sentence was relevant to the user sentence, whereas a score of 0 meant that the response sentence was not related to the user sentence. For grammatical correctness, a score of 2 indicated that the content of the response sentence was grammatically correct, while a score of 0 meant that it was not. In the empathy score, a rating of 2 indicated that the content of the response sentence reflected empathy, correctly represented the user's emotions, and provided an empathetic response. A score of 0 suggested that the content of the response sentence lacked empathy and was unable to provide the user with a sympathetic response.

In subjective evaluation, the proposed model attained the accuracies of 77.55 %, 79.47% and 75.87 % on the score of 2 for relevance, grammatical correctness, and empathy, respectively. From the evaluation results, this study showed that the **conditional Transformer with R/F + Emo class loss + updating discriminator** method could get higher scores than the other methods. In addition, the generated template obtained a more appropriate empathetic response with fewer grammatical errors.

Tables 16, 17 and 18 show the Cohen's KAPA score and the significance of the difference for different generation methods by ANOVA statistical analysis method. Cohen's kappa coefficient is a statistic that is used to measure

Table 16: The Significance Test of Relevance in Different Generation Methods.

Method	Mean	Variance	KAPA	F VALUE	P-VALUE	THRESHOLD
Transformer	1.51	0.61	0.601			
CT	1.54	0.58	0.612	**21.78**	**3.71E-10*****	3.00
CT + R/F + Emo + UD	1.65	0.48	0.641			

Table 17: The Significance Test of Grammatical Correctness in Different Generation Methods.

Method	Mean	Variance	KAPA	F value	P-value	Threshold
Transformer	1.52	0.62	0.632			
CT	1.52	0.62	0.620	**35.64**	**4.44E-16*****	3.00
CT + R/F + Emo + UD	1.68	0.44	0.612			

Table 18: The Significance Test of Empathy in Different Generation Methods.

Method	Mean	Variance	KAPA	F value	P-value	Threshold
Transformer	1.54	0.54	0.625			
CT	1.59	0.49	0.623	**11.87**	**7.13E-6*****	3.00
CT + R/F + Emo + UD	1.64	0.47	0.631			

inter-rater reliability (and also intra-rater reliability) for qualitative items [21]. In Tables 16, 17 and 18, the average Cohen's kappa coefficient shows the strength of agreement are good. The average Cohen's kappa coefficient is the average Cohen's kappa coefficient of the other six participants for the first participant. According to Table 16, the correlation score between the generated sentence and the user sentence for the **conditional Transformer with R/F + Emo class loss + updating the discriminator** method was higher than the **Transformer** and the **conditional Transformer**. It showed that our proposed method in this study obtained better correlation between the generated sentence and the user sentence. According to Table 17, the grammatical correctness score for the **conditional Transformer with R/F + Emo class loss + updating the discriminator** method was higher than the **Transformer** and the **conditional Transformer**. It showed that our proposed method in this study obtained better grammatical correctness. From Table 18, the empathy score for the **conditional Transformer with R/F + Emo class loss + updating the discriminator** method was higher than the **Transformer** and the **conditional Transformer**.

To assess the effectiveness of the proposed method in consultations, we conducted 20 brief conversations with five university students, each lasting no more than 3 minutes. Participants were between the ages of 18 and 22. Subsequently, we asked the participants to rate their perceived level of empathy

on a scale of 1 to 5. A score of 1 indicated that the participant felt no empathy during the conversation, while a score of 5 indicated a high degree of empathy. The experimental results showed that the mean score of empathy for the sentences generated from the **Transformer** was 2.8 (the other participants' KAPA scores for the first participant were 0.63, 0.679, 0.614 and 0.628), and the mean score of empathy for the **conditional Transformer** method was 2.68 (the other participants' KAPA scores for the first participant were 2.68). Participants' KAPA scores were 0.643, 0.654, 0.634 and 0.6), the mean score empathy for the **conditional Transformer with R/F** method was 3.03 (other participants' KAPA scores for the first participant were 0.649, 0.658, 0.645 and 0.643), and the mean score of empathy for the **conditional Transformer with Emo class loss** method was 3.62 (the other participants' KAPA scores for the first participant were 0.66, 0.603, 0.66 and 0.65), and the mean score of empathy for the proposed method was 3.8 (the other participants' KAPA scores for the first participant were 0.624, 0.618, 0.624 and 0.643). The experimental results show that the proposed method (**conditional Transformer with R/F + Emo class loss + updating the discriminator**) achieved the best empathy score of 3.8 (average KAPA score was 0.627), which was better than the other methods.

