frontiers | Frontiers in Physics

TYPE Original Research
PUBLISHED 03 October 2022
DOI 10.3389/fphy.2022.1019969

OPEN ACCESS

EDITED BY
Fei Xiong,
Beijing Jiaotong University, China

REVIEWED BY
Qian Liu,
Nanyang Technological University, Singapore
Zhonggui Ma,
University of Science and Technology Beijing, China
Weijia You,
Beijing Forestry University, China

*CORRESPONDENCE
Haitao Xiong,
xionghaitao@btbu.edu.cn

SPECIALTY SECTION
This article was submitted to Social Physics, a section of the journal Frontiers in Physics

RECEIVED 15 August 2022
ACCEPTED 14 September 2022
PUBLISHED 03 October 2022

CITATION
Cai Y, Zuo M and Xiong H (2022), Modeling hierarchical attention interaction between contexts and triple-channel encoding networks for document-grounded dialog generation. *Front. Phys.* 10:1019969. doi: 10.3389/fphy.2022.1019969

Modeling hierarchical attention interaction between contexts and triple-channel encoding networks for document-grounded dialog generation

Yuanyuan Cai[1,2], Min Zuo[1,2] and Haitao Xiong[1,3]*

[1]National Engineering Research Centre for Agri-Product Quality Traceability, Beijing Technology and Business University, Beijing, China, [2]School of E-Business and Logistics, Beijing Technology and Business University, Beijing, China, [3]School of International Economics and Management, Beijing Technology and Business University, Beijing, China

Dialog systems have attracted attention as they are promising in many intelligent applications. Generating fluent and informative responses is of critical importance for dialog systems. Some recent studies introduce documents as extra knowledge to improve the performance of dialog generation. However, it is hard to understand the unstructured document and extract crucial information related to dialog history and current utterance. This leads to uninformative and inflexible responses in existing studies. To address this issue, we propose a generative model of a neural network with an attention mechanism for document-grounded multi-turn dialog. This model encodes the context of utterances that contains the given document, dialog history, and the last utterance into distributed representations *via* a triple-channel. Then, it introduces a hierarchical attention interaction between dialog contexts and previously generated utterances into the decoder for generating an appropriate response. We compare our model with various baselines on dataset CMU_DoG in terms of the evaluation criteria. The experimental results demonstrate the state-of-the-art performance of our model as compared to previous studies. Furthermore, the results of ablation experiments show the effectiveness of the hierarchical attention interaction and the triple channel for encoding. We also conduct human judgment to evaluate the informativeness of responses and the consistency of responses with dialog history.

KEYWORDS

document-grounded conversation generation, hierarchical attention interaction, semantic feature encoding, context-aware, transformer, encoder–decoder framework

1 Introduction

A dialog system, or conversational agent, is a computer system intended to communicate with human beings intelligently. Dialog systems have wide applications in various domains, such as open-domain chatbots and task-oriented online services [1,2]. For example, patients can quickly obtain an effective diagnosis with the assistance of the automatic medical consultation system. In terms of the way of response acquisition, traditional dialog systems are divided into retrieval-based and generative systems. The former needs to select appropriate responses from a set of candidate facts to match user requests [3], while the latter directly generates more free responses according to the questions.

In general, the task of dialog response generation (DRG) faces more challenges than answer selection, as it is difficult to generate diverse and informative responses. Thus, recent works try to introduce background knowledge such as explicit commonsense [4], keywords, or knowledge graphs [5] into dialog response generation for improving response quality [6].

Document-grounded dialog generation is a typical task in knowledge-based conversations [7]. It expects to use a great amount of relevant information learned from the given unstructured document(s) to limit the range of the output responses [8]. It commonly consists of multi-turn questions and answering, in which the history of dialog can also be used for generating responses constantly. Table 1 shows an example of multi-turn dialog in the openly available document-grounded dialog dataset CMU_DoG [9]. As shown in Table 1, document-grounded dialog generation expects smooth interlacing between multi-turn task-oriented QA (with underline) and casual chat (with underline and italics) taking a document as the background. In particular, this task requires two speakers chatting surrounding a special topic while taking many turns. The generated replies in document-grounded dialog are freer than QA, which may be a safe sentence such as "I do not know" or "I think so" for unanswerable questions, and sometimes maybe even a rhetorical question or new subject of a talk. Some studies consider that document-grounded dialog generation resembles machine reading comprehension (MRC) in the introduction of the unstructured document(s) as supplementary knowledge. However, document-grounded dialog generation comprises multi-turn conversation while the MRC involves only single-turn QA.

The impressive success of deep learning techniques in natural language processing has advanced document-grounded multi-turn dialog. In spite of the explicit knowledge resources that play an important role in generating responses, neural network models have proved effective in generative dialog systems for their strong capabilities to understand a conversation and generate fluent responses [9,10]. Neural network-based response generation does not rely on a specific answer database or template but carries out dialog generation according to the language understanding ability acquired from a large number of corpora. Most of the studies on dialog generation utilize hierarchical encoder–decoder [11,12] or Seq2Seq architecture [13] based on deep neural networks. Chen et al. proposed a dialog pre-training framework named DialogVED, which introduces continuous latent variables into the enhanced encoder–decoder framework to increase the relevance and diversity of responses [1]. These existing research works mainly use two representation methods of dialog utterances. One is treating dialog history and current dialog as an integrated sequence fed into a single encoder [7,9], while the other is recurrently encoding multi-turn utterances by hierarchical encoders, such as the RNN (recurrent neural network) and transformer [1]. Although existing works made some exploration in the conversation structure, they ignore the features of human participating in a multi-turn conversation. As shown in Table 1, a speaker may give multiple utterances

TABLE 1 Example in CMU_DoG.

Movie	**Jaws**
Document	A beach party takes place at dusk on Amity Island, where a woman named Chrissie Watkins goes skinny dipping in the ocean. She is violently pulled under. The next day, her partial remains are found on the shore. The medical examiner rules the death a shark attack which leads to Police Chief Martin Brody closing the beaches. Mayor Larry Vaughn overrules him, fearing it will ruin the town's summer economy. The coroner now concurs with the mayor's theory that Watkins was killed in a boating accident. Brody reluctantly accepts their conclusion until another fatal shark attack occurs shortly after. . .
Conversation	u1: What is it about?
	u2: I meant to say Jaws lol.
	u1: I think I have seen it at some point. Is it a scary movie?
	u2: It is a movie made in the late 70s where people are been hunted by a large shark
	u2: Well, more like action; seems people are always being attacked by the shark
	u1: I have definitely seen it. I always think about it when I am at the beach
	u2: *Really? Wow, it is kind of scary if you are in the deep*
	. . .

constantly in some scenarios. Moreover, different speakers have different backgrounds and emotional attitudes. Thus, we consider that it is unreasonable to predict the response without specifying the speaker.

