IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 17, 2021, accepted December 27, 2021, date of publication January 6, 2022, date of current version January 13, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3140820

Neural Network With Hierarchical Attention Mechanism for Contextual Topic Dialogue Generation

XIAO SUN[1,2], (Member, IEEE), AND BINGBING DING[1]

[1]School of Computer Science and Information Engineering, Hefei University of Technology, Hefei, Anhui 230009, China
[2]Anhui Artigicial Intelligence Laboratory, Hefei Comprehensive National Science Center, Institute of Artificial Intelligence, Hefei, Anhui 230601, China

Corresponding author: Xiao Sun (sunx@hfut.edu.cn)

This work was supported in part by the General Program of the National Natural Science Foundation of China under Grant 61976078, and in part by the National Key Research and Development Program of China under Grant 2019YFA0706203.

ABSTRACT The encoder-decoder model has achieved remarkable results in natural language generation. However, in the dialogue generation work, we often ignore the influence of the dialogue context information and topic information in the generation, resulting in the generated replies not close to the context or lack of topic information leads to general responses. In this work, we study the generation of multi-turn dialogues based on a large corpus and take advantage of the context information and topic information of the conversation in the process of dialogue generation to generate more coherent context-sensitive responses. We improve upon existing models and attention mechanisms and propose a new hierarchical model to better solve the problem of dialogue context (the HAT model). This method enables the model to obtain more contextual information when processing and improves the ability of the model in terms of contextual relevance to produce high-quality responses. In addition, to address the absence of topics in the responses, we pre-train the LDA(Latent Dirichlet Allocation) topic model to extract topic words of the dialogue content and retain as much topic information of dialogue as possible. Our model is extensively tested in several corpora, and the experiments illustrate that our model is superior to most hierarchical and non-hierarchical models with respect to multiple evaluation metrics.

INDEX TERMS Dialogue context, dialogue generation, hierarchical attention, topic.

I. INTRODUCTION

Conversation systems are widely used in a range of applications, from technical support services to language learning tools and entertainment,such as Microsoft's XiaoIce, Apple's Siri and Google's Google Assistant. Dialogue systems can be categorized as single-turn dialogue and multi-turn dialogue. The generation of single-turn dialogue has made great progress in recent years [1]. A single-turn conversation is characterized by considering only the current utterance to generate a response. For example, the classic Seq2Seq model and subsequent work that improved the encoder-decoder structure have achieved good results in a single turn of dialogue. However, the format of a single turn of dialogue is not consistent with the habits of daily human communication: human communication typically consists of multiple turns of sentence communication, so multi-turn dialogue generation requires further investigation. Compared with a single round of dialogue, multi-turn dialogue faces several challenges. Instinctively, humans tend to adapt conversations to their interlocutor not only by looking at the last utterance but also by considering information and concepts covered in the conversation history [2]. The generation of multi-turn dialogue is not limited to consideration of the current statement but should consider the overall contextual information of the dialogue to obtain more semantic information when generating a reply. Current solutions to this problem fall into two categories: (1) improving the Seq2Seq model to generate diversified replies [3]; (2) adding additional information, such as structured world knowledge, personality or emotion [4], to the training corpus to enrich the content of the generated dialogue.

In this paper, we study the multi-turn dialogue response generation of an open domain session in a conversation system and attempt to train the response generation model based on the response and its context. The context refers

The associate editor coordinating the review of this manuscript and approving it for publication was Kathiravan Srinivasan.

the current sentence and previous dialogue information. Serban *et al.* introduced the HRED model to a dialogue system and improved the classical Seq2Seq model [5]. HRED uses a hierarchical structure to construct multiple turns of dialogue. Additionally, the encoder RNN was used to encode the input sentences,the hidden vector at the last moment is regarded as the encoding vector of the input sentence and as the input vector of the next layer of the RNN. A context RNN in the middle layer is used to encode dialogue-level information, such as the state and intention of the overall dialogue, while the first-layer RNN is used to encode the sentence-level information of a sentence. The sentence representation vector of the first layer output in the middle layer is input at each moment. Therefore, the hidden layer vector of the context RNN can remember the previous dialogue information as a context vector. Finally, the vector encoding the previous dialogue information is used as the input vector of the decoder RNN; So, in the decoding process, in addition to the information of the answer sentence itself, the dialogue context information is included. Serban *et al.*subsequently improved the HRED model, proposed the VHRED model, introduced the idea of variational coding on the basis of the HRED model [6], and added a Gaussian random variable to the context RNN link to enhance the diversity.Tang H et al proposed the structural features are captured by the sequential variational autoencoder component, and the topic modeling component based on Gaussian distribution is used to enhance the recognition of text semantics [7].Although the previous work has begun to consider the influence of the content of the context in the generation, the effect obtained is not very obvious. From the perspective of the generated effect, the context information is not fully utilized.However, how to obtain a better contextual representation is currently the main problem. Some researchers use cosine similarity to define contextual relevance. Xing *et al.* introduced the traditional attention mechanism to the hierarchical model [8]. Compared with the previous model, the new model improved the contextual relevance. Kong *et al.* proposed the HSAN,which use a hierarchical encoder to update the word and utterance representations with their position information respectively [9].We tested the HAN and HSAN model on multiple datasets and output the weight of the attention mechanism for most sentences. Higher attention weights were found to be assigned to words or sentences that are not important in context, which is not ideal for the use of context information and results in considerable loss of the main body information of the sentence. Therefore, in this work, we redesigned the hierarchical attention model to model the context. At the same time, we took into account the problem of topic missing in the generated response, We adopt a multi-task learning approach to use the LDA model to model the context topic, which effectively improves the effect of multi-round dialogue generation. The contributions of this paper are as follows:

(1) The importance of context in multi-turn dialogue generation is illustrated, and an improved solution is proposed.

(2) The current hierarchical model is improved, and a more complete attention mechanism is designed to make the model learn more fine-grained information and maintain contextual relevance.

(3) The importance of dialogue topic information is demonstrated, and adopting a multi-task method to use the pre-trained LDA model to extract topics from the content of multiple turns of dialogue to retain more topic information when generating replies.