To evaluate the textual diversity of generated responses, the metric "Distinct" calculates the number of distinct n-grams. Meanwhile, "Lexical diversity" measures the lexical richness by comparing the ratio of unique tokens to the total number of tokens. This study compares the distinct and diversity scores of generated sentences and correct answers. The transformer method has a diversity score of 0.79 and a DIST-2 score of 0.05. The conditional transformer method has a diversity score of 0.72 and a DIST-2 score of 0.03. The conditional transformer with R/F loss method has a diversity score of 0.71 and a DIST-2 score of 0.05. Lastly, the conditional transformer with EMO loss method has a diversity score of 0.71 and a DIST-2 score of 0.05.

The above experiments confirmed that the generative model with conditional adversarial learning could generate empathic sentences. In addition, the results generated by the discriminator-guided generator could reflect the user's emotions. Table 19 shows the generated examples of the generative model during the test phase.

6 Conclusion and Future Work

This study proposes a conditional adversarial learning approach for empathetic dialogue response generation. The empathetic dialogue system begins by obtaining the user's personal medical history and detecting their emotions to enhance the conversation experience with empathic responses. To achieve this,

Table 19: Generated Examples of the Generative Model.

Speaker	Text	Personal Information
User	大便都大不出來，真叫人受不了！(I can't defecate smoothly, and I can't take it anymore.)	便秘 (constipation)
Golden Response	我能理解大便都大不出來讓你不開心，便秘建議你有便意就要及時如廁哦！(I can understand that having difficultly during defecation makes you unhappy. Because you have constipation, I suggest that you must go to the toilet regularly!)	
Generated	我能理解大便都大不出來讓你生氣，便秘建議你營養均衡不偏食哦！(I can understand that having difficultly during defecation makes you angry. Because you have constipation, I suggest consuming a balanced diet!)	
User	最近一直鼻塞，真難過！(I am having nasal congestion recently. It feels bad.)	過敏性鼻炎 (allergic rhinitis)
Golden Response	鼻塞感覺真難過，你有接近什麼過敏原嗎？(Having nasal congestion feels bad. Have you come in contact with any allergens?)	
Generated	我懂你因為鼻塞難過的心情，你有過敏性鼻炎，要養成健康的生活習慣對你有幫助。 (I understand that you feel sad because of nasal congestion. You have allergic rhinitis, and it is helpful to develop a healthy lifestyle.)	
User	好愉悅哦！我現在不會失眠了，我要去騎腳踏車囉！(I am so happy! I don't have insomnia now. I'm going to ride a bike)	高血壓 (hypertension)
Golden Response	我知道你很愉悅，雖然你現在不會失眠，不過還是定期到醫院做健康檢查哦！ (I know that you are happy, although you got rid of insomnia, you still need to go to the hospital for health checkups regularly!)	
Generated	我知道你很愉悅，雖然你現在不會失眠，不過還是適度的規律運動以維持理想體重哦！(I know that you are happy, although you got rid of insomnia, you still need regular moderate exercise to maintain the ideal weight!)	