The existing generative models can be divided into parallel models [14] and incremental models [15], according to the way of representing the given document and historical dialog. The incremental one preserves the temporal relationship of utterances and encodes each historical utterance in turn for response generation. Inspired by the study of incremental response generation, this work models a triple-channel encoder with self-attention networks and an incremental decoder with hierarchical attention interaction between the context of dialog for document-grounded multi-turn response generation. The proposed model uses a transformer framework [16] for the encoder and decoder, which depends entirely on an attention mechanism. The attention mechanism relates different positions of a single sequence to learn and optimize the representation of this sequence [16]. We take the given document, dialog history, and current utterance as significant context clues to create responses. Then we model the hierarchical attention interaction between contexts with self-attention under an encoder–decoder framework for document-grounded multi-turn dialog generation. The contributions of this work are as follows:

- *We propose a novel generative model within the encoder–decoder framework, which introduces a triple-channel encoder to capture global attention between documents, dialog history, and the last utterance for document-driven conversation.*
- *The triple-channel in the encoder is used to learn the distributed representation of conversational context synchronously and then integrate them with the self-attention mechanism.*
- *A hierarchical attention interaction structure is built in the decoder, which can preserve the temporal relationship and the coherence of contexts for response generation. On the basis of this structure, context knowledge is integrated within layer-wise attention to increase the relevance and diversity of next-turn responses.*

The rest of this study is organized as follows. The next section reviews related work. Our proposed model is introduced in Section 3, then the experiments and analyses are presented in Section 4, followed by the conclusion in Section 5.

2 Related work

Traditional dialog systems consist of single-turn conversations [17,18] and multi-turn conversations [12,19] according to the independence of inquiries. In multi-turn dialogs, different inquiries have certain correlations; thus, not only the current quire but also previous utterances are used as model inputs for predicting the next utterance. Wu et al. proposed a multi-hop attention with a termination mechanism in the generative neural network for multi-turn reasoning [20]. However, these previous studies only take utterances as model input regardless of the external knowledge.

In terms of the purpose of conversation, dialog systems can be also divided into task-oriented, QA, conversational recommendation, and chit-chat [21]. Task-oriented systems aim to accomplish users' goals (e.g., online shopping or restaurant reservation) through an optimal decision-making process in multi-turn dialogs. QA systems commonly retrieve useful information from background knowledge to answer single-turn user requests. Conversational recommendation systems offer a useful service or a good product without receiving explicit commands [22,23], while chit-chat systems are designed to meet the users' emotional and social needs. However, most authentic dialog scenarios consist of more than one purpose, which makes recent studies introduce external unstructured or structured knowledge [24] to generate more informative responses for various conversational purposes. Some studies [25] have proved that it is more reasonable to divide dialog systems by background knowledge to reflect the dialog tasks and datasets. Lian et al. used external knowledge to build an end-to-end neural network for single-turn dialog generation [26]. Ghazvininejad et al. extended an end-to-end neural network with real-world facts or conversation history for single-turn conversation generation [27]. Some other studies focus on multi-turn dialog with capturing useful information from the given knowledge [28].

A document-grounded dialog system is one of the representative knowledge-based systems, which uses relevant information derived from the given unstructured text to obtain (generate) responses. Generally, document-grounded dialog systems include reading comprehension in the form of QA [29] and multi-turn dialog in the form of chit-chat. Li et al. proposed a transformer-based architecture (incremental transformer) with a two-way deliberation decoder to encode utterances along with textual knowledge for multi-turn document-grounded conversations [15]. Following their works, Li et al. designed a D3G model with a doc-reader mechanism to locate the information related to the user's questions in a given document [7].

For conversational response generation, sequence-to-sequence or encoder–decoder frameworks with sequential neural networks are widely used to construct generative models for safe and ordinary responses. Qin et al. employed an RNN with memory as the decoder [30], while Shang et al. combined global and local context information in the RNN for a one-round conversation [17]. Vinyals and Quoc explored the LSTM (long short-term memory) network to produce sequential multi-turn conversations [18]. Li et al. improved the LSTM-based generator by using maximum mutual information as the

