APSIPA Transactions on Signal and Information Processing, 2024, 13, e100

Original Paper

Informative and Long-Term Response Generation using Multiple Suggestions and User Persona Retrieval in a Dialogue System

Jia-Hao Hsu[1], Tsai-Yi Chen[2] and Chung-Hsien Wu[1,2*]

[1] *Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan*
[2] *Graduate Program of Artificial Intelligence, National Cheng Kung University, Tainan, Taiwan*

ABSTRACT

Enhancing user satisfaction in dialogue systems relies on their ability to understand users and generate responses that meet their expectations. This study proposes a dialogue system that incorporates the Multi-Suggestions Transformer (MST) to generate informative and long-term responses. The MST combines empathy suggestions, system persona suggestions, and knowledge suggestions to produce comprehensive and informative responses. Additionally, the system employs a persona detection model and a persona extraction model to extract the user persona from current sentences and retrieve the most suitable user persona from the dialogue history. This facilitates long-term conversations by enabling the system to remember and respond to sentences relevant to the user persona. The proposed MST-based dialogue system outperforms the baseline in terms of informativeness, as evidenced by higher scores in BLEU, BERT-score, Distinct-n, and Perplexity on the Blended Skill Talk and Multi Session Chat datasets. Furthermore, two novel evaluation metrics, PerP and PerB, introduced

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

Received 26 June 2023; Revised 01 August 2023
ISSN 2048-7703; DOI 10.1561/116.00000145

in this study demonstrate the system's effective utilization of the user persona for achieving long-term dialogue. Human subjective evaluation indicates that our model consistently outperforms the baseline, achieving superior scores of 68%, 56%, 52%, and 64% in the four subjective metrics.

Keywords: Dialogue system, user persona, multi-suggestions transformer, informative and long-term responses.

1 Introduction

1.1 Motivations

In everyday conversations, humans naturally gauge the character of others through their words and provide informative responses. When familiar individuals engage in conversation, they typically discuss topics of mutual interest [7, 37]. Once preferences and backgrounds are understood [30], people consider how to respond in order to meet the expectations of the other party, as depicted in Figure 1. For instance, we respond to individuals under stress with comforting and empathetic content, and provide information to those facing difficulties, rather than offering perfunctory or vague responses such as "I am sorry to hear that" and "I see" [38, 45]. Moreover, our responses are influenced by past utterances. Such interactions also foster interpersonal attraction [28]. Building upon these principles, this study introduces a long-term dialogue system capable of actively remembering user preferences and personalities during conversations, generating informative responses that align with the user's interests [9, 24]. Despite years of research on dialogue systems, many existing studies continue to grapple with the issue of generating poorly informed and irrelevant responses in open-domain settings [21, 41]. Addressing these challenges is the primary motivation behind this study.

1.2 Background

Common chit-chat dialogue systems can be categorized into four types: empathetic dialogue [31, 39, 42], persona-aware dialogue [22, 34, 40], knowledge-based dialogue [16, 19], and open-domain dialogue systems [2, 46]. Empathetic dialogue systems primarily prioritize emotional performance during interactions with individuals [18, 20, 46]. For example, EmpTransfo utilized emotions, actions, and topics from the EmpatheticDialogues dataset to provide additional information to the model, leading to improved results [42]. Persona dialogue systems mainly focus on aligning the system's responses with a given persona,

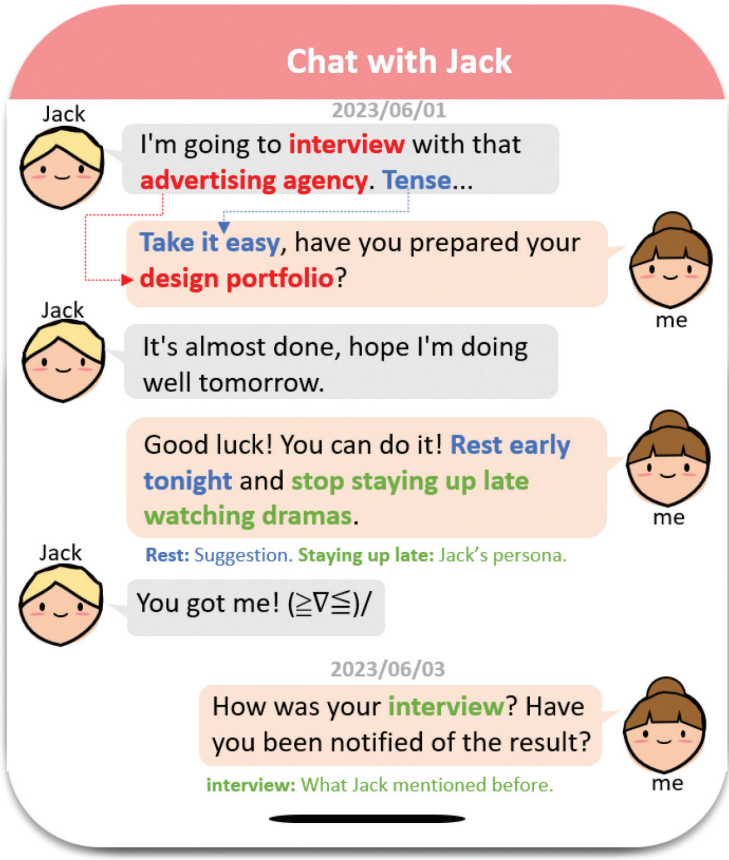

Figure 1: The example of a daily conversation.

which contains personal details [3]. However, in the majority of studies, personas are directly used from the dataset itself. Only a few studies focus on generating or predicting personas for unfamiliar users. Lu *et al.* [23] proposed a method to generate personas from user sentences, and their approach led to an enhancement in succeeding dialogue response generation. On the other hand, Cao *et al.* [6] suggested a different approach of editing existing personas to expand user personas. This solution addressed the problem of limited persona-based dialogue data available for experimentation. Song *et al.* [34] implemented a three-stage framework: (1) generating a response, (2) verifying if the response incorporates the system's persona, and (3) subsequently revising the response if necessary. Knowledge-based dialogue systems concentrate on enabling the system to utilize external knowledge. Lian *et al.* [19] employed

user input sentences, prior knowledge distributions, response tokens, and posterior knowledge distributions to facilitate knowledge selection. Open-domain dialogue systems aim to enable the system to engage in general-purpose chat across various topics. Blender [29], for instance, employed a Transformer model and demonstrated that setting a minimum length in beam search can effectively enhance performance. These studies highlight the value of providing the system with additional reference information to generate satisfactory responses.

However, open-domain dialogues lack these reference sources. Blended Skill Talk (BST) [33] serves as an open-domain dialogue corpus that combines the three characteristics of EmpatheticDialogues (ED) [27], PersonaChat (PC) [43] (convAI2 [11]), and WizardofWikipedia (WoW) [12], enabling seamless switching between the three dialogue skills. Leveraging the perfect performance of the Transformer model in the field of deep learning [8, 13, 14, 35, 39], we modified the structure of the Transformer to obtain a greater number of suggested responses before generating the final response, thereby increasing the amount of information. In the encoding stage, the Transformer's encoders are trained to extract additional features from the suggested sentence that is relevant to the user input sentence. By acquiring multiple suggested sentence features prior to decoding, the Transformer can generate responses with richer information.