we first use the maximum matching method to extract the user's slot-value table from their sentence input. Then, BERT-MLP-based models are utilized to identify the emotions in the user's sentences and determine the appropriate system dialogue acts. For event reflection in empathy, we generate templates and fill in the words using the slot-value table, which is created by replacing the slot values of sentences in the database with the corresponding slot label. We train the template generation model using the maximum matching method on the template samples. The generated templates are used to reflect user

events in the system response. For emotion reflection, we employ conditional adversarial learning to guide the system response in emotional reflection. The experimental results showed that the proposed method of applying conditional adversarial learning achieved the best BLEU score (41.3%), the best BERTSCORE (−5.84 for evaluation on question sentences and generative sentences; −4.15 for evaluation on answer sentences and generative sentences) and emotion reflection rate (86.4%), which outperformed the Transformer-based and the conditional Transformer-based methods. This study also conducted subjective evaluations, and the proposed model achieved 77.55%, 79.47%, and 75.87% accuracy in the scores of relevance, grammatical correctness, and empathy, respectively. In the significance test and Cohen's KAPA score of relevance, grammatical correctness and empathy, the proposed method was better than the Transformer- and conditional Transformer-based methods. In addition, in the consulting performance evaluation, the experimental results showed that the proposed method (CT + R/F + Emo + UD) achieved the best empathy score of 3.8 (average KAPA score was 0.627), which was better than the other methods.

This study suggests several avenues for further research. Firstly, we propose incorporating new topics such as food and family to enrich the conversational content. Secondly, in future studies, we aim to consider additional personal medical histories, such as family medical history, to provide users with more personalized and empathetic responses. Furthermore, our future work will emphasize enhancing our proposed approach through the exploration of more advanced models and the integration of knowledge graph models. Lastly, we plan to conduct user evaluations by facilitating interaction between individuals and the complete system, thereby increasing its practical applicability.

Financial Support

This work was supported in part by the Ministry of Science and Technology, Taiwan, under Contract No. MOST 111-2221-E-006-150-MY3.

Biographies

Ming-Hsiang Su received the B.S. degree in computer science and information engineering from the Tunghai University, Taichung, Taiwan, in 2001, the M.S. degree in management information systems from the National Pingtung University of Science and Technology, Pingtung, Taiwan, in 2003, and the Ph.D.

degree in computer science and information engineering from National Chung Cheng University (NCCU), Tainan, Taiwan, in 2013. He was a Postdoctoral Fellow with the Department of Computer Science and Information Engineering, NCKU, from 2013 to 2020. Since 2022, he has been with the Department of Data Science, Soochow University (SCU), Taiwan. His research interests include e-learning, artificial intelligence, machine learning, multimedia signal processing, and personality detection.

Chung-Hsien Wu received his B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981. He went on to earn his M.S. and Ph.D. degrees in electrical engineering from National Cheng Kung University (NCKU), Tainan, Taiwan, in 1987 and 1991, respectively. Since 1991, he has been a member of the Department of Computer Science and Information Engineering at NCKU, where he was appointed as Chair Professor in 2017. From 2009 to 2015, he served as the Deputy Dean of the College of Electrical Engineering and Computer Science at NCKU. In the summer of 2003, he worked as a Visiting Scientist at the Computer Science and Artificial Intelligence Laboratory of the Massachusetts Institute of Technology (MIT), Cambridge, MA, USA. He has served as an Associate Editor of several journals, including IEEE Transactions on Audio, Speech and Language Processing (2010–2014), IEEE Transactions on Affective Computing (2010–2014), ACM Transactions on Asian and Low-Resource Language Information Processing, and APSIPA Transactions on Signal and Information Processing (2014 $\sim$ 2020). He was also a member of the APSIPA BoG from 2019 to 2021. He received the 2018 APSIPA Sadaoki Furui Prize Paper Award and the Outstanding Research Award from the Ministry of Science and Technology in Taiwan in 2010 and 2016. His research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing.

Chia-Yu Liao received the M.S. degree in computer science and information engineering from National Cheng Kung University, Tainan, Taiwan, in 2019. Her research interests include multimedia signal processing and natural language processing.

References

[1] A. Agrawal and A. An, "Unsupervised Emotion Detection from Text Using Semantic and Syntactic Relations," in *Proceedings of IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, 2012, 346–53.

[2] R. R. Carkhuff, "Helping and Human Relations: A Primer for Lay and Professional Helpers: I. Selection and Training," *Holt. Rinehart and Winston*, 1969.

[3] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "SemEval-2019 Task 3: EmoContext Contextual Emotion Detection in Text," in *Proceedings of the 13th International Workshop on Semantic Evaluation*, 2019, 39–48.