II. RELATED WORK

With the maturation of the single-turn dialogue system, multi-turn dialogue generation has attracted increased attention.Chen, H. used adversarial training to generate dialogue [10]. Multi-turn dialogue generation has more challenges and problems than single-turn dialogue generation, but it also has more directions and innovations. In the generation of multiple turns of dialogue, not only the current words but also the previous dialogue context must be considered to generate responses, and the context and themes of the overall dialogue must be taken into account. Therefore, multiple rounds of dialogue result in difficulties that remain to be addressed. Serban *et al.* proposed a hierarchical encoder-decoder model to model the context and then proposed variants to the HRED model, named VHRED and MrRNN [11]. On the basis of the HRED model, Gaussian random variables were introduced to enhance the generation of responses. HVMN adds a memory network to VHRED to improve the quality of dialogue responses, but the appropriateness of the answers is weakened due to the lack of long-term memory [12]. To address this problem, some researchers have attempted to define the relevance of a context using a similarity measure, such as cosine similarity [13].However, these similarities are defined on either word or sentence level, which cannot well tackle the topic lacked problem in multi-turn dialogue generation. Moreover, attention mechanisms have been widely used in natural language processing [14]. Attention can be used to process a single sentence to capture the keywords, but the structure and coherence of the entire dialogue cannot be captured. Therefore, the intention network was used to capture the high-level structure and coherence of real dialogue information [15]. Tian *et al.* proposed the WSeq model and added a weighted attention mechanism to HRED [16]. It is more about how to use the hierarchical method to integrate the representation of each utterance sentence into a representation method. Then, a deep dialogue integration model (DUA) was proposed to solve the problems of noise and redundancy in the direct splicing of past conversations as contextual information in multiple turns of dialogue. DUA(Deep Utterance Aggregation) uses an attention mechanism to mine key information from dialogue and replies, highlighting key information and ignoring redundant information, to obtain a matching score for utterances and the response [17]. In theory, DUA can match more contextual information, but all this requires a huge corpus to support it to achieve the desired effect. Zhou *et al.* proposed a two-hop graph attention mechanism

to enhance the semantic representation of the context so as to better extract context information [18].Zeng *et al.*proposed a multi-task learning method to label conversation and text to improve the quality of generation [19].Tang *et al.*use the Gaussian distribution to enhance the recognition of text semantics [7].Zhang et al proposed a large and adjustable neural dialogue generation model named DalioGPT to ensure the richness of generated content [20].Lewis M et al present BART,a denoising autoencoder for pretraining sequence-to-sequence models [21].Bao *et al.*introduced discrete latent variables to solve the inherent one-to-many mapping problem in response generation [22] and Devlin J et al proposed the BERT,BERT is designed to pre-train deep bidirectional representations from unlabeled text by jointly conditioning on both left and right context in all layers [23].Feng s *et al.* use GAN network to provide information and coherence through two complementary evaluation viewpoints [24].Although these works have made great efforts in multiple rounds of dialogue generation, they have not been able to solve the problem of lack of topics in dialogue generation. Zhang *et al.* proposed a topic drift method to simulate the scene of topic transfer in human daily conversations [25], and Sevegani K made the context coherent by predicting and generating a ''bridging'' utterance connecting the new topic to the topic of the previous conversation turn [26].However, the current dialogue generation model still suffers from several challenging problems: (1) Commonly repeated replies. Despite the neural network being trained on a large corpus, repeated meaningless replies, such as ''I don't know'' and ''how are you?'', occur. (2) Context independence. Even if the context and topic information are given, the generated response may be irrelevant to the context. (3) Loss of contextual topic information. Although considerable work has improved the content generated by multiple turns of dialogue, the topic information retention of multiple turns of dialogue remains a problem to be solved.

In this work, we studied the generation of multiple turns of dialogue while simultaneously improving the existing attention mechanism so that the model can retain more fine-grained information to assist in response generation without deviating from the context.At the same time, we use the pre-trained LDA topic model to control the generation of conversation topics through multi-task learning. Attention mechanisms were first proposed in the field of machine translation to enhance the relevance of the context [16], [27]. Seo *et al.* proposed a hierarchical attention network [28] to accurately focus on objects of different scales and shapes in an image. Li *et al.* proposed a deep reinforcement learning method to generate meaningful and diverse responses or increase the length of the generated dialogue [4]. Dhingra *et al.* proposed an end-to-end dialogue system for information reception [29]. The algorithm, called KB-infobot, is based on knowledge base (KB) reinforcement learning. On the basis of these works, we use the advantages of the current mainstream models and also remodel the discourse and context. In addition to focusing on fine-grained information, we should consider the loss of contextual information during dissemination. To overcome the lack of topic information in the generated response, we adopt the pre-trained LDA model to assign topics to the conversation history to improve the topic relevance of the model's responses.Therefore, we designed a more complete attention mechanism to model the contextual information, and verified its effectiveness through a large number of experiments, and we adopted a multi-task learning method to use the LDA model to capture the topic information in the dialogue for use the final generation.

III. THE PROPOSED MODEL

The proposed method is based on the encoder-decoder framework. Suppose that the dataset format can be described as $D = \{(C_i, Y_i)\}_{i=1}^{N}$. For any turn of dialogue,$(C_{\mathrm{i}}, Y_{\mathrm{i}})$ contains the corresponding target response $\mathrm{Y_i}$, and the composition of $\mathrm{Y_i}$ can be defined as $Y_i = \left(y_{i,1} \ldots y_{i,T\ i}\right)$, where $\mathrm{y_{i,j}}$ is the j-th word of this sentence. The context $\mathrm{C_i} = \left(c_{\mathrm{i},1}, \cdots c_{i,m}\right)$,$c_{i,m}$ is the message information and $\left(c_{\mathrm{i},1}, \ldots c_{i,m-1}\right)$ is the information from the previous turns of dialogue. Given the contextual dialogue corpus $C = (c_1, c_2, \ldots, c_m)$, we use the bidirectional gated unit recurrent neural network (BiGRU) [16] and encode every c_i as a hidden vector $\left(\mathrm{h}_{i,1}, \ldots \mathrm{h}_{i,T\ i}\right)$. Suppose that $c_{\mathrm{i}} = \left(\mathrm{w}_{\mathrm{i},1}, \ldots . w_{\mathrm{i},T\ i}\right)$, $\forall t \in \{1, \ldots T_i\}$, $\vec{h}_{i,t}$ is the t-th hidden−state of the forward GRU, $\overleftarrow{h}_{i,t}$ is the t-th hidden state of a back-ward GRU,$\vec{h}_{i,0}$ is initialized with a isotropic Gaussian distribution,and $w_{i,T\ i}$ is the Ti-th word in the sentence,$\mathbf{e}_{i,t}$ is the embedding of $w_{\mathbf{i},T\ i}$, z_t and r_t are an update gate and a reset gate respectively. The calculation is conducted as follows:

$$\mathbf{h}_{i,t} = concat\left(\vec{h}_{i,t}, \overleftarrow{h}_{i,t}\right) \tag{1}$$

$$\mathrm{z}_t = \sigma\left(\mathbf{W}_z \mathbf{e}_{i,t} + \mathbf{V}_z \overrightarrow{\mathbf{h}}_{i,t-1}\right) \tag{2}$$

$$\mathbf{r}_t = \sigma\left(\mathbf{W}_r \mathbf{e}_{i,t} + \mathbf{V}_r \overrightarrow{\mathbf{h}}_{i,t-1}\right) \tag{3}$$

$$\mathbf{s}_t = \tanh\left(\mathbf{W}_s \mathbf{e}_{i,t} + \mathbf{V}_s\left(\overrightarrow{\mathbf{h}}_{i,t-1} \mathbf{r}_t\right)\right) \tag{4}$$

$$\overrightarrow{\mathbf{h}}_{i,t} = (1 - \mathbf{z}_t)^{\circ} \mathbf{s}_t + \mathbf{z}_t^{\circ} \overrightarrow{\mathbf{h}}_{i,t-1} \tag{5}$$

where $\sigma(\cdot)$ is the sigmoid function and Wz, Wr, Ws, Vz, Vr, and Vs are parameters.