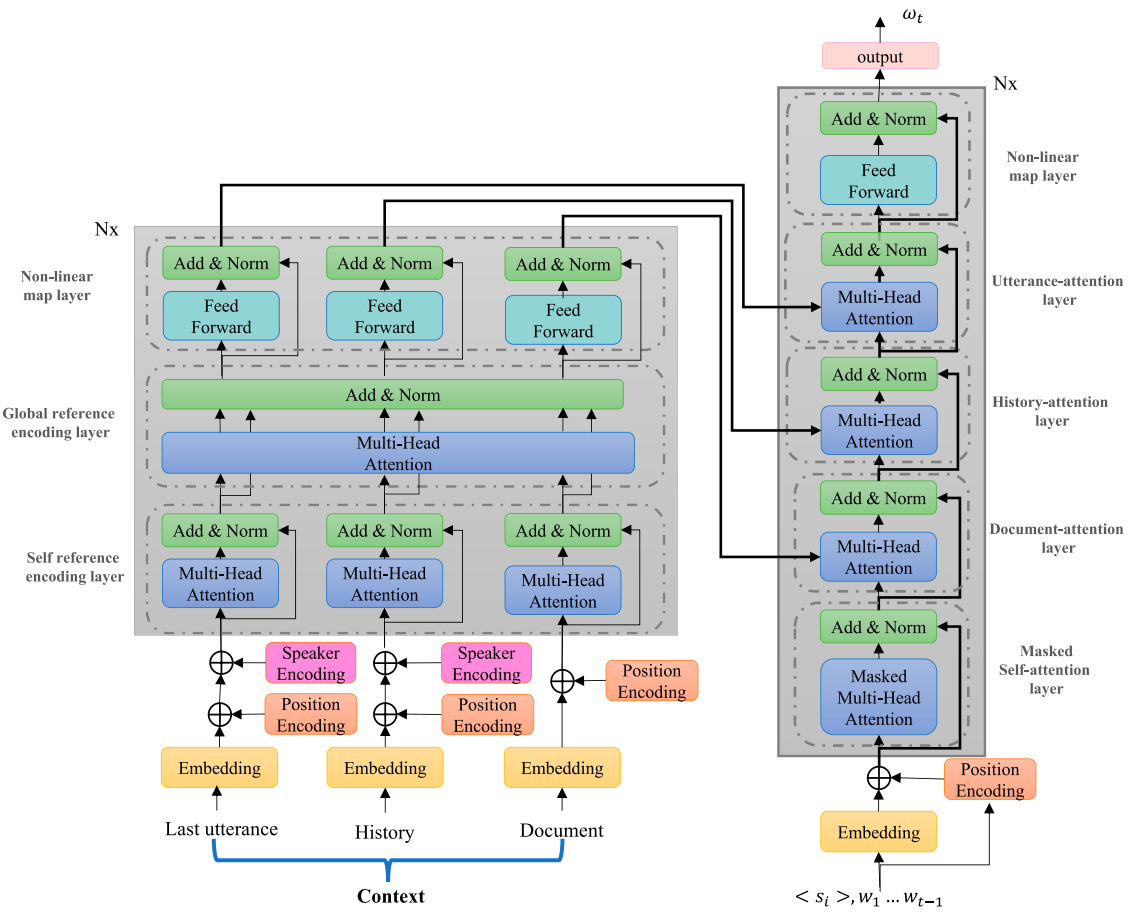

FIGURE 1
Framework of the model. The left is the context encoder with the triple channel. The right is the decoder interacted with contextual encoding for response generation.

objective function [19]. Tang et al. used a one-layer transformer as the decoder [14]. To capture the textual dependencies and key information in the sequences, attention mechanisms [16] have been used to enhance the neural networks of dialog generation and improve response quality. Andrea et al. used a multi-hop attention mechanism over memories with pointer networks to effectively incorporate external knowledge into dialog generation [31]. To emphasize the correlation between contexts, Wu et al. also proposed a multi-hop attention mechanism to learn a single-context vector by computing attention scores [20]. Xing et al. combined utterance-level attention with word-level attention in a neural network to draw the important parts for generating response [32].

3 Methodology

In this work, we propose a novel transformer-based model for document-grounded dialog, and the overall structure of the model is shown in Figure 1. It follows the encoder–decoder framework and consists of three parts: 1) a multi-view embedding module that concatenates position features, speaker features, and word-level semantic features together as input; 2) a context encoding module that learns the semantic information of the document and dialog and captures their information interaction with the triple channel; 3) a hierarchical decoding module that generates context-aware response according to the interaction with outputs of the encoder.

3.1 Problem definition

Given 1) a document D that provides the knowledge associated with the dialog, 2) the corresponding dialog history H presented as a sequence of utterances, and 3) the current (last) utterance u_{k+1}, the motivation of this work is to simulate human reading comprehension to generate a proper response utterance R and keep the conversation going.

Assume the document is a sequence of tokens with length m, which is denoted as $D = \{w_1^d, w_2^d, \ldots, w_m^d\}$. A dialog history is generally treated as a sequence of k utterances $\{u_1, u_2, \ldots, u_k\}$, and the current utterance is naturally denoted as u_{k+1}. Each utterance in the dialog $u_i = \{w_{i,1}^u, w_{i,2}^u, \ldots, w_{i,l_i}^u\}$ is a token-level sequence with length l_i. The dialog history is formalized as

$$H = \left\{w_{1,1}^u, w_{1,2}^u, \ldots, w_{1,l_1}^u, \ldots, w_{k,1}^u, w_{k,2}^u, \ldots, w_{k,l_k}^u\right\}, \quad (1)$$

where w_{k,l_k}^u denotes the kth utterance in the dialog history which has l_k tokens in total. For convenience, we further use $l = \sum_{i=1}^{k} l_i$ to denote the total length of the dialog history sequence. The generated response of our model is a sequence $R = \{w_1, w_2, \ldots, w_T\}$, where T is the length of the response sequence.

Overall, the goal of our model is to generate a reasonable reply by maximizing the conditional probability P. For all possible responses $\{R_j\}$, it can be formalized as

$$R = \arg\max_{R_j} P\left(R_j | D, H, u_{k+1}, \Theta\right), \quad (2)$$

where Θ indicates all trainable parameters in the generative model.

3.2 Multi-view embedding module

This module aims to map the symbolic representations of the given document, dialog history, and current utterance (input query) to distribution representations. In this module, each token is represented with three kinds of embedding features. As shown on the left of Figure 1, the features are listed as follows:

1) **Token embedding**: token embedding is learned to capture the lexical semantics in numerical form. According to the distributed hypothesis, the semantic dependency between words can be efficiently calculated in low-dimensional vector space, i.e., similar words have a closer distance in vector space. In our work, we use matrix $E_w \in R^{(|V|+4)\times d_e}$ to denote the word embedding of the sequence, where $|V|$ is the vocabulary size and d_e is the dimension of the embedding. The first four lines of E_w represent some special tokens: $[PAD]$ for padding the sequence to a fixed length, $[UNK]$ for representing the words out of the vocabulary, $[GO]$ is the start flag of a sequence, and $[EOS]$ is the end flag of a sequence. The rest lines in E_w are the semantic embeddings of words in the vocabulary.
2) **Position embedding**: it is obvious that the temporal information of utterance is important for semantic encoding. Existing recurrent neural networks model the temporal information with its recurrent structure. However, there is a lack of the ability to implicitly modeling the sequence information in the self-attention mechanism and feed-forward networks. Therefore, following the transformer model [16], we introduce an additional position embedding mechanism to supply the required position information for each token in the input sequence. The vector representation for each position pos is defined as

$$PE(pos, d) = \begin{cases} sin\left(pos/10000^{d/d_e}\right) & \text{if } d \text{ is even} \\ cos\left(pos/10000^{(d-1)/d_e}\right) & \text{otherwise} \end{cases}, \quad (3)$$

where d is the dth dimension of the representation.

3) **Speaker embedding**: in general, one dialog consists of multiple utterances from at least two speakers who may have different attitudes toward the same question. It is unreasonable for us to model all utterances as equals, and we consider that the same token in the historical sequence H should have different representations if derived from different speakers. Therefore, we introduce a speaker embedding for integrating the speaker feature into the token representations. In particular, we use $E_s \in R^{n\times d_e}$ to specify the speaker information, where n is the total number of speakers, and each line in E_s represents a speaker.