Recently, Xu *et al.* noted that while many existing dialogue datasets exhibit good quality, the length of the dialogues is often too short, which poses a challenge for dialogue systems to effectively remember the dialogue history [41]. To address this issue, they introduced the Multi Session Chat (MSC) dataset, which consists of long-distance dialogues. Notably, the system responses within the MSC dataset include information about the users' personas. In our approach, we train the models to extract the user persona from the user input sentence. This enables the system to remember the extracted user personas. When generating a response sentence, the system retrieves appropriate and relevant personas from these characters and takes them into consideration. By incorporating user personas into the dialogue generation process, our system enhances its ability to generate contextually appropriate and personalized responses.

1.3 Contributions

The objective of this study is to enhance the generation of informative responses and the understanding and generation of user persona-related responses in dialogues. To achieve this goal, we adopted the data collection methods employed in BST and MSC to develop a dialogue system based on a novel approach called the Multi-Suggestion Transformer. The main contributions of this study are outlined below.

1.3.1 Multiple Suggestions for Informative Response

The proposed system extends the Transformer model by incorporating multiple suggestions. These suggestions are encoded based on the dialogue skills from ED, PC, and WoW, which are relevant to the current user sentence. By integrating these suggestions, the system generates informative responses. To evaluate the performance of our proposed system, we employed metrics such as BLEU, BERT-score, Distinct-N, and Perplexity, which indicate the system's ability to generate responses that are more informative.

1.3.2 Retrieval of User Persona for Long-Term Dialogue

The user persona is continuously extracted and stored in the dialog history, referred to as the user persona list. This enables the system to retrieve the appropriate persona from the history when generating responses. In the experimental results, we utilized metrics such as Persona percentage and Persona BLEU to demonstrate that our proposed system generates responses that contain more user persona-related phrases.

In summary, this study introduces a Multi-Suggestion Transformer-based dialogue system that generates informative responses and utilizes user persona for long-term dialogues. The experimental results indicate the effectiveness of our approach in terms of response informativeness and incorporation of user persona.

2 Related Works

As far as we know, ChatGPT [4], Bard [26], LaMDA [36] and BlenderBOT are all natural language processing models developed by OpenAI and Google. ChatGPT stands out as one of the largest language models to date, with 1.75 trillion parameters, making it very powerful in generating language. However, due to its extensive training, ChatGPT may sometimes generate false or untrustworthy information.

In comparison, LaMDA and Bard are known for being more fluent and efficient than traditional NLP models. However, since they might still be in the experimental stage, there could be some limitations and issues to consider. Notably, Bard appears to be more suitable for the function of a personal assistant.

On the other hand, BlenderBOT is specifically optimized for multi-turn dialogue and excels in handling complex dialogue situations. Given its model size and flexibility in fine-tuning, particularly its excellent performance in multi-turn dialogue and open-domain situations, this study has chosen BlenderBOT as the baseline model.

The BlenderBot model has emerged as a state-of-the-art system for open-domain dialogue [1], showcasing its prowess across various benchmarks. BlenderBot 1.0 is a deep learning model designed to engage in conversations and respond like a conversational agent [29]. Building upon this foundation, BlenderBot 2.0 utilizes a standard architecture that incorporates seq2seq models and Transformers to generate responses. The Transformer-based architecture serves as the basis for creating long-term memory chatbots, with three different sizes available (90M, 2.7B, and 9.4B). These models can access and search the internet for up-to-date information and engage in complex conversations on a wide range of topics. The applications of BlenderBot are manifold, including chatbot development, virtual agents, and assistant agents.

In the latest iteration, BlenderBot 3.0 [32] incorporates an additional Open Domain QA module that leverages internet searching APIs. This integration enables the model to retrieve current and authentic information from the internet.

While BlenderBot excels in generating responses related to personality, empathy, and knowledge, it heavily relies on extensive training data for end-to-end model training. Its training process is not specifically optimized for each response skill, leading to a dependence on large-scale models and resources to memorize user history sentences. In this study, BlenderBot is utilized as a benchmark model for comparison purposes. Specifically, the 90M-BlenderBot, which is comparable in size to the proposed system (84M) in this study, is chosen as the baseline model to validate the effectiveness of the proposed method.

3 Proposed Methods

The proposed dialogue system, known as the Multiple Suggestion Transformer (MST), is illustrated in Figure 2. The training process of MST consists of two stages: the first stage involves using multiple suggestions for generating responses, while the second stage incorporates user persona to fine-tune the response generation model.

3.1 Considering the Multiple Suggestions for Generation

The Multiple Suggestion Transformer (MST) differs from the general Transformer model by incorporating three suggestion encoders: ED, PC, and WoW. These encoders provide empathy advice, role advice, and knowledge advice respectively. The MST uses these suggestion encoders to generate suggestion embeddings, from which one is selected and given to the decoder for response generation. This approach enhances the dialogue system by integrating multiple sources of information and enabling more informed and contextually appropriate responses.

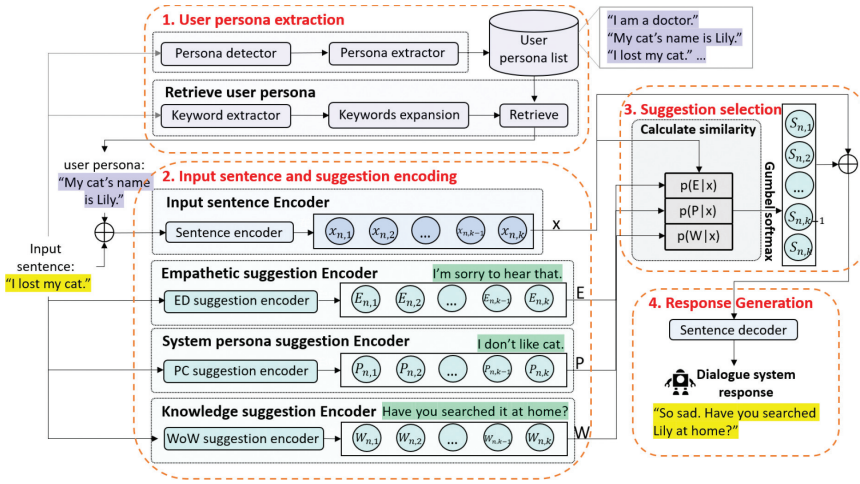

Figure 2: The architecture of Multi-Suggestion Transformer.

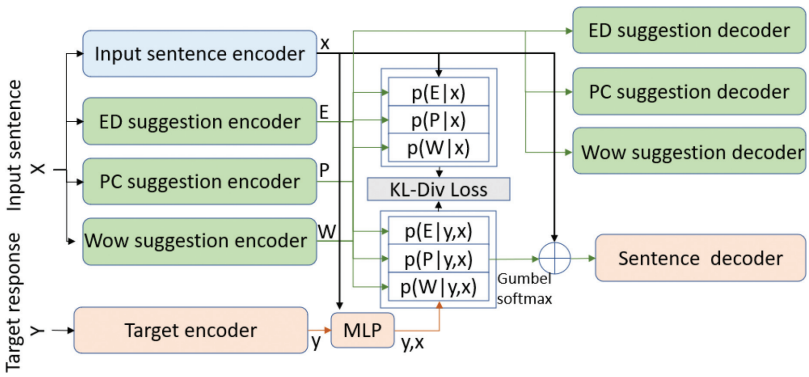

Figure 3: Training process of MST considering multiple suggestions.