A. WORD-LEVEL ATTENTION

For the input hidden layer vector $\mathrm{h}_{i,j}$, we calculated the attention weight of each word as $\left(\alpha_{i,t,1}, \ldots, \alpha_{i,t,T\ i}\right)$. Although considerable training time is required to use this method to calculate the attention of the sentence, this approach can retain fine-grained information. Utterance c_{i} is converted into a vector $r_{i,t}$, for $\forall \mathrm{i} \in \{1, \ldots, m\}$,$r_{i,t}$ can be obtain by the following formulas, where $\alpha_{i,t,j}$ is the calculated weight value.

$$e_{i,t,j} = \eta\left(\mathbf{s}_{t-1}, \mathbf{l}_{i+1,t}, \mathbf{h}_{i,j}\right) \tag{6}$$

$$\alpha_{i,t,j} = \frac{\exp\left(e_{i,t,j}\right)}{\sum_{k=1}^{T_i} \exp\left(e_{i,t,k}\right)} \tag{7}$$

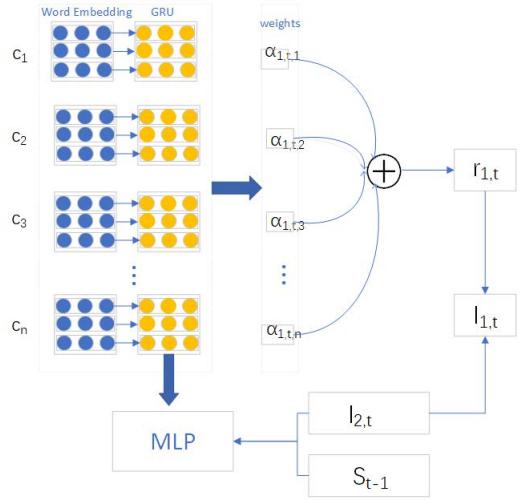

FIGURE 1. Structure of the word-level attention module.

$$\mathbf{r}_{i,t} = \sum_{j=1}^{T_i} \alpha_{i,t,j} \mathbf{h}_{i,j} \tag{8}$$

where $l_{m+1,t}$ is the initialized Gaussian distribution, s_{t-1} is the hidden state of the t-1 decoder, and $\eta(\cdot)$ is a multilayer perceptron with tanh activation function. Then, $\{\{r_{i,t}\}_{i=1}^{m}$ is initialized as a sentence-level encoder and transformed into a context representation vector $(l_{1,t}, \ldots . l_{m,t})$. Figure 1 illustrates the word-level model.

B. CONTEXT-LEVEL ATTENTION

We initialized a GRU unit to encode each sentence, and for each utterance, we used word-level attention to calculate the attention weight of each word. A new representation of the sentence $l_{m,t}$ is obtained after calculating the weight of each word. Inspired by many current works, we improved the current sentence-level attention mechanism to obtain a better contextual representation. The hierarchical attention mechanism proposed by the HRAN model has achieved relatively good results in context modelling. We took advantage of the hierarchical attention and further improve the context-level attention mechanism. We also improved the sentence-level attention. For each sentence, we calculated the context-level attention. First, we calculated the importance of each input sentence, obtained a new representation of each sentence, and then set its attention value to a constant that was not dynamically updated during the decoding process. Therefore, the result based on the weight of the hidden layer state used for decoding is calculated as follows:

$$e_i = V^T \tanh(Wh_i + Ch_s) \tag{9}$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_i \exp(e_i)} \tag{10}$$

$$\gamma_t = \sum_i \alpha_{i,t} h_i \tag{11}$$

where h_i and h_s, respectively, represent the hidden state of the i-th sentence and the last sentence. V, W, and C are parameters. After the weight of each sentence α_i is calculated, the weights do not change throughout the rest of the decoding process. In the decoding process, the hidden layer state st at the t-th step can be calculated as follows:

$$s_t = f(y_{t-1}, s_{t-1}, \gamma) \tag{12}$$

where y_{t-1} is the output of decoding step t-1, the hidden layer state of step t-1 of s_{t-1} decoding, and f is GRU. The weight calculated via attention for each sentence is fixed and is not updated when the parameters are updated.

After the above attention calculation, each sentence is assigned a weight value. Then, we recalculate the attention weights and constantly update the weight of each sentence, as follows:

$$e_{i,t} = V^T \tanh(Wh_i + Cs_{t-1}) \tag{13}$$

$$\alpha_{i,t} = \frac{\exp(e_{i,t})}{\sum_i \exp(e_i)} \tag{14}$$

$$\gamma_t = \sum_i \alpha_{i,t} h_i \tag{15}$$

where V, W, and C are parameters independent of the former attention mechanism. $e_{i,t}$ and $\alpha_{i,t}$ are calculated during the t-th step of decoding, and the hidden layer state at the t-th step is calculated by the following formula,T denotes the transposition operation of V.s_{t-1} is the is the hidden state of t-1-th time step in decoding:

$$s_t = f(y_{t-1}, s_{t-1}, \gamma_t) \tag{16}$$

C. TOPIC ATTENTION

In response to the improvement of the conversation topic, to obtain the topic of the conversation history, we extract topic words to respond to the topic relevance. We use a pre-trained LDA model to assign the topic T to the conversation context. LDA is a document-driven probabilistic model [30]. In our work, we treat the historical information of the conversation as a document, so the most likely topic is sufficient to model the conversation. After the historical dialogue is entered and the subject words of the entire history are obtained, we select n words with the highest probability under T (n = 100 in our experiment). We linearly combine the subject terms $\{t_1, t_2, \ldots, t_n\}$ into a fixed-length vector k, and the subject term attention value is calculated as follows:

$$\beta_{i,t} = \frac{\exp\left(\eta\left(s_{t-1}, t_i, o_{N,t}\right)\right)}{\sum_{j=1}^{n} \exp\left(\eta\left(s_{t-1}, t_j, o_{N,t}\right)\right)} \tag{17}$$

where $i \in \{1, \ldots, n\}$,$o_{n,t}$ is the last hidden state of the context encoding and s_{t-1} is the hidden state of the decoder at t-1. At the same time, we consider the transfer of dialogue information and changes in dialogue state and adverse reactions when multiple topic words are generated. We use the last hidden state of $o_{N,t}$ for topic calculation assignment.