[4] B. Chen and C. Cherry, "A Systematic Comparison of Smoothing Techniques for Sentence-Level BLEU," in *Proceedings of the 9th Workshop on Statistical Machine Translation*, 2014, 362–7.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," 2018, *arXiv preprint* [Online], arXiv: 1810.04805.

[6] S. Dutta and D. Klakow, "Evaluating a neural multi-turn chatbot using BLEU score," *Univ. Saarl*, 10, 2019, 1–12.

[7] P. Fung, D. Bertero, Y. Wan, A. Dey, R. H. Y. Chan, F. B. Siddique, Y. Yang, C.-S. Wu, and R. Lin, "Towards Empathetic Human-robot Interactions," in *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, 2016, 173–93.

[8] P. Fung, D. Bertero, P. Xu, J. H. Park, C. S. Wu, and A. Madotto, "Empathetic Dialog Systems," in *Proceedings of International Conference on Language Resources and Evaluation*, 2018.

[9] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Communications of the ACM*, 63(11), 2020, 139–44.

[10] D. Guo, K. Wang, J. Yang, K. Zhang, X. Peng, and Y. Qiao, "Exploring Regularizations with Face, Body and Image Cues for Group Cohesion Prediction," in *Extraction of International Conference on Multimodal Interaction*, 2019, 557–61.

[11] M. Hasan, E. Rundensteiner, and E. Agu, "Emotex: Detecting Emotions in Twitter messages," in *Proceedings of ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference*, 2014.

[12] J.-H. Hsu, J. Chang, M.-H. Kuo, and C.-H. Wu, "Empathetic Response Generation based on Plug-and-Play Mechanism with Empathy Perturbation," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 31, 2023, 2032–42.

[13] J.-H. Hsu, M.-H. Su, C.-H. Wu, and Y.-H. Chen, "Speech Emotion Recognition Considering Nonverbal Vocalization in Affective Conversations," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 29, 2021, 1675–86.

[14] L. H. Lee and Y. Lu, "Multiple embeddings enhanced multi-graph neural networks for Chinese healthcare named entity recognition," *IEEE Journal of Biomedical and Health Informatics*, 25(7), 2021, 2801–10.

[15] Q. Li, H. Chen, Z. Ren, P. Ren, Z. Tu, and Z. Chen, "EmpDG: Multi-resolution Interactive Empathetic Dialogue Generation," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, 4454–66.

[16] Q. Li, P. Li, Z. Ren, P. Ren, and Z. Chen, "Knowledge Bridging for Empathetic Dialogue Generation," in *Proceedings of the 36th AAAI conference on Artificial Intelligence*, 2022, 10993–1001.

[17] Z. Lin, A. Madotto, J. Shin, P. Xu, and P. Fung, "MoEL: Mixture of Empathetic Listeners," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 2019, 121–32.

[18] Z. Lin, P. Xu, G. I. Winata, F. B. Siddique, Z. Liu, J. Shin, and P. Fung, "CAiRE: An End-to-End Empathetic Chatbot," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 13622–3.

[19] D. Liu, J. Fu, Q. Qu, and J. Lv, "BFGAN: Backward and Forward Generative Adversarial Networks for Lexically Constrained Sentence Generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(12), 2019, 2350–61.

[20] L. Mary, "Significance of Prosody for Speaker, Language, Emotion, and Speech Recognition," in *Extraction of Prosody for Automatic Speaker, Language, Emotion and Speech Recognition*, 2019, 1–22.

[21] M. L. McHugh, "Interrater Reliability: The Kappa Statistic," *Biochemia Medica*, 22(3), 2012, 276–82.

[22] S. Mohammad, "# Emotional Tweets," in *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, 2012, 246–55.

[23] E. Mower, M. J. Matarić, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," *IEEE Transactions on Audio, Speech, and Language Processing*, 19(5), 2010, 1057–70.

[24] A. Odena, C. Olah, and J. Shlens, "Conditional Image Synthesis with Auxiliary Classifier GANs," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2017, 2642–51.