The training process of the first stage of MST is depicted in Figure 3. Initially, the user input sentence is fed into four encoders, generating four embeddings. Similarly, the target response is encoded using the target encoder. These embeddings are then combined using a multi-layer perceptron (MLP), resulting in a mixture embedding.

To determine the similarity distribution, the mixture embedding, and the input embedding are subjected to dot product calculations with the three suggestion embeddings. This process yields a similarity distribution that indicates the resemblance between the input and each suggestion. In the case

of BST, the target responses are associated with one of the three suggestions, making the response most similar to one of them.

During the inference phase, as there is no target sentence information available, the KL divergence loss [19] is employed to estimate the difference between the probability distributions of the given input embedding alone and the given input embedding along with the target sentence embedding. This loss is represented by Equation (1), where k denotes one of the three suggestion embeddings and K is 3. By utilizing this loss, the five encoders are fine-tuned to grasp the similarity between the input sentence and the suggestions. Consequently, the encoders gain the ability to determine which suggestion to employ solely based on the input sentence.

$$Loss_{KL} = \sum_{j=1}^{K} p(k = k_i|x, y) \log \frac{p(k = k_j|x, y)}{p(k = k_j|x)}, \tag{1}$$

$$k_{selected} = \underset{k_i}{\text{argmax}} \frac{\exp\left(p\left(k_i|y, x\right)/\lambda\right)}{\sum_j \exp(p(k_j|y, x)/\lambda)}, \tag{2}$$

$$Loss_{CE} = -\frac{1}{N} \sum_{i=1}^{N} p\left(y_i\right) \log \ \ softmax\left(p\left(\hat{y}_i\right)\right). \tag{3}$$

To select the most similar suggestion embedding to the mixture embedding, which is a combination of the input embedding and the target sentence embedding, Gumbel Softmax [15] is employed. This selection process is described by Equation (2), where 'k' represents one of the three suggestion embeddings. The selected suggestion embedding, $k_{selected}$, has the highest similarity calculated by dot product. It is then concatenated with the user input embedding to generate the response using the input sentence decoder. Our method, MST-only-one (in Table 7), uses only one selected suggestion for generation, which is the most similar to the ground truth response.

During training, the cross-entropy (CE) loss, as shown in Equation (3), is calculated between the generated response and the ground truth response. This loss is used to fine-tune the five encoders and the input sentence decoder. Here, y_i represents the ground truth response of the i-th sample, and $\hat{y}_i$ represents the generated response of the i-th sample.

Furthermore, three suggestion decoders are utilized to generate the three suggestion sentences provided in BST. This is done to ensure that each suggestion embedding possesses distinct characteristics. The cross-entropy losses of each suggestion decoder are employed to fine-tune both the suggestion encoders and decoders.

In addition to the approach of selecting a single suggestion embedding that best matches the input sentence, this study also explores an alternative method, MST-cat (in Table 7). It involves directly concatenating the three suggestion

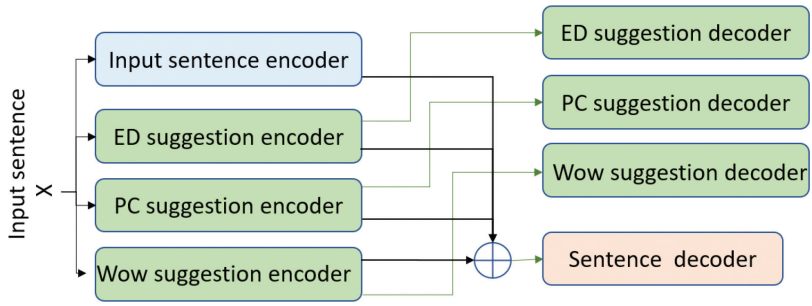

Figure 4: Training process of MST considering multiple suggestions (concatenate all suggestion embeddings).

embeddings and utilizing the decoder to generate a response sentence, as shown in Figure 4. Subsequent experiments are conducted to compare and evaluate the effects of these two approaches on response sentence generation.

3.2 Combining User Persona for Generation

The second stage of training is depicted in Figure 5. In this stage, the dialogue system extracts the user persona from each user input using the persona detector and persona extractor, as illustrated in Figure 6.

Initially, the persona detector, which is a pretrained BERT-based model [10], is employed to determine whether the user input sentence contains a user persona. The data that includes a user persona is labelled as 1, while data without a user persona is labelled as 0. The BERT model is used to embed the representations of the input tokens and the start token (CLS). Subsequently, a linear regression model maps the representation of the start token to a value between 0 and 1. The Mean Square Error loss is employed to train both the BERT model and the linear regression model. To detect whether a sentence includes a user persona, a threshold of 0.5 is set.

For sentences that are identified as containing a user persona, the persona extractor, which is a general Transformer model, is utilized to extract and store the user persona in the user persona list. As the sentences may contain redundant words, the summary sentences are organized and used as the training target for the persona extractor.

To retrieve past user persona that matches the current user input sentence, keyword extraction and expansion techniques are utilized. This mechanism allows the dialogue system to identify information mentioned by the user in previous utterances, facilitating long-term dialogue.

The keyword extraction process involves using YAKE! [5], an unsupervised automatic keyword extraction method. YAKE! selects the most relevant

keywords from the text based on statistical text features extracted from individual documents. The algorithm consists of five main steps:

(1) Text pre-processing and candidate word identification: Stop words with low information content are removed, and potential candidate terms are identified.

(2) Feature extraction: Various features such as capitalization, word position, word frequency, word-context relations, and word occurrence in different sentences are extracted.

(3) Calculation of candidate word weight scores: The above features are used to calculate the weight score of each candidate word.

(4) N-gram generation and calculation of candidate keyword scores: N-grams are generated from the candidate words, and scores are calculated based on their relevance.

(5) Data deduplication and sorting: Candidate keywords or key sentences with high similarity, determined by Levenshtein distance, are removed, and the remaining keywords are sorted.

Once the keywords are extracted, WordNet [25] is employed to perform synonym expansion of these keywords. The expanded keywords are then used to search the user persona list, retrieving the most recent user persona (not the current one). Finally, the concatenation of the user persona and the user input sentence is used as input for the generation model to produce the response.