D. DECODER

The main function of the decoder is to decode a response based on the abovementioned context result and topic word assignment. Our goal is to estimate a generation probability $p(y_1, \ldots y_m \mid C)$ from dataset D. Then, in a given context dialogue C, we can generate a reply $Y = (y_1, \ldots, y_T)$. We define $p(y_i)$ as $p(y_i) = p_v(y_i) + p_k(y_i)$, where $p_v(y_i)$ and $p_k(y_i)$ are defined as follows:

$$p_V(y_i = w) = \begin{cases} \frac{1}{N} \exp(\sigma_V(\mathbf{s}_i, \mathbf{y}_{i-1}, w)), & w \in \mathbf{V} \cup \mathbf{K} \\ 0, & w \notin \mathbf{V} \cup \mathbf{K} \end{cases} \tag{18}$$

$$p_K(y_i = w) = \begin{cases} \frac{1}{N} \exp(\sigma_K(s_i, \mathbf{y}_{i-1}, d_i, w)) & w \in \mathbf{K} \\ 0, & w \notin \mathbf{K} \end{cases} \tag{19}$$

$$\mathbf{N} = \sum_{\mathbf{v} \in \mathrm{V}} \exp(\sigma_V(\mathbf{s}_i, \mathbf{y}_{i-1}, w)) + \sum_{\mathrm{k} \in \mathrm{K}} \exp(\sigma_K(\mathbf{s}_i, \mathbf{y}_{i-1}, d_i, w)) \tag{20}$$

where $S_i = \mathrm{GRU}(y_{i-1}, s_{i-1}, d_i, k)$, V represents a reply word, and K represents a topic word. σ is the tanh activation function, and N is the normalizer. Figure 2 shows the framework of the proposed hierarchical attention topic model.

IV. EXPERIMENT

A. DATASETS

1) DAILYDIALOG

[31] Dailydialog is a high-quality open-domain multi-turn dialogue corpus that covers ten major topics in daily life, reflects the human dialogue style, has a definite dialogue mode, and contains a wealth of emotional information. The dataset contains 13k dialogue sequences, each with an average of eight turns. In recent years, many studies have used this dataset to evaluate new models. Because this small chat dataset resembles daily conversation, it is not particularly prominent in terms of topics. As a result, models such as HRAN do not focus on the topic of sentences. To avoid this problem, we performed some data enhancement on this dataset to expand the number of dialogue sequences.

2) EmpChat

[32] Hannah Rashkin *et al.* proposed a new empathy dialogue generation dataset composed of 25k dialogue sequences. Many language structure training corpora are obtained from text scraping, social media dialogue or independent books with little curation. Models trained in this way often result in offensive and cold responses in the dialogue. The EmpChat dataset has been improved in this regard. Every conversation is based on a specific situation, and the participant describes the situation he is in with a specified emotion. The speaker describes his situation at the beginning of the dialogue, and the listener responds empathetically after seeing the description.

3) PersonaChat:

[33] The persona-chat dataset comes from real crowdsourced conversations between people. These crowdsourcers are randomly paired and asked to act according to a given persona. The paired crowdsourcers have a natural conversation and try to get to know each other. The persona-chat data are designed to simulate a conversation where two interlocutors meet for the first time and get to know each other. The goal is to participate in the dialogue process as much as possible, understand each other's interests, discuss their own interests and find common ground.

4) REDDIT

To train the LDA topic model, a scoring corpus that is not the corpus of the abovementioned training dialogue is used. This approach is taken to indirectly introduce the prior knowledge of other data sources to increase the diversity and information of the dialogue. We selected the Reddit dataset, which consists of posts and comments. Each comment has rich metadata (such as author, number of replies, and karma for user comments). To obtain the dataset, we selected 95 English Kanban boards from approximately 1.1 million public Kanban boards. Our selection is based on the top-ranked reddit sections. Topics discussed in these sections include news, education, business, politics, and sports. Furthermore, we categorized the Reddit dataset. According to the number of dialogue rounds, we trained the model using three, four, and five rounds of dialogue, and selected the best LDA model for topic word extraction. Table 1 shows the size and specifications of the working dataset.

TABLE 1. Shows the size of the three datasets. Turns means the number of rounds of a conversation, and length means the length of a conversation.

	turns			length			
Dataset	Max	Avg	Min	Max	Avg	Min	Vocab
Dailydialog	34	5.09	1	262	14.46	2	34989
Personachat	54	11.96	4	61	11.02	2	18096
Empchat	7	2.23	1	101	14.95	2	37109

B. BASELINE

To verify the effectiveness of our proposed model, we used six baseline models for a comparison experiment.

1) HRED

[5] The model proposed by Serban *et al.* was the first to use a hierarchical encoder-decoder structure to model multi-turn dialogue generation.

2) VHRED

[6] Serban *et al.* and others made improvements to the HRED model and proposed VHRED to overcome the difficulty of RNNLM and HRED to produce meaningful and high-quality responses. The idea of VAE is introduced. Additionally, noise is introduced in the middle layer to reconstruct the input data, so the samples from the auto-encoder have a higher globality.

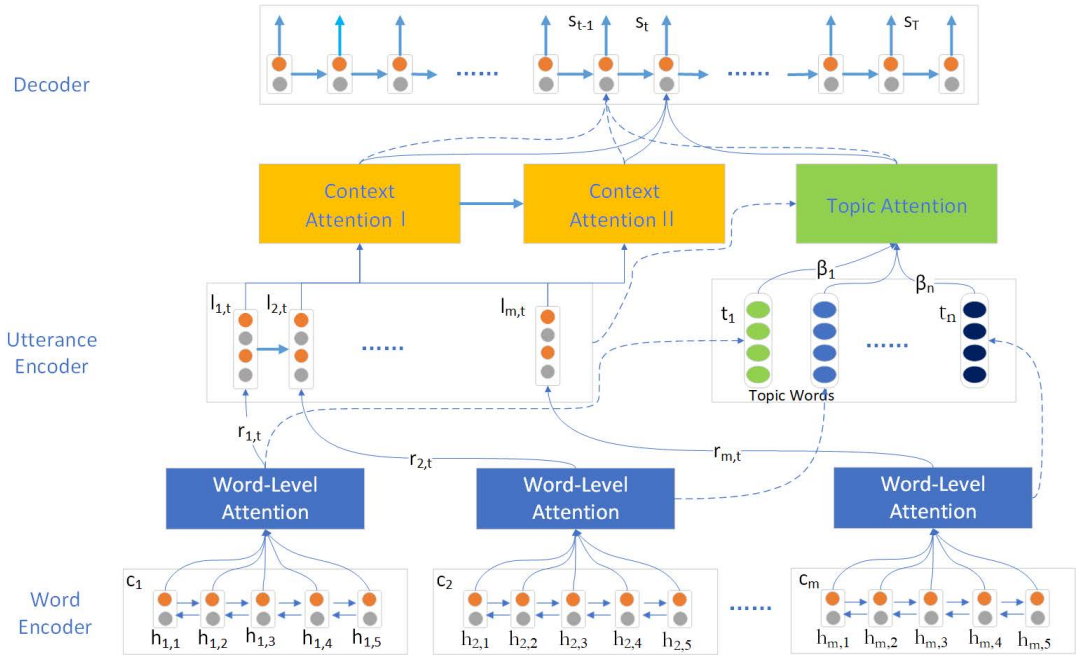

FIGURE 2. Our HAT model. We perform synchronous modelling from the hierarchical attention structure and topic information structure.

3) HRAN

[8] Xing *et al.* used a traditional attention mechanisms to learn the importance of context. HRAN includes repeated attention mechanisms, called the word-level attention mechanism and discourse-level attention mechanism, to make full use of context information. HRAN uses word-level encoders to encode the information of each utterance in the context into latent vectors. Then, when each word is generated, the hierarchical attention mechanism uses word-level attention and sentence-level attention to learn the deep relationship between sentence and sentence opinions.