The synthetic embedding representations of the input sequences are obtained by the aforementioned three kinds of embedding. For historical dialog H and current utterance u_{k+1}, we sum all three kinds of embedding features mentioned previously as their representations, which are denoted as $H_0^h \in R^{l\times d_e}$ and $H_0^u \in R^{l_{k+1}\times d_e}$, respectively. For document D, we only sum the token embedding and position embedding as its representation, which is denoted using $H_0^d \in R^{m\times d_e}$.

3.3 Context encoding module

We build an encoder with a triple channel in the context encoding module of our model. The encoder synchronously learns the document, history, and current utterance and captures the interaction information between them. Previous works generally integrate the document and history information into the hidden states of the current utterance and then feed the fused representations into the decoder. In this work, we retain the representations of all inputs and implement global reference encoding on them.

The encoder is stacked by N blocks with the same structure. The inputs of the first block are the utterance embedding H_0^u, history embedding H_0^h, and document embedding H_0^d. The outputs of it are three matrices H_1^u, H_1^h, and H_1^d, which are the encoded representations of the utterance, history, and document, respectively. For the nth block ($n > 1$), the inputs of it are the outputs of the previous block, while the outputs are also three matrices H_n^u, H_n^h, and H_n^d.

As shown in the left section of Figure 1, each block of the encoder consists of three layers. The first layer is the "self-reference encoding." The second is the "global reference encoding" layer. The last one is the "non-linear map." Following the transformer, we apply an "Add & Norm" operation in each layer, and for convenience, this operation is omitted in the study.

The "self-reference encoding" layer is proposed to learn the independent representation of the utterance, history, and document. It is implemented by a multi-head attention function [16], which is defined as *MultiHead* (·, ·, ·) and the inputs of the function are the query, key, and value, respectively[1].

For *n*th block, the self-reference encoding layer is defined as

$$\begin{aligned}\bar{H}_n^u &= MultiHead\left(H_{n-1}^u, H_{n-1}^u, H_{n-1}^u\right)\\ \bar{H}_n^h &= MultiHead\left(H_{n-1}^h, H_{n-1}^h, H_{n-1}^h\right)\\ \bar{H}_n^d &= MultiHead\left(H_{n-1}^d, H_{n-1}^d, H_{n-1}^d\right)\end{aligned}. \tag{4}$$

We concatenate the outputs of the "self-reference encoding" layer,

$$H_n = \left[\bar{H}_n^u; \bar{H}_n^h; \bar{H}_n^d\right], \tag{5}$$

and then feed them into the "global reference encoding" layer.

$$\left[\tilde{H}_n^u; \tilde{H}_n^h; \tilde{H}_n^d\right] = MultiHead\left(H_n, H_n, H_n\right). \tag{6}$$

Through the aforementioned operations, the model could capture the interaction information of the utterance, history, and document.

Furthermore, we feed the outputs of the global reference encoding layer into the "non-linear map" layer to obtain the outputs of the current block. The non-linear map layer is implemented by a position-wise feed-forward network (FFN) with two layers, which is defined as

$$FFN\left(x\right) = \sigma\left(xW_1 + b_1\right)W_2 + b_2, \tag{7}$$

where σ is a ReLU function, $W_1 \in R^{d_m \times d_{in}}$, $b_1 \in R^{d_{in}}$, $W_2 \in R^{d_{in} \times d_m}$, and $b_2 \in R^{d_e}$ are trainable parameters, d_m is the hidden size of the model, and d_{in} is the inner size of the FFN. The formalization of "non-linear map" layer is, therefore, defined as

$$\begin{aligned}H_n^u &= FFN\left(\tilde{H}_n^u\right)\\ H_n^h &= FFN\left(\tilde{H}_n^h\right)\\ H_n^d &= FFN\left(\tilde{H}_n^d\right)\end{aligned}. \tag{8}$$

Finally, we define the outputs of the last block H_N^u, H_N^h, and H_N^d as the outputs of the context encoding module.

1 Since multi-head attention is a general function, we will not introduce its implementation in this study.

3.4 Hierarchical decoding module

This module is built for generating responses according to the dialog context as shown in the right section of Figure 1. We consider that people normally read through a given document and dialog history first in reading comprehension to learn and pay close attention to the required evidence, relevant to the current utterance. Thus, we designed a hierarchical information interaction structure with a multi-head attention mechanism for the conversation context and the generated replies. The three-layer structure consists of the document-attention layer, history-attention layer, and utterance-attention layer. In this module, the context representations learned from the triple-channel encoder are integrated with the previously generated responses to predict the next target reply.

As shown in Figure 1, a decoder that has N stacks and contains five layers per stack is built in this module. The response is generated token by token, that is, we first produce the first token with a start flag [*GO*], then we concatenate the start flag and the generated first token to produce the second token, and so on, until an end flag [*EOS*] is produced. At time step t, the start flag, the previous $t-1$ tokens $w_1, w_2, \ldots, w_{t-1}$, and the output of the encoder are fed into the decoder to predict the tth token in the response.

First, we convert the start flag and the generated tokens to distributed representations by the multi-view embedding module. The representation for the start flag and previous $t-1$ tokens is defined as $H_0^{r_{t-1}}$.

After that, $H_0^{r_{t-1}}$ is fed into the stacked blocks. For the nth block, we feed the inputs (embedding representations or outputs of the previous block) into the masked self-attention layer to learn the information from the generated responses. This layer is implemented by a masked multi-head attention mechanism [16], which is defined as *MaskedMultiHead* (·, ·, ·) and the inputs of it are the same as the *MultiHead* function.

$$\bar{H}_n^{r_{t-1}} = MaskedMultiHead\left(H_{n-1}^{r_{t-1}}, H_{n-1}^{r_{t-1}}, H_{n-1}^{r_{t-1}}\right). \tag{9}$$

Then, we pass the encoded responses through the document-attention layer, history-attention layer, and utterance-attention layer sequentially.

$$\begin{aligned}\tilde{H}_n^{r_{t-1}} &= MultiHead\left(\bar{H}_n^{r_{t-1}}, H_N^d, H_N^d\right)\\ \dot{H}_n^{r_{t-1}} &= MultiHead\left(\tilde{H}_n^{r_{t-1}}, H_N^h, H_N^h\right)\\ \ddot{H}_n^{r_{t-1}} &= MultiHead\left(\dot{H}_n^{r_{t-1}}, H_N^u, H_N^u\right)\end{aligned}. \tag{10}$$

Next, $\ddot{H}_n^r$ is fed into a non-linear map layer, which is the same as in the context encoding module.

$$H_n^{r_{t-1}} = FFN\left(\ddot{H}_n^{r_{t-1}}\right). \tag{11}$$

Finally, we use the outputs of the Nth block to predict the generated token at the tth time step. The generated token is selected from the pre-defined vocabulary with the highest probability, and the distributed probability is defined as

TABLE 2 Statistics of the CMU_DoG datasets.