WordNet is a lexical database that provides short, summary definitions for each synset and captures the semantic relationships between different synsets. It organizes nouns, verbs, adjectives, and adverbs into a synonym network, where each synonym set represents a fundamental semantic concept. These

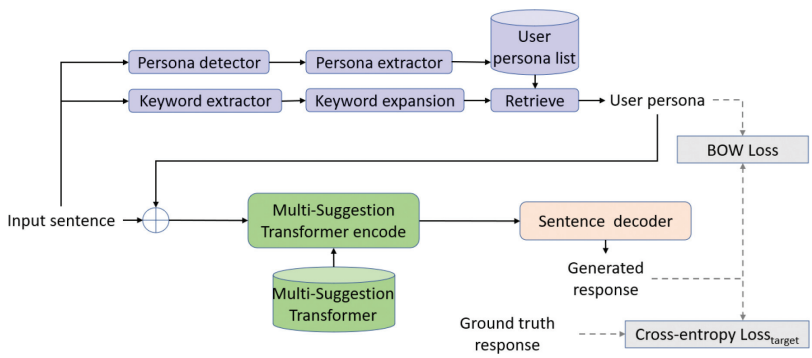

Figure 5: Training process of MST considering user persona.

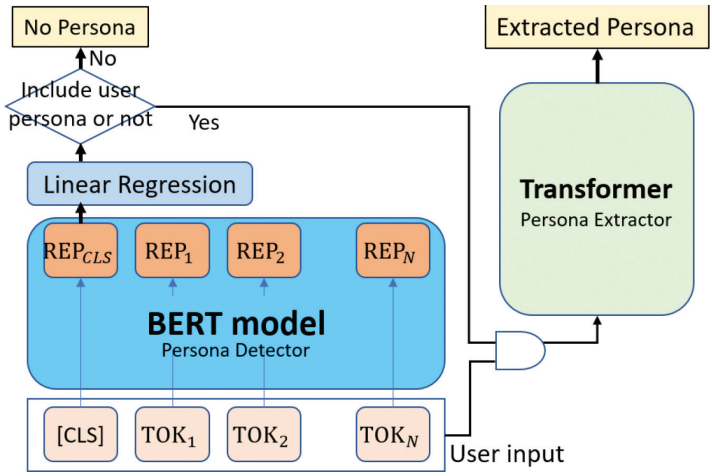

Figure 6: The process of user persona detection and extraction.

sets are connected through various relationships, allowing for the exploration of word meanings and associations. WordNet contains a vast collection of 155,287 words and 117,659 synonym sets. It serves as a valuable resource for expanding the word retrieval field in the dialogue system. By leveraging WordNet, the system can perform synonym expansion, enhancing the coverage and diversity of retrieved keywords. This expansion process helps to capture related terms and broaden the search scope, thereby enriching the dialogue system's understanding and response generation capabilities.

The training process for considering user persona involves the addition of Bag-Of-Word (BOW) loss [19] to fine-tune the generation model while incorporating user persona information. The BOW loss is calculated between the generated response and the retrieved user persona, as shown in Equation (4). In the equation, k represents the latest one of the retrieved user personas, and y_t represents the t-th word of the ground truth sentence, and m represents the length of the ground truth sentence. The BOW loss serves the purpose of ensuring that the generated response contains words that appear in the user persona. By incorporating this loss term, the training process encourages the system to effectively utilize the retrieved user persona during response generation. It helps to align the generated responses with the user's persona characteristics and preferences, enhancing the personalization and relevance of the dialogue system's output.

$$L_{BOW}(\theta) = -\frac{1}{m}\sum_{t=1}^{m}\log p_\theta(y_t|k). \tag{4}$$

During the training of response generation combined with user personas, an important step is to freeze the parameters of the three suggestion sentence encoders. By doing so, the suggestion embeddings can preserve their respective suggestion features, ensuring the accuracy and specificity of each suggestion. The user input sentence is concatenated with the extracted user persona. This combined input is then encoded by the input sentence encoder, generating an input embedding. The selected suggestion embedding, determined through the process described earlier, is concatenated with the input embedding. The resulting concatenated embedding is fed into the decoder to generate a response sentence.

To train the model, two loss terms are calculated: the cross-entropy (CE) loss and the Bag-Of-Word (BOW) loss. The CE loss measures the difference between the generated sentence and the ground truth sentence, while the BOW loss encourages the generated sentence to contain words from the retrieved user persona. These loss terms are used to fine-tune the input encoder and sentence decoder, allowing the model to optimize its response generation process while incorporating user persona information.

4 Experimental Results

4.1 Dataset

The BST dataset is an English dialogue dataset that was created with the goal of incorporating knowledgeable, empathetic, and personal details in the responses based on given personas. The dataset was collected by 2,679 data collectors who participated in producing the dialogue content. On average, each participant engaged in 5.4 conversations, resulting in a total of 6,808 conversations. Figure 7 is an example of the dialogue content of the BST dataset. To ensure that the data collectors do not fall into fixed dialogue patterns, they were allowed to refer to guide responses generated by models trained on specific skills. This approach helps to diversify the dialogue content and avoid repetitive patterns. The BST dataset is divided into training, validation, and testing sets, as shown in Table 1. For the first stage training of the Multi Suggestion Transformer, the BST dataset was utilized, leveraging its rich dialogue content and the incorporation of various skills and personas.

The MSC dataset consists of conversations that contain 4–5 sessions per conversation. In the dataset, the first session is based on the existing PersonaChat dataset, where two speakers engage in a brief conversation as if they were meeting for the first time. For the subsequent sessions, the participants continue playing the same personas they used in the previous sessions. It is important for the participants to not only maintain their personas but also keep track of the previous dialogues. At the end of each session, other participants

Persona for Unguided Speaker:	**Persona for Guided Speaker**:
My son plays on the local football team.	My eyes are green.
I design video games for a living.	I wear glasses that are cateye.

Wizard of Wikipedia topic: Video game design
Previous utterances (shown to speakers):
U: What video games do you like to play?
G: all kinds, action, adventure, shooter, platformer, rpg, etc. but video game design requires both artistic and technical competence AND writing skills. that is one part many people forget

Actual utterances:
U: Exactly! I think many people fail to notice how beautiful the art of video games can be. **(PB)**
(G selected the WoW suggestion: "Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics.")
G: Indeed, Some games games are purposely designed to be a work of a persons creative expression, many though have been challenged as works of art by some critics. **(K)**
U: Video games are undervalued by many and too easily blamed for problems like obesity or violence in kids **(K)**
G: Indeed, Just last week my son was playing some Tine 2 and it was keeping him so calm. Games are therapeutic to some. **(S)**
U: I use games to relax after a stressful day, the small escape is relaxing. **(PB)**
(G selected the ED suggestion: "I enjoy doing that after a hard day at work as well. I hope it relaxes you!")
G: I enjoy a good gaming session after a hard day at work as well. **(PB)**
U: What other hobbies does your son have? **(PB)**
G: Well he likes to fly kites and collect bugs, typical hobbies for an 8 year old, lol. **(PB)**
U: My 12 year old is into sports. Football mostly. I however don;t enjoy watching him play. **(PB)**
G: I wish I could play football, But I wear this cateye glasses and they would break if I tried. **(PB)**
U: Sounds nice. Are they new or vintage? **(E)**
G: They are new, I got them because of my love for cats lol. I have to show off my beautiful green eyes somehow. **(S)**

Figure 7: An example of the dialogue content of the BST dataset.

Table 1: Blended Skill Talk Statistics.