4) WSeq

[16] The hierarchical WSeq model explicitly weights context vectors through context query relevance (cosine similarity).

5) DSHRED

[34] DSHRED uses dynamic and static attention mechanisms to generate more context-sensitive responses. The query with dynamic attention is the hidden state of the decoder, and the query with static attention is the hidden state of the last sentence in the conversation.

6) ReCoSa

[35] ReCoSa uses multi-head self-attention to detect multiple relative utterances in the context, achieving the most advanced performance. First, the initial representation of the context of each round of dialogue is obtained through a word-level encoder; then, the updated context representation and the masked representation of the response a used simultaneously. Finally, the attention weight between each context representation and the response representation is calculated and used in the process of further decoding the response.

7) HSAN

[9] HSAN ues a hierarchical self-attention network, which attends to the important words and utterances in context simultaneously. Firstly, it use the hierarchical encoder to update the word and utterance representations with their position information respectively. Secondly, the response representations are updated by the mask self-attention module in the decoder. Finally, the relevance between utterances and response is computed by another self-attention module and used for the next response decoding process.

C. TOPIC MODEL

We pre-trained the LDA model on the Reddit dataset. The size of the training data is 1M, and we set the number of topics to 150. The value of α is set to 1/150, and the value of Υ is set to 0.01. Finally, we eliminate the 1000 words with the highest generation rate from the generated topic words to avoid repeated generation when assigning topics to sentences with high probability. Figure 3 shows the model of LDA, First sampling from the Dirichlet distribution θ_i to generate the topic distribution of document i,then sampling from the topic polynomial distribution θ_i to generate the topic $z_{i,j}$of the j-th word of the document i,sampling from Dirichlet distribution β to generate topic $Z_{i,j}$ corresponding

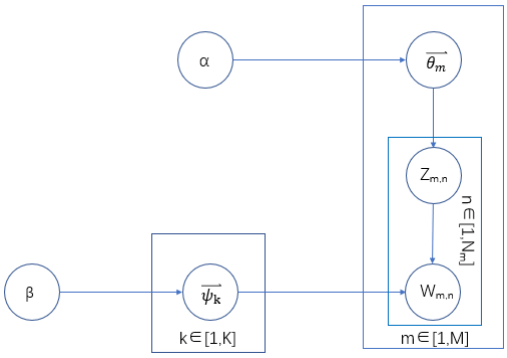

FIGURE 3. Graphic model of LDA.

word distribution $\phi_{zi,j}$,finally sampling from the polynomial distribution $\phi_{zi,j}$ of words and finally generating words $w_{i,j}$.

D. EVALUATION METRICS

1) PERPLEXITY

Perplexity measures the accuracy of a model's prediction of a human response. Perplexity can also be thought of as the average branch factor, that is, how many choices are available when predicting the next word. A lower perplexity usually indicates better generation performance.

$$\mathrm{PPL} = \exp\left\{-\frac{1}{N}\sum_{i=1}^{n}\log\left(p\left(Y_i \mid U_i\right)\right)\right\} \tag{21}$$

2) ROUGE

Evaluating text content based on the co-occurrence information of n-grams in the text is an evaluation method oriented to the recall rate of n-grams. The quality of the text is evaluated by counting the number of overlapping basic units (n-grams, word sequences, and word pairs).The calculation formula is shown in equation 22, as shown at the bottom of the next page.

3) BLEU

This metric used as an evaluation metric in the field of machine translation, and it is also commonly used to evaluate dialogue systems.The metric counts the number of occurrences of n-gram phrases in the generated response and the real response in the entire training corpus. Although use of the BLEU indicator is controversial for the evaluation of dialogue generation, because comparison models were used in their paper and considering the fairness of the comparison experiment, we decided to include this metric.

4) EMBEDDING

[36] The embedding-based metric measures performance by calculating the similarity between sentence embeddings and the generated answers.In this article, we use embedding average, vector extrema and greedy matching as three word vector evaluation indicators. Moreover, we use the Word2Vec tool to train word embeddings on the Google News Corpus dataset for evaluation.

5) DISTINCT

[37] Diversity is a major problem in dialogue generation. As in Li et al, we measure diversity as the ratio of different elemental (dist1) and diatomic (dist2) results in all generated responses. The two dist indicator calculate different unigrams and bigrams. The value of the indicator reflects whether the content generated by the model's response is bad or meaningless.

6) HUMAN ANNOTATION

In addition to the automatic metrics above, we further recruited human annotators to judge the quality of the generated responses of different models. Five labelers with rich experience were invited to do evaluation.Responses generated by different models were pooled and randomly shuffled for each labeler. Labelers referred to the test messages and judged the quality of the responses according to the following criteria:

+2: The response is not only relevant and natural, but also contained the context message and topic information.

+1: The response can be used as a reply to the message, but it is too universal like ''Yes or No'', ''Hello'' and ''I don't know''.

0: The response cannot be used as a reply to the message.

Agreements among labelers were calculated with Fleiss'kappa. [38]

E. PARAMETER SETTINGS

All models are based on pytorch. The hyperparameters of the model are shown in the table2.

TABLE 2. The settings of parameter.

Learning rate decay	0.5
GRU hidden size	512
Utterance encoder layer	2
Bidirectional encoder	True
Context encoder layer	1
Decoder layer	2
Gradient clip	3.0
Learning rate	1e-4
Optimizer	Adam
Dropout ratio	0.3
Weight decay	1e-6
Epochs	80
Word embed size	256
Seed	30
Multi-head	8

F. EVALUATION RESULTS

Tables 3 - 5 presents the results of the automatic evaluation. Our model achieves obvious index improvements on the three corpora. Our hierarchical attention topic model generates richer response content and has a better generation effect. On the dailydialog corpus, the perplexity PPL metric of our method is reduced by approximately 2.8% compared with that of HAN, which is better than the other models. Secondly, in terms of the Bleu, embedding and distinct evaluation

TABLE 3. The results of different models on the Dailydialog dataset.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
HRED	33.83	20.38	9.24	5.21	3.20	4.93	2.18	10.4	60.41	76.02	49.58
VHRED	36.33	21.61	9.991	5.78	3.77	5.13	**3.0**	14.37	61.31	77.06	49.51
HRAN	28.10	22.90	11.17	6.82	4.60	6.25	2.75	14.25	62.57	77.79	51.26
DSHRED	33.55	20.24	8.94	4.75	2.69	4.39	1.50	7.31	60.427	75.98	49.24
WSeq	33.72	21.29	9.46	5.23	3.11	4.71	1.97	9.37	61.13	76.88	49.63
Recosa	32.89	22.18	10.31	5.90	3.71	5.27	1.97	9.92	61.72	77.01	50.22
HSAN	30.52	23.24	10.48	6.13	4.21	5.97	2.06	13.59	61.89	77.09	51.26
HAT	**27.32**	**24.96**	**13.10**	**8.89**	**6.66**	**8.48**	2.98	**16.24**	**63.09**	**77.68**	**53.18**