Dataset	#Conversation	#Utterance	Average utterances per conversation	Average tokens per utterance
Train	3,373	107,792	31.96	11.03
Valid	229	7,030	30.70	12.22
Test	619	19,375	31.30	10.94
Total	4,221	134,197	31.79	11.08

$$p_t = softmax\left(H_n^{r_{t-1}}(t) \cdot E_w^T\right), \tag{12}$$

where $H_n^{r_{t-1}}(t)$ is the tth row of the output matrix $H_n^{r_{t-1}}$, indicating the output vector corresponding to the last input token, and E_w is the matrix representation of the pre-defined vocabulary.

During training, we use the negative log-likelihood of the target word as the loss at each time step, and then the final loss is defined as the summation of the losses of all time steps.

4 Experiments and result analysis

4.1 Dataset

We conduct the experiments of dialog response generation on dataset CMU_DoG which has a total of 4,221 conversations with an average of 31.79 utterances per conversation. The statistics of the dataset are shown in Table 2. This dataset presents a set of movie-themed text descriptions and their corresponding multi-turn dialogs. The extra textual descriptions contain movie names, introductions, ratings, and some other scenes. They present enough dialog-related information that may help generate context-specific and informative responses for a multi-turn conversation. The average length of the documents is approximately 200 words.

Each dialog in CMU_DoG involves two participants, and the given document is accessible to only one participant or both participants. Moreover, the dataset also provides the correspondence between each utterance and the paragraphs of the document.

4.2 Experiment settings

In our experiments, the stacks of both the encoder and decoder are set to 4. The number of attention heads is set to 8. The embedding size d_e and the hidden size d_m are set to 512, and the inner size of the FFN d_{in} is set to 2,048. We use the Adam algorithm [33] with a learning rate of 0.0001 for optimization. The batch size is set to 64, and the dropout rate is set to 0.1. In addition, we train the model for 50 epochs. To save space and training time, we take the previous three utterances as the dialog history in our experiments.

4.3 Evaluation criteria

4.3.1 Automatic evaluation

We evaluate the proposed document-driven generative model in terms of diverse metrics, including BLEU-n [34], METEOR [35], Doc_BLEU, perplexity [36], and the average length of the generated responses.

BLEU-n: it is known to correlate reasonably well with human beings on the evaluation of conversation generation. It measures n − Gram overlap between generated responses and the ground truth, which is defined as BLEU−n. We leverage multiple BLEU−n scores to evaluate the performance of generative models.

METEOR: it is also a common metric to measure the relevance between generated responses and ground truth. Compared with BLEU−n scores, METEOR pays more attention to the recall rate and applies a more generalized concept of a unigram matching method that the unigram can be matched based on their surface forms, stemmed forms, and meanings.

Doc_BLEU: we use this criterion to evaluate the relevance of generated responses with the given documents. It measures the unigram overlap between responses and documents, which is calculated by the original formula of BLEU [34] without the brevity penalty factor.

Perplexity (PPL): this indicator is used to evaluate the fluency of the response. A lower perplexity indicates the better performance of the models and higher quality of the generated sentences.

Avg_Len: generally speaking, longer sentences supply richer information. Therefore, we use the average length of generated utterances as an automatic criterion to evaluate the quality of the generated responses.

TABLE 3 BLEU–*n* scores and METEOR scores for baselines and the proposed model. The "-speaker embedding" indicates the speaker embedding is removed from the multi-view embedding module of our model.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
SEQ [37]	6.12	1.52	0.59	0.30	4.18
SEQS [37]	6.57	1.65	0.67	0.35	4.30
D3G [7]	6.32	1.71	0.71	0.41	4.17
Transformer [16]	8.55	2.49	1.12	0.60	4.53
ITDD [14]	-	-	-	0.95	-
ITDD (our impl)	8.19	2.88	1.66	0.85	5.21
BCTCE [9]	9.98	3.56	2.05	1.42	5.22
Our model	**11.24**	**4.27**	2.54	1.80	5.83
-speaker embedding	10.59	3.87	**2.66**	**2.22**	**5.86**

4.3.2 Human judgment

In addition to the aforementioned quantitative evaluation, we also augment the manual evaluation in terms of fluency and coherence of dialog and the diversity of responses. These evaluation metrics are scored 0/1/2. We randomly sample multiple conversations containing 822 utterances from the test set. We ask multiple annotators to score these utterances given their previous utterances and related documents. The final score of each utterance is the average of the scores rated by three annotators.

Fluency: it is used to indicate whether the response is natural and fluent. Score 0 shows the response is not fluent and incomprehensible; 1 shows the response is partially fluent but still comprehensible; 2 shows the response is sufficiently fluent.

Coherence: it is used to evaluate whether the response is logically coherent with the dialog. Score 0 shows the response is irrelevant with previous utterances; 1 shows the response matches the topic of previous utterances; 2 shows the response is exactly coherent with previous utterances.

Diversity: it is used to reflect the lexical diversity of the response. Score 0 represents the safe response which is applicable to almost all conversations, e.g., "I agree with you;" 1 represents the response suitable to limited conversations but plain and uninformative; 2 shows the response is evidently vivid and informative.

4.4 Baselines and result discussion

We choose several studies on document-grounded conversational generation as baselines, which contain the transformer [16], ITDD [15], D3G [7], BCTCE [9], SEQ, and SEQS [37] models. The transformer, ITDD, BCTCE, and the proposed model in this work depend entirely on attention mechanisms, while others utilize sequential neural networks as encoders and decoders. The transformer takes the dialog context as a sequence of tokens and inputs it into an encoder–decoder framework without distinction. The ITDD incrementally encodes multi-turn utterances along with the knowledge in related documents and applies a two-pass decoder that focuses on context coherence and knowledge correctness to generate responses. The BCTCE proposes a binary-channel structure for context encoding. However, our proposed model in this work uses a triple-channel structure to encode the dialog contexts in parallel and employs an incremental decoder to capture the semantic dependency of contexts in a hierarchical network for response generation.

