Hindawi
Complexity
Volume 2020, Article ID 3710104, 13 pages
https://doi.org/10.1155/2020/3710104

Research Article

A Bichannel Transformer with Context Encoding for Document-Driven Conversation Generation in Social Media

Yuanyuan Cai,[1] Min Zuo,[1] Qingchuan Zhang,[1] Haitao Xiong,[1] and Ke Li[2]

[1]*National Engineering Laboratory for Agri-Product Quality Traceability and Beijing Key Laboratory of Big Data Technology for Food Safety, Beijing Technology and Business University, Beijing 100048, China*
[2]*Beijing University of Posts and Telecommunications, No. 10 Xitucheng Road, Haidian District, Beijing, China*

Correspondence should be addressed to Min Zuo; zuomin1234@163.com

Received 7 June 2020; Accepted 30 July 2020; Published 17 September 2020

Guest Editor: Liang Wang

Along with the development of social media on the internet, dialogue systems are becoming more and more intelligent to meet users' needs for communication, emotion, and social intercourse. Previous studies usually use sequence-to-sequence learning with recurrent neural networks for response generation. However, recurrent-based learning models heavily suffer from the problem of long-distance dependencies in sequences. Moreover, some models neglect crucial information in the dialogue contexts, which leads to uninformative and inflexible responses. To address these issues, we present a bichannel transformer with context encoding (BCTCE) for document-driven conversation. This conversational generator consists of a context encoder, an utterance encoder, and a decoder with attention mechanism. The encoders aim to learn the distributed representation of input texts. The multihop attention mechanism is used in BCTCE to capture the interaction between documents and dialogues. We evaluate the proposed BCTCE by both automatic evaluation and human judgment. The experimental results on the dataset CMU_DoG indicate that the proposed model yields significant improvements over the state-of-the-art baselines on most of the evaluation metrics, and the generated responses of BCTCE are more informative and more relevant to dialogues than baselines.

1. Introduction

Dialogue systems such as Siri, Cortana, and Duer have been widely used to facilitate interactions between humans and intelligence devices as virtual assistants and social Chatbots. For example, people can conveniently make airline reservations with the help of an intelligent agent in social media. Conversational response generation, as a challenging task in natural language processing, plays a critical role in conversational systems.

Conversational generation aims to produce grammatical, coherent, and plausible responses in accordance with the input from users. Previous studies on dialogue generation mainly focus on either one-round conversation [1] or multiturn conversation [2]. One-round conversation tasks commonly determine responses on the basis of a single current query, while a multiturn conversation that consists of context-message-response triples commonly builds context-sensitive generators according to the dialogue history [2, 3]. Multiturn conversation tasks tend to generate a variety of correlative responses in either goal-driven customer services [4–6] or chit-chat without predefined goals [7].

In the previous studies, the sequence-to-sequence (seq2seq) framework [8] with the attention mechanism has commonly been used to generate conversational responses and has achieved remarkable success in various domains [9–14]. The seq2seq models map a type of sequential syntactic structure to another without explicitly defining structural features by building an end-to-end neural network [2, 15]. Most seq2Seq models use a recurrent neural network [16, 17] as the encoder and decoder to capture the sequential dependency. However, hierarchical recurrent neural networks, which suffer from time-consuming training, have difficulty in solving the problem of long-distance textual semantic dependency.

Since lacking of a knowledge background, the previous studies on conversational generation may suffer from the safe and generic responses such as "That's all right" and "Yes." The

uninformative responses are hard to match the relative content in the given document and satisfy the demand of users. To address this challenge, some knowledge-based methods have been proposed in recent works for conversational response generation. In these works, external knowledge is leveraged to facilitate conversation understanding and generation [18, 19], which includes structured data such as knowledge graphs [20, 21], unstructured textual knowledge [22], and visual knowledge [23]. With the development of the internet and big data, unstructured knowledge is more accessible than structured knowledge, which is constructed manually and depends heavily on the experience of experts. Therefore, some recent works take conversation-related documents and texts as the background knowledge to enrich useful information in conversations to generate more informative and interesting responses [24].

Our work is inspired by the recent success of the transformer framework [25], which is entirely based on attention mechanisms in end-to-end natural language processing tasks and eliminates complex recurrent and convolution network architectures [26]. We propose a transformer-based model for multiturn document-driven conversation. The proposed model encodes the conversational context and the current utterance, respectively. It also incorporates the multihop attention mechanism into the encoder and decoder to capture the correlative content for response generation, which draws global dependencies between documents, utterances, and responses. We conduct experiments on the conversation generation task regarding many metrics, including BLEU [27], METEOR [28], NW [24], and perplexity [19]. The experimental results indicate that the proposed model significantly outperforms the state-of-the-art methods. We also conduct ablation experiments to indicate the effects of the input elements fed into the encoder. The human judgment on various ablation models shows that the responses generated by the BCTCE model are more relevant to the context (document and dialogue history), more informative, and fluent than its several variants.

The contributions of this work are as follows:

We propose a novel BCTCE based on the transformer framework to build an encoder-decoder generator for document-driven conversation. The experimental results show that our model achieves new state-of-the-art performance.

The BCTCE learns the distributed representation of conversational context by encoding the document and dialogue utterances in parallel and integrating them within the interattention mechanism.