TABLE 4. The results of different models on the EmpChat dataset.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
HRED	50.34	19.03	7.72	3.52	1.81	2.85	0.56	3.45	69.59	83.99	52.20
VHRED	50.28	18.90	7.57	3.49	1.83	2.71	0.75	4.28	69.58	84.05	51.88
HRAN	46.64	19.11	7.80	3.57	1.84	2.95	0.71	4.19	69.37	83.84	52.23
DSHRED	47.88	18.63	7.44	3.30	1.64	2.74	0.47	3.08	69.61	83.94	51.94
WSeq	47.55	18.26	7.38	3.38	1.74	2.76	0.46	2.73	69.69	83.81	52.00
Recosa	**45.82**	18.82	7.53	3.35	1.68	2.73	0.43	2.88	69.14	83.71	51.95
HSAN	51.16	18.95	7.62	3.64	1.72	2.90	0.77	4.39	69.77	**84.14**	52.97
HAT	47.01	**20.85**	**8.73**	**4.50**	**1.88**	**3.05**	**1.66**	**5.15**	**69.82**	84.08	**53.39**

TABLE 5. The results of different models on the PersonaChat dataset.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
HRED	33.64	21.35	10.28	5.33	2.86	4.74	0.39	2.30	62.58	83.32	48.46
VHRED	**31.93**	20.94	9.77	4.78	2.37	4.18	0.23	0.87	57.05	81.62	45.96
HRAN	32.92	**22.21**	10.63	5.41	2.87	4.78	0.66	3.93	63.20	83.40	49.24
DSHRED	33.73	21.55	10.28	5.25	2.78	4.59	0.43	2.58	62.85	83.36	48.55
WSeq	34.56	21.45	10.28	5.26	2.79	4.69	0.38	2.16	62.45	83.27	48.29
Recosa	36.53	21.30	10.13	5.14	2.74	4.62	0.59	3.69	63.51	83.32	48.83
HSAN	35.48	21.84	10.35	5.22	2.85	4.87	0.62	3.77	63.82	83.62	48.90
HAT	33.86	22.14	**11.60**	**6.41**	**3.87**	**4.97**	**0.79**	**4.97**	**64.74**	**84.46**	**50.54**

metrics, our model outperforms the other models to varying degrees. In terms of Distinct-1, our model and the VHRED model have the same results (VHRED introduces Gaussian variables to enhance the diversity of responses), and in terms of Distinct-2, our model has achieve an approximately 13% compared to the second most effective VHRED model. Compared with the latest Recosa model for multiple turns of dialogue, our model achieves an improvement of approximately 33%. The results confirm the correctness of our method of introducing topic information to enrich the content of dialogue generation. Second, in terms of the embedding, because the HRED uses a hierarchical structure, it does not use an attention mechanism to model the context, resulting in a considerable loss of original text information when generating responses. The HAN mechanism greatly improves this metric, and our HAT model improved the context-level attention compared to HRAN's hierarchical attention. Then, we demonstrate the performance of each model based on two major evaluation metrics: diversity and relevance. Figure 4 and Figure 5 show the model comparison. Our approach outperforms HRAN, indicating that including the hierarchical attention mechanism in the modelling of contextual information will lead to better results.

TABLE 6. The result on Human annotation.

Models	+2	+1	0	Kappa
HRED	32.7%	23.4%	43.9%	0.7126
VHRED	33.9%	23.6%	42.5%	0.7218
HRAN	41.4%	23.2%	35.4%	0.8125
DSHRED	40.3%	25.3%	34.4%	0.8058
WSeq	31.8%	24.6%	43.6%	0.6986
Recosa	34.4%	24.7%	40.9%	0.7416
HSAN	36.2%	26.9%	36.9%	0.7928
HAT	45.6%	26.2%	28.2%	0.8462

Table 6 shows the results of human annotation.We randomly selected 200 samples from the responses generated by each model on three dataset. Obviously, compared with the baseline model, the responses generated by the HAT

$$\text{ROUGE - N} = \frac{\sum_{S\in\{\text{ ReerenceSummaries }\}}\sum_{\text{gram}_N\in S}\sum_{\text{Count }_{\text{match}}}\left(\text{ gram }_N\right)}{\sum_{S\in\{\text{ ReferenceSummaries }\}}\sum_{\text{gram }_N\in S}\text{ Count }\left(\text{ gram }_N\right)} \tag{22}$$

TABLE 7. Case study of a round of dialogue in the Dailydialog corpus.

Message
A: Good evening, **madam**. Can I help you?
B: Yes. I bought this **sheet** here yesterday.
A: Anything **wrong** with it?
B: Yes. When I spread it out back home, I **found** this hole.
A: Oh, how awful! May I have a **look** at the **invoice**?
B: Here it is.
A: Well. Please accept our sincere **apologies**, I'll be glad to change it for another one if you wish.
Other: Oh, thanks for you, I will get it.
Ours: That's very kind of you. I am willing to accept this way, and will consider shopping at your home next time.

TABLE 8. Case study of a round of dialogue in the EmpChat corpus.

Message
A: A few years ago, there was a time that I really needed a **car**. I had to go to school and my husband worked full time.
B: Did you not have a **car**? That's hard. I don't have one now and use Uber and Lyft all the time.
A: We had one car we shared but I had a little **girl** and we were just stuck at the house or got rides. We didn't really have uber and Lydt yet. But someone gave me a **car** out of the blue!
B: That is incredible! I swear some things in life are determined by fate. I'm glad you were able to get a **car**.
A: Thanks me too. It blew me away that **people** can be so **kind**.
Other: Yes,there are many kind people.
Ours: Congratulations, you have your own car and thanks to the kind man.

model can make better use of context and topic information (responses marked as +2) and fewer general responses, achieving the best performance. Among them, compared with HRED, it increased the "+2" response by 12.9% and decreased the "0" response by 15.7%. This means that compared with the traditional attention model, the hierarchical attention model has been well strengthened. At the same time, compared with HRAN, the "+2" response of HAT has increased by 4.2%, and the "0" response has been reduced by 7.2%. At the same time, the Kappa value of the HAT model is also the highest compared to the baseline model. This is a good indication that HAT can better generate sentences containing topic information.

V. DISCUSSION

A. CASE STUDY

We used a sample from the three corpus test set to compare the effectiveness of our model and that of the baseline model. Given the context of seven rounds to predict the generated result, the examples are shown in the Table 6 - 8. The generated content indicates that our model is comparable to the baseline model. The topic words extracted from the context can be used and applied to the generated replies when the prediction is generated so that more information can be reserved for response generation. On the dailydialog corpus,the content of the conversation indicates that the conversation occurred in a store and covered product after-sales issues. HAT can understand the content of the conversation and knows that salesperson B apologized for his product and proposed a solution. The response given by HAT indicates that the solution was accepted. For the prediction of the topic word, HAT predicts "accept" based on "apologize" in the context and then further predicts the word "shopping" according to the content.Then,on the Personchat corpus, the scene in this case should be two friends introducing themselves, including personal information and some basic preferences exchanges. From the generated results, we can see that our model also takes into account the contextual information of the dialogue, including the topic] word information such as "music", "book", "hobbies", etc. appearing above for generation, and the final generated results effectively cover the contextual information and themes. Information, among which topic word information such as "softball", "volleyball", "music", and "together" appear in the generated sentences, which increases the coherence and quality of responses.On the EmpChat corpus, the model has also achieved relatively good results. The sample shows that the generated response contains contextual information and subject word information such as "car" and "kind", which enriches the dialogue Generate effect.