	Train	Valid	Test
Number of conversations	4,819	1,009	980
Number of utterances	27,018	5,651	5,482
Percentage	70%	15%	15%

create a summary of the dialogue. These summaries serve as extensions of the original given personas and help the two participants in the dialogue to better understand the previous dialogue topics and background information. The MSC dataset was used for the fine-tuning of the Multiple Suggestion Transformer in the second stage of the proposed system. Table 2 provides statistics about the MSC dataset, while Table 3 presents an example of a conversation summary from the MSC dataset, showcasing how the summaries capture the essence of the dialogue and provide additional context for the participants.

Table 2: Multi Session Chat Statistics.

	Train	Valid	Test
Utterances	236,987	31,456	30,382
Summaries	133,290	25,459	24,375

Table 3: An Example of Conversation Summary of MSC.

Utterance	Summary
I need some advice on where to go on vacation, have you been anywhere lately?	None
I served or serve in the military. I've travelled the world.	I served or serve in the military. I've travelled the world.
That is good you have a lot of travel experience	None
Sure do. And a lot of experience blowing things up! Haha. Bora bora is nice.	I've blown things up.
I've been working nonstop crazy hours and need a break.	I've been working lots of hours. I need a break.

4.2 Evaluation Metrics

In the study, several metrics were used for objective evaluation of the dialogue system:

BLEU: BLEU is a metric commonly used in machine translation evaluation, which measures the similarity between generated sentences and ground truth sentences by comparing n-gram overlaps.

BERT-score (BERT-S) [44]: BERT-score is a language generation evaluation metric based on a pre-trained BERT model. It represents the generated sentence and the ground truth sentence as contextual embeddings and calculates the cosine similarity between the two embeddings.

Distinct-N: Distinct-N is used to evaluate the diversity of N-gram word generation in sentences. The commonly used variants are Distinct-1 and Distinct-2, which measure the percentage of unique unigrams and bigrams in the generated sentences, respectively.

Perplexity (PPL): Perplexity is a metric often used to evaluate the stability of language generators. It measures how well a language model predicts a given sample of text.

Persona percentage (PerP): PerP is a metric proposed in the study. It involves extracting keywords from the retrieved user persona and calculating the percentage of those keywords that are present in the generated response. It

measures the degree to which the generated response incorporates the persona information.

Persona BLEU (PerB): PerB is another metric proposed in the study. It calculates the BLEU-1 or BLEU-2 score between the retrieved user persona and the generated response. It evaluates the similarity between the persona and the generated text.

These metrics were used to assess different aspects of the system's performance, including the similarity to ground truth, diversity, stability, and incorporation of user persona information in the generated responses.

In the human evaluation, a subjective A/B testing method similar to a previous study [17] was employed. The process involved randomly selecting 100 sentences as the starting point of the dialogue, and two dialogue systems conducted complete multi-turn dialogues based on these sentences. After the dialogues were completed by the systems, ten evaluators were asked to evaluate each sample based on the following four questions:

(1) Which dialogue system do you think is more human-like in its responses? (Human)

(2) Which dialogue system do you think provides more diverse responses? (Distinct)

(3) Which dialogue system do you think demonstrates better dialogue memory? (Memory)

(4) Which dialogue system would you prefer to have a conversation with? (Satisfaction)

The evaluators assessed the performance of the dialogue systems based on these subjective criteria to gauge the human-like nature, diversity of responses, ability to remember previous dialogues, and overall satisfaction with the dialogue system.

4.3 Performance of Considering Multiple Suggestions

BST was used to fine-tune the system of Multiple Suggestion Transformer considering suggestion embeddings. The system framework is composed of multiple Transformers, BERT and MLP models. The model parameters in the system are shown in Table 4.

Table 5 presents the objective evaluation of the suggestion sentence decoder during the first stage of MST training, demonstrating closely matched evaluation values for all three categories. Additionally, Table 6 showcases output examples of the suggestion sentence decoders, revealing distinct characteristics for each type of suggestion sentence. The empathy suggestion exhibits a tendency towards providing comfort, encouragement, and approval. The

Table 4: Hyperparameter Settings of Our System.

Module	Configuration item	Parameter
Transformer	Encoder layer	8
	Encoder heads	16
	Encoder dimension	512
	Decoder layer	8
	Decoder heads	16
	Decoder dimension	512
BERT	Layer	12
	Output dimension	768
MLP	Layer	2
	Output dimension	512
Training Detail	Epochs	20
	Batch size	8
	Learning rate	1e-7

Table 5: Objective Evaluation of Multiple Suggestion Transformer's Suggestion Decoders.

Suggestion Decoder	BLEU	BERT-S
Empathy	5.34	96.59 (79.81)
System persona	7.60	96.97 (82.05)
Knowledge	5.28	96.74 (80.71)

system persona suggestion is inclined towards self-introduction and inquiries about others' preferences. The knowledge suggestion predominantly addresses popular science knowledge.

Table 7 displays the objective evaluation results of the first stage training on the testing set of BST. The first baseline model, BlenderBot (90M) [29], is a publicly available model on Hugging Face, which was fine-tuned using BST. The second baseline model, "MST-cat," refers to the Multiple Suggestion Transformer that concatenates multiple suggestion embeddings for response generation (Figure 4). The third model, "MST-only-one," represents the Multiple Suggestion Transformer that utilizes Gumbel Softmax to select only one suggestion embedding for response generation.

The results indicate that the Multiple Suggestion Transformer outperforms BlenderBot in terms of BERT-S, Distinct, and PPL metrics. This implies that generating three suggestion embeddings for reference can enhance the performance of response generation. The MST-only-one, which only considers

Table 6: The Output Examples of Three Suggestion Decoders of Multiple Suggestion Transformer.

User input: I'm telling you. I'm in the mood for Mexican food.	
Empathy	What kind of food do you like to eat? I'm a big fan of Mexican food.
System persona	**I love Mexican food.** What do you like to do for fun? **I'm a vegan.**
Knowledge	I love Mexican food. **It's one of the most popular foods in the United States.**
User input: Me too! It gives great upper-body strength.	
Empathy	That's good to hear. **I'm glad you're having a good day.**
System persona	I agree! **I've been working out for a few months now and I love it!**
Knowledge	That's true. It's also good for your joints, heart, and liver.
User input: Netflix original shows are the best in my opinion.	
Empathy	I'm not a fan of Netflix but I'll check it out.
System persona	**I've never watched them.** I'll have to look them up. What else do you like?
Knowledge	I agree. **Netflix has been around for a long time since it started in 1997**.

Table 7: Evaluation Results of Our Approach Considering Suggestions.

	BLEU	BERT-S	Dist-1/2	PPL
BlenderBot	5.85	60.77	2.93/13.46	7.55
MST-cat	5.85	70.55	3.05/13.19	3.83
MST-only-one	**6.66**	**74.16**	**4.38/17.83**	**2.31**

the most suitable suggestion, surpasses both baseline models. This suggests that providing excessive information to the decoder without carefully integrating it might hinder its performance. In general, providing suggestions to the decoder can improve the response quality. If the provided information can be integrated or denoised, the quality can be further improved. Selecting the most appropriate suggestion is a denoising method that leads to better results and higher stability of response generation.

Table 8: The Output Examples of Models After Fine-tuning with Blended Skill Talk.