TABLE 4 Document relevance and response quality for baselines and the proposed model. The "-speaker embedding" indicates the speaker embedding is removed from the multi-view embedding module of our model.

Models	Doc_BLEU	Avg_Len	PPL
SEQ [37]	24.88	7.31	15.62
SEQS [37]	27.96	7.21	19.53
D3G [7]	26.76	6.83	18.40
Transformer [16]	27.55	7.91	13.70
ITDD [14]	-	-	15.10
ITDD (our impl)	26.96	8.52	**11.01**
BCTCE [9]	28.23	9.16	17.80
Our model	**29.10**	10.22	20.06
-speaker embedding	26.42	**11.10**	24.68

The results in Tables 3, 4 show the comparison of our model with baseline models on the dataset CMU_DoG. It can be seen that our model significantly outperforms all baselines according to the automatic evaluation criteria except PPL and achieves state-of-the-art performance. The new state-of-the-art

TABLE 5 Human evaluation of baselines and our proposed model. The "-speaker embedding" means the speaker embedding is removed from the multi-view embedding module in our model.

Models	Fluency	Dialog coherence	Diversity
SEQ [37]	1.27	0.81	0.42
SEQS [37]	1.13	0.96	0.71
D3G [7]	1.29	1.12	0.84
Transformer [16]	1.34	1.17	0.90
ITDD (our impl)	1.35	1.27	0.93
BCTCE [9]	1.34	1.29	0.95
Our model	**1.42**	**1.35**	**0.97**
-speaker embedding	1.38	1.32	0.94

performance achieved by our model indicates that the responses generated by our model correspond more with the ground truth. However, our original model gets lower scores in terms of BLEU−3, BLEU−4, and METEOR than the one without speaker embedding. In spite of its worse performance, the effectiveness of speaker information could be demonstrated by the following case study.

Moreover, we also conducted some human judgment to evaluate our model and the results are shown in Table 5. It shows that the responses generated by our model are more relevant to the context (document and dialog history) with better diversity, compared to baselines.

According to Tables 4, 5, our model has better fluency but worse PPL scores than baselines. Both PPL and fluency are used to measure whether the generated response is natural and fluent. However, PPL is hard to accurately evaluate the fluency of the response. Some research studies find that a sentence with a low PPL would not be in accord with natural language and an informative sentence usually has a higher PPL than a common sentence [38]. This fact indicates that the models with a low PPL tend to produce more generic responses. Thus, our model would generate fluent responses in spite of its higher PPL.

4.5 Ablation study

To validate the effectiveness of each layer for contextual attention interaction in the hierarchical decoder, we also conduct ablation experiments on the dataset CMU_DoG. First, we change the decoding order of the context. Some instances are explained as follows.

1) **Document=>utterance=>history:** the last utterance is exchanged with the dialog history upon our model, which means that the attention interaction between previously generated responses with it is conducted before with the dialog history.
2) **History=>document=>utterance:** we feed the embedding of the dialog history, document, and the last utterance into the decoder for the attention interaction by turn.
3) **Parallel decoder:** we replace the incremental decoder in our model with a parallel decoder. The concatenate of embedding of the dialog history, document, and the last utterance is fed into the decoder for attention interactions with previously generated responses.

The results in Tables 6, 7 show that the original model outperforms the ablation models in terms of most metrics, which indicates the decoding order used in our model is more effective. The reason is that the order is in more accordance with human custom. Human beings commonly read through the given documents to acquire background knowledge before finding out the core information of dialog by understanding historical utterances and then focus on the current utterance to give its response.

In addition, the results can also show the effectiveness of hierarchical attention interaction with a reasonable order for utterance decoding, despite of the concatenation of the context within parallel attention interaction achieving the highest BLEU−4 score.

Furthermore, we remove an attention layer from the decoder in the second ablation experiment. The results given in Tables 8, 9 also show the effectiveness of all attention interactions in the decoder. Compared to other variants, our model has better performance in terms of BLEU−*n*, METEOR, PPL, and Doc_BLEU.

4.6 Case study

To demonstrate the role of speaker embedding, we analyze two conversations in CMU_DoG shown in Table 10. In the first case, our model generates a proper response for the speaker "u2" to reply to the question from the speaker "u1," while the model without speaker embedding produces a repeated question from the speaker "u2" in the last dialog turn. In the second case, the speaker "u1" gives two continuous utterances, where the first one is a question and the second one answers the previous question from the speaker "u2." Our model generates a more proper response for the speaker "u2" to reply to the first utterance of the speaker "u1," compared to the model without speaker embedding. The case study shows that speaker embedding is an effective component for our model to identify the speaker of each utterance and improve the consistency of generated responses in a multi-turn dialog.

TABLE 6 Comparison of various decoding orders of the context on the BLEU–*n* score and METEOR. The symbol " =>" represents the decoding order from the bottom up.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Document=>history=>utterance(Ours)	**11.24**	**4.27**	**2.54**	1.80	**5.83**
Document=>utterance=>history	10.12	3.45	2.21	1.73	5.45
History=>document=>utterance	10.23	3.42	2.11	1.68	5.62
History=>utterance=>document	10.34	3.41	2.14	1.65	5.60
Utterance=>history=>document	9.96	3.40	2.16	1.72	5.44
Utterance=>document=>history	10.20	3.42	2.18	1.70	5.55
Parallel decoder	10.43	3.67	2.37	**1.91**	5.69

TABLE 7 Comparison of various decoding orders of the context on document relevance and response quality. The symbol " =>" represents the decoding order from the bottom up.

Model	Doc_BLEU	Avg_Len	PPL
Document=>history=>utterance(Ours)	**29.10**	10.22	**20.06**
Document=>utterance=>history	27.29	10.44	23.07
History=>document=>utterance	28.08	**11.41**	23.78
History=>utterance=>document	27.53	11.31	22.40
Utterance=>history=>document	26.90	10.41	22.73
Utterance=>document=>history	28.48	10.97	24.01
Parallel decoder	27.47	11.07	23.51

TABLE 8 Ablation study for the hierarchical decoding module on BLEU–*n* scores and METEOR. The symbol " =>" represents the decoding order from the bottom up.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Document=>history=>utterance(Ours)	**11.24**	**4.27**	**2.54**	**1.80**	**5.83**
Document=>history	10.35	3.43	2.11	1.60	5.63
Document=>utterance	10.20	3.34	2.04	1.58	5.47
History=>utterance	10.19	3.41	2.09	1.64	5.49
History=>document	9.71	3.29	2.07	1.62	5.49
Utterance=>history	10.41	3.53	2.23	1.77	5.61
Utterance=>document	10.06	3.39	2.09	1.62	5.22

TABLE 9 Ablation study for the hierarchical decoding module on document relevance and response quality. The symbol " =>" represents the decoding order from the bottom up.