B. MODEL ABLATION

The experimental results in Tables 3 - 5 show that our model achieves substantial improvements with respect to multiple metrics and datasets, because of the hierarchical attention model and topic structure. To verify the impact of the introduction of the topic information structure and hierarchical attention mechanism on the generated results, we performed

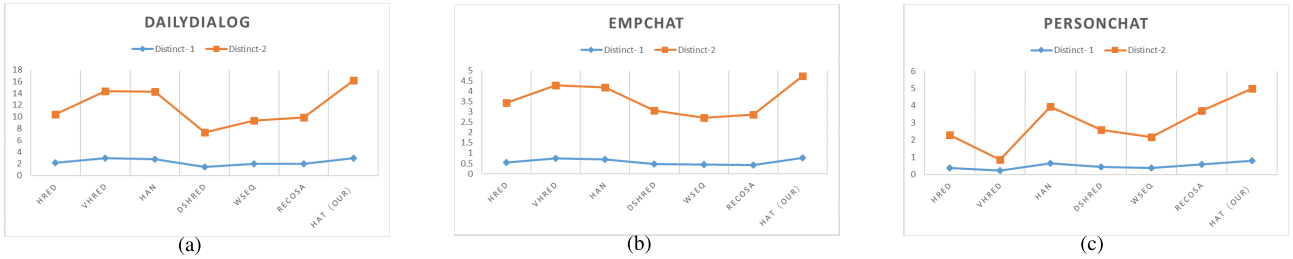

FIGURE 4. Comparison of diversity on different corpora.

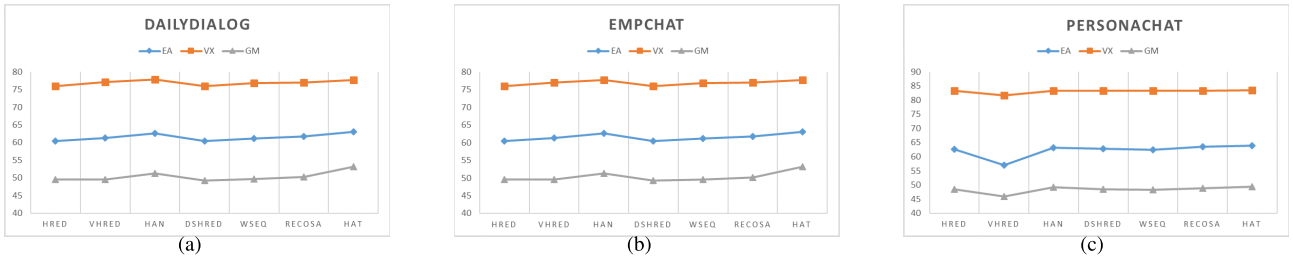

FIGURE 5. Comparison of relevance on different corpora.

TABLE 9. Case study of a round of dialogue in the PersonaChat corpus.

Message
A: I love playing **sports** and being active . I love rap **music** . I love to hang out with my friends . I am in college studying education.i am a 22 year old girl .
B: Hi , I sell **shoes** online , a meat eater and love the rush band . and you ?
A: My favorite **music** is rap , tupac is my favorite.
B: I design **book** covers in my spare time . you write books ?
A: I also enjoy hanging out with **friends** , how about you ?
B: I **love** a grilled steak . my best friends got married last week . you married ?
A: Oh how **nice** , no not yet.
B: love dear mama by pac .
A: One of my favorite **songs** . What kind of **book** covers do you design ?
B: <chic lit>, You work or have any special **hobbies** ?
A: I like to play **softball** and **volleyball** . I guess that is considered a **hobby** .
Other: yeah,I think it is good.
Ours: oh,I like softball and volleyball too,at the same time,I love the music and read book, Can we play it together?

TABLE 10. The result of model ablation on No-Att or No-topic on Dailydialog.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
No-topic	29.22	22.29	11.06	6.90	4.74	6.32	2.73	14.37	62.04	77.10	50.89
No-Att	28.93	22.35	10.92	6.78	3.90	5.92	1.97	13.92	61.84	77.02	50.26

an ablation analysis on the three corpus and removed the hierarchical attention mechanism and topic structure to create new models, respectively, named No-Att and No-Topic. The results in Table 10 - 12 show that after removing the topic prediction module or the hierarchical attention mechanism respectively, our model sufferes different degrees of decline in the three corpus, which confirmed that our model with the topic information and hierarchical attention mechanism can improve the response generation results. In the table 13 - 15, we test the model after removing the topic prediction module and the hierarchica attention mechanism named No-topic/Att.Compared with the original model, the obtained data has different degrees of decline, which verifies that the HAT model has a better improvement in contextual information association and topic prediction.

C. ERROR ANALYSIS

We conducted a statistical analysis of the HAT generation effect on the dailydialog corpus. We randomly selected 100 rounds of dialogue from the test set responses generated

TABLE 11. The result of model ablation on No-Att or No-topic on PersonaChat.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
No-topic	36.25	19.20	8.25	4.72	2.35	4.79	0.61	14.37	62.69	83.35	47.56
No-Att	38.43	19.06	8.02	3.98	2.32	4.72	0.51	13.86	62.57	82.59	47.23

TABLE 12. The result of model ablation on No-Att or No-topic on Empchat.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
No-topic	49.25	18.82	7.68	3.54	1.79	2.94	0.71	3.52	69.60	82.69	52.25
No-Att	49.97	18.36	7.06	2.97	1.62	2.89	0.68	3.48	68.98	82.36	51.27

TABLE 13. The result of model ablation on No-Att and No-topic on Dailydialog.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
No-topic/Att	35.62	17.26	8.72	5.49	2.86	4.67	1.13	11.29	59.24	76.06	49.52

TABLE 14. The result of model ablation and No-Att or No-topic on PersonaChat.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
No-topicAtt	41.56	18.03	7.15	2.87	1.96	3.88	0.49	12.96	60.58	81.26	46.06

TABLE 15. The result of model ablation and No-Att or No-topic on EmpChat.

Model	PPL	BLEU1	BLEU2	BLEU3	BLEU4	Rouge	Dis-1	Dis-2	EA	VX	GM
No-topic/Att	56.38	16.25	6.53	2.02	1.14	2.26	0.59	2.98	67.54	81.69	50.73

by HAT. The statistical results showed that approximately 20% of the sentence responses lacked correct subject information. Potential reasons for this issue include the following: (1) The influence of the dailydialog corpus itself because the corpus contains daily dialogue and does not have strong topicality. Second, the conversational sentence topic changes frequently, and the model cannot learn the context and obtain subject information and contextual information effectively. (2) This model generates replies for topics to predict multiple topic words. However, due to the increase in the number of conversations in multiple turns of a dialogue, the topics discussed at first and the topics discussed later may be completely irrelevant, which can cause deviation in the model predictions. These problems represent directions for future work. Potential ways to address these issues include (1) labelling the topic of the corpus, (2) optimizing the current topic model to improve the contextual topic logic and (3) Improve the model for more tasks.