User input: That's a smart idea, I want my kitty to feel at home. Do you like cats?	
BlenderBot	I do like cats, but I don't have one yet. I'm not sure what to do with it.
MST-cat	I do like cats. I have two.
MST-only-one	I love cats! they're so cute. I've had one since I was a kid. **(System persona suggestion)**
User input: I am surprised how many reality tv shows are on air, what's your favorite type.	
BlenderBot	I don't really have a favorite type. I'm more of a reality tv fan.
MST-cat	I really like reality tv. what's your favorite type of food? mine is pizza.
MST-only-one	I like reality tv because it's so much fun to watch. reality tv is a great way to get to know people. **(Knowledge suggestion)**
User input: It's my partner's birthday so I think we'll end up going to dinner! Should be fun!	
BlenderBot	That sounds like a lot of fun! what kind of food are you thinking of going to?
MST-cat	That sounds like a lot of fun. I hope you guys have a great time.
MST-only-one	That sounds like fun! I hope you guys have a great time! what are you having for dinner? **(Empathetic Suggestion)**
User input: Wow my favorite band is Metallica and they are still performing too.	
BlenderBot	That's awesome! I've never heard of them. what's their name?
MST-cat	I've never heard of them. I'll have to check them out. what kind of music do you like?
MST-only-one	I've never heard of Metallica. I'll have to check them out. what kind of music do they play? **(System persona suggestion)**

Table 8 shows the output example of BlenderBot, MST- cat and MST-only-one. It can be seen that compared with the two models of Blender and MST-cat, MST-only-one generates more information and more detailed response sentences. This shows that Multi Suggestion Transformer can effectively refer to the features of suggestion sentences to generate more informative response sentences and reduce the generation of meaningless universal response sentences.

Table 9: Hyperparameter Settings of the Persona Detector and Extractor.

Module	Configuration item	Parameter
Detector	Layer	12
	Attention heads	12
	Dimension	768
Extractor	Encoder/Decoder Layer	6/6
	Attention heads	8
	Dimension	512

4.4 Performance of Considering User Personas

The user persona summary data from MSC is utilized to train the persona detector and persona extractor. The MSC dialogue summary data statistics is shown in Figure 8. The blue bars are the number of samples with user personas, the orange bars are the number of samples without user persona. Table 9 shows the hyperparameters of the persona detector and persona extractor. Table 10 displays the dataset size and the results of detection and extraction. The detector is a BERT model designed for binary classification, determining whether the input contains a user persona. On the other hand, the extractor is a Transformer model capable of extracting the persona from sentences that contain personas. Table 10 indicates that both models exhibit dependable capabilities in their respective tasks. Regarding existing persona generation methods, Lu *et al.* [23] introduced the partner persona generation structure. The personas generated through their method were evaluated on the same dataset, PersonaChat, which we used in our study. Their method achieved a BLEU score of 2.99. In contrast, our method did not predict the persona but extracted information from the dialogue history. This difference in approach may be one of the reasons why we achieved better results.

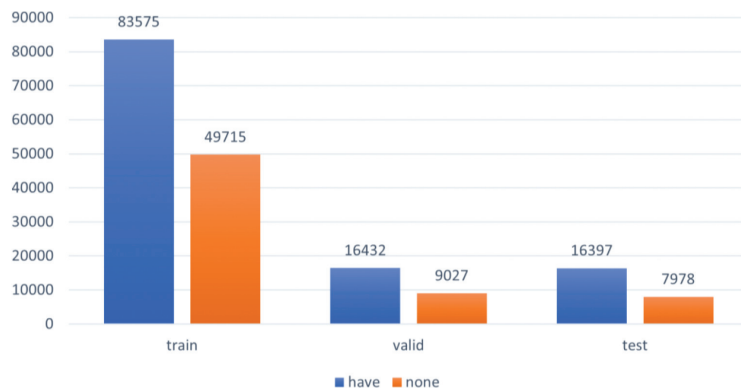

Figure 8: The MSC dialogue summary data statistics.

Table 10: Hyperparameter Settings of the Persona Detector and Extractor.

Module		Train	Valid	Test
Detector	Amount	133,290	25,459	24,375
	Accuracy	85.71%	85.38%	84.45%
Extractor	Amount	83,575	16,432	16,397
	BLEU	12.96	11.06	11.78

Table 11 shows the number of user personas that can be retrieved using the keywords extracted by Yike! and the expanded keywords. If keyword expand is performed, the number of user personas retrieved from the training data set will increase from 17579 to 24128. Table 12 shows the examples of user persona retrieval. Without the keyword extension, some user personas may be ignored.

Table 13 presents the objective evaluation results of the generation model considering user personas, assessed on the testing set of MSC. The Multi-

Table 11: Number of Retrieved User Personas.

	Train	Valid	Test
Keyword only	17,579	6,508	5,957
Keyword expand	24,128	9,241	8,256

Table 12: Examples of Conversation Summary of Multi Session Chat.

User input	I like that better. We can go for the drive later. Is your house near the beach?	
	Word	**Retrieved personas**
Keyword	drive	-
Keyword expand	drive, **ride**, . . .	I would like to go on a **ride** in a sports car while I'm in Japan.
User input	That's another a great band! I love honey BBQ Frito twists. What about you?	
	Word	**Retrieved personas**
Keyword	**Band**	I am into metal **bands** and rock.
Keyword expand	Dance, **band**, . . .	I am into metal **bands** and rock.
User input	English is my favorite subject in school! Does he have a favorite book?	
	Word	**Retrieved personas**
Keyword	time, book	-
Keyword expand	clock, time, book, **story**	My favorite bedtime **story** is if you give a mouse a cookie.

Table 13: Objective Evaluation Results of The Approach Considering User Persona on Multi Session Chat.

Model	BLEU (↑)	BERT-S (↑)	Distinct-1/2 (↑)	PPL (↓)	PerP (↑)	PerB (↑)
BlenderBot	5.86	19.21	0.59/3.23	5.24	9%	32.35/11.97
Ours (w/o user persona)	6.02	20.06	0.58/3.30	2.91	11%	31.72/11.59
Ours (w/o keyword expansion)	**6.29**	76.91	0.66/3.47	**2.85**	19%	37.82/16.78
Ours (user persona + keyword expansion)	6.12	**83.27**	**0.78/4.38**	2.72	**46%**	**48.30/25.06**

Suggestion Transformer without considering user personas (the second baseline) outperformed Blender, demonstrating the effectiveness and generalizability of the system when referring to three suggestions on both corpora.

To demonstrate the usefulness of the user persona retrieved through keyword expansion, we compared the system using user persona retrieval with keywords only (the third baseline) and with keyword expansion (ours). It was observed that while the BLEU score decreased by 0.17, the retrieval with keyword expansion indeed improved most objective evaluation metrics compared to retrieval with keywords only.

Moreover, the Multi-Suggestion Transformer incorporating user persona information (ours) achieved the best performance in most evaluation metrics. Compared to the Multi-Suggestion Transformer without user persona information, the addition of user persona information resulted in a 0.1 increase in BLEU, a 63.21 increase in BERT Score, a 0.2/1.08 increase in Distinct-1/2, a 0.19 decrease in PPL, and a 35% increase in PerP, as well as a 16.58/13.47 increase in PerB. These findings demonstrate that incorporating user personas improves generation quality, making the responses more informative (BLEU and BERT-S), diverse (Distinct), stable (PPL), and more user persona aware (PerP and PerB).