Model	Doc_BLEU	Avg_Len	PPL
Document=>history=>utterance(Ours)	**29.10**	10.22	**20.06**
Document=>history	27.27	**11.54**	21.86
Document=>utterance	27.88	10.71	21.50
History=>utterance	27.35	10.84	22.79
History=>document	26.33	9.75	20.60
Utterance=>history	27.34	10.94	21.65
Utterance=>document	26.62	10.84	21.76

TABLE 10 Sample responses of model variants.

Document	Dialog	Responses
Home Alone is a 1990 American comedy film written and produced by John Hughes and directed by Chris Columbus. The film stars Macaulay Culkin as Kevin McCallister, a boy who is mistakenly left behind when his family flies to Paris for their Christmas vacation. Kevin initially relishes being home alone, but soon has to contend with two would-be burglars played by Joe Pesci and Daniel Stern. The film also features Catherine O'Hara and John Heard as Kevin's parents	**u1:** Hi **u2:** have you seen Home Alone? **u1:** Yes, I love that movie **u2:** [predicted response]	**ground truth:** There is a second one, isn't there? **w/ speaker:** I like the actors in it too **w/o speaker:** Have you seen it before?
La la land is a 2016 American musical romantic comedy-drama film written and directed by Damien Chazelle. It stars Ryan Gosling as a jazz pianist and Emma Stone as an aspiring actress, who meet and fall in love in Los Angeles while pursuing their dreams. The film's title refers simultaneously to music, the city of Los Angeles, and to the idiom for being out of touch with reality	**u2:** Do you like this movie? **u1:** The movie is called La La Land. Have you seen it? **u1:** Yes, I like the movie. I saw it in the theater **u2:** [predicted response]	**ground truth:** I like it. Yes, I have seen **w/ speaker:** I have not seen it yet **w/o speaker:** It is a great movie to see

5 Conclusion

In this study, we present a novel neural model for dialog response generation in document-oriented dialog systems. This model utilizes unstructured document information for response generation to exploit the knowledge hidden in the text, which will be the inevitable trend of the open domain dialog systems. To improve the quality of generated responses, we model a triple-channel encoding and a hierarchical attention interaction between dialog contexts. Comparative experiments are conducted on a public dataset CMU_DoG to evaluate the proposed model. The results show the greater performance of our model than that of several relevant models. The experiment results also confirm the effectiveness of hierarchical attention interactions between contexts with a triple channel for multi-turn dialog generation.

Data availability statement

The dataset used in our experiments is publicly available at https://github.com/festvox/datasets-CMU_DoG.

Author contributions

YC, HX, and MZ contributed to the conception and design of the study. MZ performed the statistical analysis. YC wrote the first draft of the manuscript. HX and MZ wrote sections of the manuscript. All authors agreed to be accountable for the content of the work. All authors contributed to manuscript revision and approved the submitted version.

Funding

This research was supported by the R&D Program of the Beijing Municipal Commission of Education (Grant No. KM202010011011), National Natural Science Foundation of China (Grant Nos. 72171004 and 61873027), Beijing Natural Science Foundation (Grant No. 4202014), and Humanity and Social Science Youth Foundation of the Ministry of Education of China (Grant Nos. 21YJCZH186 and 20YJCZH229).

Acknowledgments

The authors would like to thank the associate editor and the reviewers for their useful feedback that improved this study.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors, and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

1. Chen W, Gong Y, Wang S, Yao B, Qi W, Wei Z, et al. Dialogved: A pre-trained latent variable encoder-decoder model for dialog response generation. In: S Muresan, P Nakov, A Villavicencio, editors. *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2022, dublin, Ireland, may 22-27, 2022*. Stroudsburg, PA: Association for Computational Linguistics (2022). p. 4852–64.

2. Adiwardana D, Luong M-T, So D R, Hall J, Fiedel N, Thoppilan R, et al. *Towards a human-like open-domain chatbot* (2020). doi:10.48550/arXiv.2001.09977

3. Park Y, Ko Y, Seo J. Bert-based response selection in dialog systems using utterance attention mechanisms. In: *Expert systems with applications* (2022).

4. Zhou H, Young T, Huang M, Zhao H, Xu J, Zhu X. Commonsense knowledge aware conversation generation with graph attention. In: *Twenty-seventh international joint conference on artificial intelligence IJCAI-18* (2018).

5. Xu J, Wang H, Niu Z, Wu H, Che W. Knowledge graph grounded goal planning for open-domain conversation generation. *Proc AAAI Conf Artif Intelligence* (2020) 34:9338–45. doi:10.1609/aaai.v34i05.6474

6. Xu Y, Ishii E, Liu Z, Winata G I, Su D, Madotto A, et al. Retrieval-free knowledge-grounded dialog response generation with adapters. In: *Proceedings of the second DialDoc workshop on document-grounded dDialog and conversational question answering*. Dublin, Ireland: Association for Computational Linguistics (2022). p. 93–107. DialDoc@ACL 2022May 26, 2022.

7. Li K, Bai Z, Wang X, Yuan C. A document driven dialog generation model. In: *Chinese computational linguistics - 18th China national conference, CCL 2019*. Kunming, China (2019). p. 508–20. October 18-20, 2019, Proceedings.

8. Gao C, Zhang W, Lam W. Unigdd: A unified generative framework for goal-oriented document-grounded dialog. In: S Muresan, P Nakov, A Villavicencio, editors. *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 2: Short papers), ACL 2022, dublin, Ireland, may 22-27, 2022*. Stroudsburg, PA: Association for Computational Linguistics (2022). p. 599–605.

9. Cai Y, Zuo M, Zhang Q, Xiong H, Li K. A bi-channel transformer with context encoding for document-driven conversation generation in social media. *Complexity* (2020) 1–13. doi:10.1155/2020/3710104

10. Zhang H, Lan Y, Pang L, Guo J, Cheng X. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialog generation. In: *Proceedings of the 57th annual meeting of the association for computational linguistics* (2019).

11. Zeng H, Liu J, Wang M, Wei B. A sequence to sequence model for dialog generation with gated mixture of topics. *Neurocomputing* (2021) 437:282–8. doi:10.1016/j.neucom.2021.01.014

12. Serban I V, Sordoni A, Bengio Y, Courville A, Pineau J. *Building end-to-end dialog systems using generative hierarchical neural network models* (2016). p. 3776–84.