VI. CONCLUSION

In this paper, we proposed a topic dialogue generation model based on a hierarchical attention mechanism to solve the problem of how to generate a contextual content response in multiple turns of dialogue generation. At the same time, our pre-trained topic model can be assigned topic words in the generation through multi-task learning, which improves the topic relevance of the dialogue.The improvement of our model compared with the baseline model is considerable. Moreover,in the ablation experiment, we demonstrated the importance of introducing themes in dialogue generation and hierarchical attention mechanism and provided a better reference for multiple turns of dialogue generation. In future work, we will further consider the utilization of contextual information and the controllability of generated topics to improve the effect of conversational data in other fields and be suitable for more tasks.

REFERENCES

[1] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 3104–3112.
[2] C. Danescu-Niculescu-Mizil and L. Lee, "Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs," in *Proc. 2nd Workshop Cogn. Modeling Comput. Linguistics*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2011, pp. 76–87.
[3] I. Serban *et al.*, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017.
[4] J. Li *et al.*, "Deep reinforcement learning for dialogue generation," 2016, *arXiv:1606.01541*.
[5] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proc. AAAI Conf. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2015, pp. 1–8.
[6] I. Serban *et al.*, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Proc. AAAI Conf. Artif. Intell.*, vol. 31, no. 1, 2017.
[7] H. Tang, M. Li, and B. Jin, "A topic augmented text generation model: Joint learning of semantics and structural features," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, 2019, pp. 5090–5099.

[8] C. Xing *et al.*, "Hierarchical recurrent attention network for response generation," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
[9] Y. Kong, L. Zhang, C. Ma, and C. Cao, "HSAN: A hierarchical self-attention network for multi-turn dialogue generation," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7433–7437.
[10] J. Li, W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky, "Adversarial learning for neural dialogue generation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2017, pp. 2157–2169.
[11] I. V. Serban *et al.*, "Generative deep neural networks for dialogue: A short review," 2016, *arXiv:1611.06216*.
[12] H. Chen, Z. Ren, J. Tang, Y. E. Zhao, and D. Yin, "Hierarchical variational memory network for dialogue generation," in *Proc. World Wide Web Conf.*, 2018, pp. 1653–1662.
[13] T. Shen *et al.*, "Disan: Directional self-attention network for RNN/CNN-free language understanding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 32, no. 1, 2018.
[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
[15] K. Yao, G. Zweig, and B. Peng, "Attention with intention for a neural network conversation model," 2015, *arXiv:1510.08565*.
[16] Z. Tian, R. Yan, L. Mou, Y. Song, Y. Feng, and D. Zhao, "How to make context more useful? An empirical study on context-aware neural conversational models," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics*, 2017, pp. 231–236.
[17] Z. Zhang, J. Li, P. Zhu, Z. Hai, and G. Liu, "Modeling multi-turn conversation with deep utterance aggregation," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 3740–3752.
[18] S. Zhou, W. Rong, J. Zhang, Y. Wang, L. Shi, and Z. Xiong, "Topic-aware dialogue generation with two-hop based graph attention," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 7428–7432.
[19] Y. Zeng and J.-Y. Nie, "A simple and efficient multi-task learning approach for conditioned dialogue generation," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2021, pp. 4927–4939.
[20] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," 2019, *arXiv:1911.00536*.
[21] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.
[22] S. Bao, H. He, F. Wang, H. Wu, and H. Wang, "PLATO: Pre-trained dialogue generation model with discrete latent variable," 2019, *arXiv:1910.07931*.
[23] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.
[24] S. Feng, H. Chen, K. Li, and D. Yin, "Posterior-GAN: Towards informative and coherent response generation with posterior generative adversarial network," in *Proc. AAAI Conf. Artif. Intell.*, 2020, vol. 34, no. 5, pp. 7708–7715.
[25] K. Sevegnani, D. M. Howcroft, I. Konstas, and V. Rieser, "OTTers: One-turn topic transitions for open-domain dialogue," 2021, *arXiv:2105.13710*.
[26] H. Zhang, Y. Lan, L. Pang, H. Chen, Z. Ding, and D. Yin, "Modeling topical relevance for multi-turn dialogue generation," 2020, *arXiv:2009.12735*.
[27] K. Cho, A. Courville, and Y. Bengio, "Describing multimedia content using attention-based encoder-decoder networks," *IEEE Trans. Multimedia*, vol. 17, no. 11, pp. 1875–1886, Nov. 2015.
[28] P. H. Seo *et al.*, "Progressive attention networks for visual attribute prediction," 2016, *arXiv:1606.02393*.
[29] B. Dhingra *et al.*, "Towards end-to-end reinforcement learning of dialogue agents for information access," 2016, *arXiv:1609.00777*.
[30] W. X. Zhao, J. Jing, J. Weng, J. He, and X. Li, "Comparing Twitter and traditional media using topic models," in *Proc. Eur. Conf. Adv. Inf. Retr.*, 2011, pp. 338–349.
[31] Y. Li *et al.*, "Dailydialog: A manually labelled multi-turn dialogue dataset," 2017, *arXiv:1710.03957*.
[32] H. Rashkin *et al.*, "Towards empathetic open-domain conversation models: A new benchmark and dataset," 2018, *arXiv:1811.00207*.
[33] S. Zhang *et al.*, "Personalizing dialogue agents: I have a dog, do you have pets too?" 2018, *arXiv:1801.07243*.
[34] W. Zhang *et al.*, "Context-sensitive generation of open-domain conversational responses," in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 2437–2447.
[35] H. Zhang, Y. Lan, L. Pang, J. Guo, and X. Cheng, "ReCoSa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3721–3730.
[36] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proc. Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2016, pp. 2122–2132.
[37] J. Li *et al.*, "A diversity-promoting objective function for neural conversation models," 2015, *arXiv:1510.03055*.
[38] J. L. Fleiss and J. Cohen, "The equivalence of weighted Kappa and the intraclass correlation coefficient as measures of reliability," *Educ. Psychol. Meas.*, vol. 33, no. 3, pp. 613–619, Oct. 1973.

XIAO SUN (Member, IEEE) was born in 1980. He received the M.E. degree from the Department of Computer Sciences and Engineering, Dalian University of Technology, in 2004, and a dual Doctor's degree from the University of Tokushima, Japan, in 2009, and the Dalian University of Technology, China, in 2010. His field of study was natural language processing. He is currently working as a Professor with the Anhui Province Key Laboratory of Affective Computing and Advanced Intelligent Machine, Hefei University of Technology. His research interests include affective computing and natural language processing.

BINGBING DING was born in 1996. He is currently pursuing the master's degree with the School of Computer Science and Information Engineering, Hefei University of Technology. His research interests include natural language processing, sentiment analysis, and neural networks.