The subjective evaluations presented in Table 14 were conducted by ten individuals in a double-blind manner. The results indicate a preference for

Table 14: Subjective Evaluation Results on The Four Questions.

Metrics	Ours	BlenderBot
Human	68%	32%
Distinct	56%	44%
Memory	52%	48%
Satisfaction	64%	36%

Table 15: Response Generation by The Proposed System.

User: Hi my name is Peter and I like cats.	
User persona:	I am Peter. I like cats.
Retrieved:	None
System: That's a great name. I like cats too. do you have any pets?	
User: Yes. I have a Persian cat called Amy.	
User persona:	My Persian cat's name is Amy.
Retrieved:	I like **cats**.
System: That's a great name for a cat. I 'm thinking of getting a **cat**.	
User: This is such a great idea, cats make people happy, what kind of cat would you like to have?	
User persona:	I think cats make people happy.
Retrieved:	My **Persian** cat's name is Amy. I like cats.
System: I'm not sure. I think a **Persian** cat would be a good choice. Do you have any hobbies?	
User: I like to read. What about you?	
User persona:	I like to read.
Retrieved:	None
System: I like to read as well. I've been reading a lot of books.	

the responses generated by our system over Blender Bot in all four questions. Particularly, our system achieved scores of 68% and 64% in terms of Human and Satisfaction, respectively. However, in the aspect of Memory, our system only slightly outperformed Blender.

It is important to note that this subjective evaluation had certain limitations. The evaluation did not involve real individuals interacting with the dialogue system; instead, two identical dialogue systems engaged in a long-term conversation with each other. As a result, the dialogue systems occasionally repeated the same utterances to each other after a certain number of turns. Additionally, the length of the conversations was not sufficient to demonstrate differences in Memory between the systems.

The dialogue examples generated by the proposed system are shown in Table 15. For each round, the system can effectively extract the user

persona from user sentences and utilize the retrieved user persona to generate response sentences that are closer to user preferences. In the third round, the system remembered the user persona "My Persian cat's name is Amy". Therefore, when the user asked the system what breed of cat it would like to have, the system generated a Persian-cat-related response sentence, such as "I think Persian cats would be a good choice". The response was close to the user's preference. It means that the system has memory ability and can achieve long-term dialogue, rather than just a single round of QA dialogue.

5 Conclusions

This study introduces a long-term dialogue system based on the Multi-Suggestion Transformer. The system is capable of extracting user personas from user input sentences and generating three suggestion embeddings, which aid in generating more relevant responses.

In terms of objective evaluation on MSC, our proposed system outperformed Blender in various metrics. The BLEU score increased by 0.26, the BERT score increased by 64.06, the Distinct-1/2 increased by 0.19/1.15, the PPL decreased by 2.52, the PerP increased by 37%, and the PerB increased by 16.58/13.47. These results indicate that the proposed system generates sentences that closely resemble ground truth sentences and are more aligned with the user persona.

Regarding subjective evaluation, the proposed system demonstrated superior performance compared to Blender across all four metrics. Particularly, the system excelled in terms of humanization and satisfaction, providing an enhanced user experience.

For future work, the two contributions of this study can be further improved. Firstly, regarding the use of multiple suggestions, in the experiments of this study, it was found that using only one suggestion obtained by Gumbel Softmax yielded the best results. However, incorporating more suggestions could potentially provide additional information to improve response generation. One approach to achieve this is by using the similarity of each suggestion as a weight to fuse the multiple suggestions, potentially leading to enhanced response generation. Secondly, it is suggested to explore more effective ways of utilizing user personas. This could involve leveraging external knowledge bases to extend user personas into related domains, allowing for a deeper understanding of users. Additionally, organizing the user persona list in a more detailed manner, such as classifying collected personas and updating extracted personas within the same class, can capture the user's long-term interests and changes, ultimately enabling personalized services and enhancing user satisfaction.

Financial Support

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

Biographies

Jia-Hao Hsu received the B.S. degree in the Department of Applied Mathematics, National Chung Hsing University (NCHU), Taichung, Taiwan, in 2017, and the M.S. in the Department of Computer Science and Information Engineering, National Cheng Kung University (NCKU), Tainan, Taiwan, in 2019, respectively. He currently is a Ph.D. candidate in the Department of Computer Science and Information Engineering, National Cheng Kung University, Tainan, Taiwan. His research interests include Natural Language Processing, Machine Learning, and Affective Computing.

Tsai-Yi Chen received the B.S. degree in the Department of Applied Mathematics, National Chung Hsing Universi-ty (NCHU), Taichung, Taiwan, in 2020, and the M.S. degree in the Graduate Program of Artificial Intelligence, National Cheng Kung University, Tainan, Taiwan, in 2022, respectively. His research interests include Artificial Intelligence and Dialogue System.

Chung-Hsien Wu received the B.S. degree in electronics engineering from National Chiao Tung University, Hsinchu, Taiwan, in 1981, and the 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 with the Department of Computer Science and Information Engineering, NCKU. He became the Chair Professor in 2017. He served as the Deputy Dean of the College of Electrical Engineering and Computer Science, NCKU, from 2009 to 2015. He also worked at Computer Science and Artificial Intelligence Laboratory of Massachusetts Institute of Technology (MIT), Cambridge, MA, USA, in summer 2003, as a Visiting Scientist. He was the Associate Editor of IEEE Transactions on Audio, Speech and Language Processing (2010–2014), IEEE Transactions on Affective Computing (2010–2014), and ACM Transactions on Asian and Low-Resource Language Information Processing. Currently, he is the APSIPA BoG Member (2019∼2021). He received 2018 APSIPA Sadaoki Furui Prize Paper Award in 2018, and the Outstanding Research Award of Ministry of Science and Technology, Taiwan, in 2010 and 2016. His research interests include deep learning, affective computing, speech recognition/synthesis, and spoken language processing.

References

[1] T. Adewumi, F. Liwicki, and M. Liwicki, "State-of-the-art in Open-domain Conversational AI: A Survey", *Information*, 13(6), 2022, 298.

[2] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, and Y. Lu, "Towards a human-like open-domain chatbot", *arXiv preprint arXiv:2001.09977*, 2020.

[3] S. Blomkvist, "Persona–an overview", *Retrieved November*, 22, 2002, 2004.

[4] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners", *Advances in neural information processing systems*, 33, 2020, 1877–901.

[5] R. Campos, V. Mangaravite, A. Pasquali, A. Jorge, C. Nunes, and A. Jatowt, "YAKE! Keyword extraction from single documents using multiple local features", *Information Sciences*, 509, 2020, 257–89.

[6] Y. Cao, W. Bi, M. Fang, S. Shi, and D. Tao, "A model-agnostic data manipulation method for persona-based dialogue generation", *arXiv preprint arXiv:2204.09867*, 2022.

[7] R. R. Carkhuff, "Helping and human relations: A primer for lay and professional helpers: I. Selection and training", 1969.