13. Sutskever I, Vinyals O, Le Q V. Sequence to sequence learning with neural networks. In: *NIPS* (2014).

14. Tang X, Hu P. Knowledge-aware self-attention networks for document grounded dialog generation. In: *The 12th international conference on knowledge science, engineering and management* (2019).

15. Li Z, Niu C, Meng F, Feng Y, Li Q, Zhou J. Incremental transformer with deliberation decoder for document grounded conversations. In: *Proceedings of the 57th conference of the association for computational linguistics, ACL 2019*, 1. Florence, Italy (2019). p. 12–21. July 28- August 2, 2019Long Papers.

16. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, et al. Attention is all you need. In: *Advances in neural information processing systems 30: Annual conference on neural information processing systems 2017*. Long Beach, CA, USA (2017). p. 5998–6008. 4-9 December 2017.

17. Shang L, Lu Z, Li H. Neural responding machine for short-text conversation. In: *Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural l anguage processing of the asian federation of natural l anguage processing*, 1. Beijing, China (2015). p. 1577–86. *ACL 2015, July 26-31, 2015*Long Papers.

18. Vinyals O, Le Q. A neural conversational model. *Comp Sci* (2015).

19. Li J, Galley M, Brockett C, Gao J, Dolan B. A diversity-promoting objective function for neural conversation models. *Comp Sci* (2015).

20. Wu X, Martinez A, Klyen M. Dialog generation using multi-turn reasoning neural networks. In: *NAACL-HLT. Association for computational linguistics* (2018). p. 2049–59.

21. Chiu S, Li M, Lin Y, Chen Y. Salesbot: Transitioning from chit-chat to task-oriented dialogs. In: S Muresan, P Nakov, A Villavicencio, editors. *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers), ACL 2022, dublin, Ireland, may 22-27, 2022*. Stroudsburg, PA: Association for Computational Linguistics (2022). p. 6143–58.

22. Xiong F, Wang X, Pan S, Yang H, Zhang C. Social recommendation with evolutionary opinion dynamics. In: *IEEE transactions on systems, man, and cybernetics: Systems* (2018). p. 1–13.

23. Li Z, Xiong F, Wang X, Chen H, Xiong X. Topological influence-aware recommendation on social networks. *Complexity* (2019) 1–12. doi:10.1155/2019/6325654

24. WanyunXiao Y, Wang H, Song Y, won Hwang S, Wang W. Kbqa: Learning question answering over qa corpora and knowledge bases. *Proc VLDB Endowment* (2017) 10:565–76. doi:10.14778/3055540.3055549

25. Ma L, Zhang W, Li M, Liu T. *A survey of document grounded dialog systems (dgds)* (2020). doi:10.48550/arXiv.2004.13818

26. Lian R, Min X, Fan W, Peng J, Hua H. Learning to select knowledge for response generation in dialog systems. In: *Twenty-eighth international joint conference on artificial intelligence IJCAI-19* (2019).

27. Ghazvininejad M, Brockett C, Chang M, Dolan B, Gao J, Yih W, et al. A knowledge-grounded neural conversation model. In: S A McIlraith K Q Weinberger, editors. *Proceedings of the thirty-second AAAI conference on artificial intelligence, (AAAI-18), the 30th innovative applications of artificial intelligence (IAAI-18), and the 8th AAAI symposium on educational advances in artificial intelligence (EAAI-18)*. New Orleans, Louisiana, USA: AAAI Press (2018). p. 5110–7. February 2-7, 2018.

28. Xu F, Zhou S, Wang X, Ma Y, Zhang W, Li Z (2022). Open-domain dialog generation grounded with dynamic multi-form knowledge fusion. doi:10.48550/arXiv:2204.11239v1

29. Lan Y, Jiang J. Modeling transitions of focal entities for conversational knowledge base question answering. In: *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural l anguage processing*, 1 (2021). Long Papers.

30. Qin L, Galley M, Brockett C, Liu X, Gao X, Dolan B, et al. Conversing by reading: Contentful neural conversation with on-demand machine reading. In: *Proceedings of the 57th annual meeting of the association for computational linguistics* (2019).

31. Madotto A, Wu C, Fung P. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In: *Proceedings of the 56th annual meeting of the association for computational linguistics, ACL 2018*, 1. Melbourne, Australia (2018). p. 1468–78. July 15-20, 2018Long Papers.

32. Xing C, Wu Y, Wu W, Huang Y, Zhou M. Hierarchical recurrent attention network for response generation. In: S A McIlraith K Q Weinberger, editors. *Proceedings of the thirty-second AAAI conference on artificial intelligence, new orleans, Louisiana, USA, february 2-7, 2018* (2018). p. 5610–7.

33. Kingma D P, Ba J. Adam: A method for stochastic optimization. In: *3rd international conference on learning RepresentationsICLR*. San Diego, CA, USA (2015). *May 7–9*.

34. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: A method for automatic evaluation of machine translationk. In: Proceedings of the 40th annual meeting on association for computational linguistics; 2002 July 6–12; Philadelphia, PA. Stroudsburg, PA: Association for Computational Linguistics (2002). p. 311–8.

35. Denkowski M, Lavie A. Meteor universal: Language specific translation evaluation for any target language. In: *Proceedings of the ninth workshop on statistical machine translation*. Baltimore, Maryland USA (2014). p. 376–80. *June 26–27*.

36. Dinan E, Roller S, Shuster K, Fan A, Auli M, Weston J. Wizard of wikipedia: Knowledge-powered conversational agents. In: *7th international conference on learning RepresentationsICLR*. New Orleans, LA, USA (2019). *May 6–9*.

37. Zhou K, Prabhumoye S, Black A W. A dataset for document grounded conversations. In: *2018 conference on empirical methods in natural l anguage processing*. Brussels, Belgium (2018). p. 708–13. October 31–November 4.

38. Kuribayashi T, Oseki Y, Ito T, Yoshida R, Asahara M, Inui K. Lower perplexity is not always human-like. In: C Zong, F Xia, W Li, R Navigli, editors. *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural l anguage processing, ACL/IJCNLP 2021, (volume 1: Long papers), virtual event, august 1-6, 2021*. Stroudsburg, PA: Association for Computational Linguistics (2021). p. 5203–17. doi:10.18653/v1/2021.acl-long.405