[25] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proceedings of the 40th annual meeting on association for computational linguistics, Philadelphia*, 2002, 311–8.

[26] T. T. D. Pham, S. Kim, Y. Lu, S. W. Jung, and C. S. Won, "Facial Action Units-Based Image Retrieval for Facial Expression Recognition," *IEEE Access*, 7, 2019, 5200–7.

[27] M. Purver and S. Battersby, "Experimenting with Distant Supervision for Emotion Classification," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, 2012, 482–91.

[28] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, "I Know the Feeling: Learning to Converse with Empathy," 2019, *arXiv preprint* [Online], Available: arXiv: 1811.00207.

[29] H. Rashkin, E. M. Smith, M. Li, and Y. L. Boureau, "Towards Empathetic Open-domain Conversation Models: A New Benchmark and Dataset," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, 5370–81.

[30] K. Roberts, M. A. Roach, J. Johnson, J. Guthrie, and S. Harabagiu, "EmpaTweet: Annotating and Detecting Emotions on Twitter," in *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, 2012, 3806–13.

[31] J. Shin, P. Xu, A. Madotto, and P. Fung, "Generating Empathetic Responses by Looking Ahead the User's Sentiment," in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, 7989–93.

[32] F. B. Siddique, O. Kampman, Y. Yang, A. Dey, and P. Fung, "Zara Returns: Improved Personality Induction and Adaptation by an Empathetic Virtual Agent," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics-System Demonstrations*, 2017, 121–6.

[33] M.-H. Su, C.-H. Wu, and Y. Chang, "Follow-Up Question Generation using Neural Tensor Network-based Domain Ontology Population in an Interview Coaching System," in *Proceedings of INTERSPEECH*, 2019, 4185–9.

[34] M.-H. Su, C.-H. Wu, and L.-Y. Chen, "Attention-based Response Generation Using Parallel Double Q-Learning for Dialog Policy Decision in a Conversational System," *IEEE/ACM Trans. Audio, Speech, and Language Processing*, 28(1), 2020, 131–42.

[35] M.-H. Su, C.-H. Wu, K.-Y. Huang, Q.-B. Hong, and H.-H. Huang, "Follow-up Question Generation Using Pattern-based Seq2seq with a Small Corpus for Interview Coaching," in *Proceedings of INTERSPEECH*, 2018, 1006–10.

[36] S. Tahara, K. Ikeda, and K. Hoashi, "Empathic Dialogue System Based on Emotions Extracted from Tweets," in *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 2019, 52–6.

[37] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, 2017, 5998–6008.

[38] Y.-H. Wang, J.-H. Hsu, C.-H. Wu, and T.-H. Yang, "Transformer-based Empathetic Response Generation Using Dialogue Situation and Advanced-Level Definition of Empathy," in *Proceedings of ISCSLP2021*, Hong Kong, January 2021.

[39] N. Weinstein, H. S. Hodgins, and R. M. Ryan, "Autonomy and Control in Dyads: Effects on Interaction Quality and Joint Creative Performance," *Personality and Social Psychology Bulletin*, 36(12), 2010, 1603–17.

[40] G. I. Winata, O. Kampman, Y. Yang, A. Dey, and P. Fung, "Nora the Empathetic Psychologist," in *Proceedings of INTERSPEECH*, 2017, 3437–8.

[41] Y. Yang, X. Ma, and P. Fung, "Perceived Emotional Intelligence in Virtual Agents," in *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, 2017, 2255–62.

[42] R. Zandie and M. H. Mahoor, "EmpTransfo: A Multi-Head Transformer Architecture for Creating Empathetic Dialog Systems," in *Proceedings of the 33rd International Flairs Conference*, 2020, 276–81.

[43] H. Zhang, V. Sindagi, and V. M. Patel, "Image De-Raining Using a Conditional Generative Adversarial Network," *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11), 2019, 3943–56.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERT-Score: Evaluating Text Generation with BERT," in *Proceedings of the 8th International Conference on Learning Representations, Virtual Conference*, 2020.