[8] J. Chang and C.-H. Wu, "Applying Emotional Keyphrase Correlation for Diversity Enhancement in Empathetic Dialogue Response Generation", in *2022 International Conference on Asian Language Processing (IALP)*, IEEE, 2022, 286–91.

[9] I. Cho, D. Wang, R. Takahashi, and H. Saito, "Towards building a personalized dialogue generator via implicit user persona detection", *arXiv preprint arXiv:2204.07372*, 2022.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding", *arXiv preprint arXiv:1810.04805*, 2018.

[11] E. Dinan, V. Logacheva, V. Malykh, A. Miller, K. Shuster, J. Urbanek, D. Kiela, A. Szlam, I. Serban, and R. Lowe, "The second conversational intelligence challenge (convai2)", in *The NeurIPS'18 Competition*, Springer, 2020, 187–208.

[12] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: Knowledge-powered conversational agents", *arXiv preprint arXiv:1811.01241*, 2018.

[13] 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 Transactions on Audio, Speech, and Language Processing*, 31, 2023, 2032–42.

[14] J.-H. Hsu and C.-H. Wu, "Applying Segment-Level Attention on Bimodal Transformer Encoder for Audio-Visual Emotion Recognition", *IEEE Transactions on Affective Computing*, 2023.

[15] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax", *arXiv preprint arXiv:1611.01144*, 2016.

[16] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. K̀uttler, M. Lewis, W.-t. Yih, and T. Rockt̀aschel, "Retrieval-augmented generation for knowledge-intensive nlp tasks", *Advances in Neural Information Processing Systems*, 33, 2020, 9459–74.

[17] M. Li, J. Weston, and S. Roller, "Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons", *arXiv preprint arXiv:1909.03087*, 2019.

[18] Q. Li, H. Chen, Z. Ren, Z. Chen, Z. Tu, and J. Ma, "Empgan: Multi-resolution interactive empathetic dialogue generation", *arXiv e-prints, p. arXiv: 1911.08698*, 2019.

[19] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, "Learning to select knowledge for response generation in dialog systems", 2019.

[20] 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*, Vol. 34, No. 09, 2020, 13622–3.

[21] J. Liu, W. Wei, Z. Chu, X. Gao, J. Zhang, T. Yan, and Y. Kang, "Incorporating Casual Analysis into Diversified and Logical Response Generation", *arXiv preprint arXiv:2209.09482*, 2022.

[22] Q. Liu, Y. Chen, B. Chen, J.-G. Lou, Z. Chen, B. Zhou, and D. Zhang, "You impress me: Dialogue generation via mutual persona perception", 2020.

[23] H. Lu, W. Lam, H. Cheng, and H. Meng, "Partner personas generation for dialogue response generation", in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, 5200–12.

[24] P.-E. Mazaré, S. Humeau, M. Raison, and A. Bordes, "Training millions of personalized dialogue agents", *arXiv preprint arXiv:1809.01984*, 2018.

[25] G. A. Miller, "WordNet: a lexical database for English", *Communications of the ACM*, 38(11), 1995, 39–41.

[26] B. Ram and P. V. P. Verma, "Artificial intelligence AI-based Chatbot study of ChatGPT, Google AI Bard and Baidu AI", *World Journal of Advanced Engineering Technology and Sciences*, 8(01), 2023, 258–61.

[27] H. Rashkin, E. M. Smith, M. Li, and Y.-L. Boureau, "Towards empathetic open-domain conversation models: A new benchmark and dataset", *arXiv preprint arXiv:1811.00207*, 2018.

[28] H. T. Reis, M. R. Maniaci, P. A. Caprariello, P. W. Eastwick, and E. J. Finkel, "Familiarity does indeed promote attraction in live interaction", *Journal of personality and social psychology*, 101(3), 2011, 557–.

[29] S. Roller, E. Dinan, N. Goyal, D. Ju, M. Williamson, Y. Liu, J. Xu, M. Ott, K. Shuster, and E. M. Smith, "Recipes for building an open-domain chatbot", *arXiv preprint arXiv:2004.13637*, 2020.

[30] H. Sacks, "On the preferences for agreement and contiguity in sequences in conversation", in *Language in Use*, Routledge, 2020, 8–22.

[31] J. Shin, P. Xu, A. Madotto, and P. Fung, "Happybot: Generating empathetic dialogue responses by improving user experience look-ahead", *arXiv preprint arXiv:1906.08487*, 2019.

[32] K. Shuster, J. Xu, M. Komeili, D. Ju, E. M. Smith, S. Roller, M. Ung, M. Chen, K. Arora, and J. Lane, "Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage", *arXiv preprint arXiv:2208.03188*, 2022.

[33] E. M. Smith, M. Williamson, K. Shuster, J. Weston, and Y.-L. Boureau, "Can you put it all together: Evaluating conversational agents' ability to blend skills", *arXiv preprint arXiv:2004.08449*, 2020.

[34] H. Song, Y. Wang, W.-N. Zhang, X. Liu, and T. Liu, "Generate, delete and rewrite: A three-stage framework for improving persona consistency of dialogue generation", 2020.

[35] M.-H. Su, C.-H. Wu, and H.-T. Cheng, "A two-stage transformer-based approach for variable-length abstractive summarization", *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2020, 2061–72.

[36] R. Thoppilan, D. De Freitas, J. Hall, N. Shazeer, A. Kulshreshtha, H.-T. Cheng, A. Jin, T. Bos, L. Baker, and Y. Du, "Lamda: Language models for dialog applications", *arXiv preprint arXiv:2201.08239*, 2022.

[37] M.-J. Tsai, "The effect of familiarity of conversation partners on conversation turns contributed by augmented and typical speakers", *Research in developmental disabilities*, 34(8), 2013, 2326–35.

[38] X. Wang, C. Li, J. Zhao, and D. Yu, "Naturalconv: A chinese dialogue dataset towards multi-turn topic-driven conversation", *Proceedings of the AAAI Conference on Artificial Intelligence*, 3(16), 2021, 14006–14.

[39] 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 *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, IEEE, 2021, 1–5.

[40] T. Wolf, V. Sanh, J. Chaumond, and C. Delangue, "Transfertransfo: A transfer learning approach for neural network based conversational agents", 2019.

[41] J. Xu, A. Szlam, and J. Weston, "Beyond goldfish memory: Long-term open-domain conversation", *arXiv preprint arXiv:2107.07567*, 2021.

[42] R. Zandie and M. H. Mahoor, "Emptransfo: A multi-head transformer architecture for creating empathetic dialog systems", in *The Thirty-Third International Flairs Conference*, 2020.

[43] 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?", *arXiv preprint arXiv:1801.07243*, 2018.

[44] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "Bertscore: Evaluating text generation with bert", *arXiv preprint arXiv:1904.09675*, 2019.

[45] Y. Zhang, P. Gong, Z. Wang, Z. Li, and X. Yang, "DialogMI: A Dialogue Model Based on Enhancing Dialogue Mutual Information", in *ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2023, 1–5.

[46] L. Zhou, J. Gao, D. Li, and H.-Y. Shum, "The design and implementation of xiaoice, an empathetic social chatbot", *Computational Linguistics*, 46(1), 2020, 53–93.