The BCTCE leverages layer-wise multihop attention mechanisms to gradually enhance the interaction between inputs, where the dialogue utterances and the document which can provide supplementary knowledge are used to generate the context-aware and dialogue-consistent responses.

The BCTCE can reduce the time of training and inference compared to the recurrent network-based response generators.

We review the related work in Section 2 and present the details of the proposed model in Section 3. Section 4 shows the experimental process, including datasets and evaluation criteria. The result analysis is also given in Section 4. Finally, we conclude this work and present future work with a brief summary in Section 5.

2. Related Work

Previous models for conversation are generally divided into rule-based, retrieval-based, and generation-based models. The rule-based and retrieval-based models depend on handcrafted rules or existing knowledge bases to match the correct answer, while the generation-based models require less manual effort by leveraging data-driven training of the algorithm on a noisy but large-scale corpus.

Recently, deep neural networks have been widely used for both response retrieval [29, 30] and response generation [31]. Some retrieval-based works determine the correct responses by the semantic similarity between the representations of a require and its candidate answers learned by neural networks. Sequence-to-sequence (seq2seq) frameworks [8] that have achieved success in many domains, such as machine translation [9, 32], have been commonly used for response generation [2, 15, 32]. In particular, seq2seq-based models play an important role in studies on multiturn conversation [4], which commonly build encoder-decoder networks for response generation. They map a sequential syntactic structure to another without explicitly defining features, where recurrent neural networks (RNNs) such as long short-term memory (LSTM) and gated recurrent units (GRUs) [33] are commonly employed as the kernel unit. Vinyals and Quoc explored the LSTM network to produce sequential responses end-to-end for the multiturn conversation [4]. Shang et al. combined global and local context information on the basis of the original RNN for a one-round conversation. Sordoni et al. encoded the semantic information of the context and message by a multilayered nonlinear forward network and took RNN as a decoder to generate responses [3]. Chen et al. utilized a memory network to preserve more historical information in a multiturn dialogue [34]. RNNs are commonly used to sequentially encode each word in the input context and produce the response word-by-word during decoding. However, they were limited by the long time required for sequential training resulting from exploding or vanishing gradient. In addition, the model may suffer from information loss due to hardly capturing long-term semantic dependencies between utterances.

Attention mechanisms have become an integral part of sequence models in response generation, modeling the textual dependencies in the input or output sequences without regard to the position information [35, 36]. In previous works on neural response generation, the attention mechanism was incorporated into the encoder-decoder framework to preserve the key semantic information in sentences [1, 37]. Vaswani et al. proposed a neural network with self-attention and multihead attention to emphasize

different positions of a single sequence and the correlation between each word and its context [25]. Wu et al. proposed an utterance-level attention network combined with a word-level attention network to draw different important parts for the response [38]. Wu et al. proposed a multihop attention mechanism to learn a single context vector by repeatedly computing attention scores [37].

However, previous works merely produce general, rigid, and stylized responses without the natural variation in the language [39]. To address this issue, some studies have proposed context-aware conversational generators to produce more diverse and meaningful responses. Li et al. improved the LSTM-based generator by simply taking maximum mutual information as the objective function [15]. Xing et al. proposed a topic-aware neural generator that leverages topic information to simulate prior knowledge of humans by a joint attention mechanism and a biased generation probability [40]. More works focus on employing extra knowledge to guide the generation and hence tend to generate meaning and context-related responses. Liu et al. presented a neural knowledge diffusion model to introduce knowledge into dialogue generation [41]. Young et al. incorporated common sense knowledge about the concepts covered in utterances into end-to-end conversational models [30]. Madotto et al. used a multihop attention mechanism over memories with pointer networks to effectively incorporate knowledge base information in generative dialogue systems [42]. Moon et al. combined a knowledge graph with conversational utterances to infer the correct entity as the output response [43]. Lian et al. focused on the selection of knowledge for conversational response generation [21]. Both Li et al. [44] and Li et al. [26] proposed document-grounded dialogue generation models to form informative and interesting multiturn responses.

3. Model Architecture

This work proposes a novel transformer-based model which leverages joint encoding of a given document and dialogue for response generation, as shown in Figure 1. It follows the encoder-decoder framework by only using stacked layers, each of which consists of a multihead attention mechanism and position-wise connection network. Encoder-decoder neural networks with attention functions have been widely leveraged for solving sequential language generation [45].

In a multiturn conversational generation task, a dialogue is commonly considered as a sequence of K utterances $\{u_1, u_2, \ldots, u_k\}$, which contains the dialogue history $\{u_1, u_2, \ldots, u_{k-1}\}$ and the current utterance u_k, where $u_i = \{w_1^{u_i}, w_2^{u_i}, \ldots, w_{|u_i|}^{u_i}\}$ denotes the i-th utterance in the multiturn dialogue and $w_j^{u_i}$ denotes the j-th word in the i-th utterance. In this work, we denote the dialogue as a token-level sequence $U = \{w_1^u, w_2^u, \ldots, w_{l_u}^u\}$, where l_u denotes the length of the dialogue sequence. The given document for response generation is denoted as $D = \{w_1^d, w_2^d, \ldots, w_{l_d}^d\}$, where l_d is the length of the document sequence. The dialogue utterances and document are fed into the encoders to learn their distributed representation, which is illustrated in Section 3.2. The output of the decoder in our model is a generated response $R = \{w_1^r, w_2^r, \ldots, w_T^r\}$, where T is the length of the response.

3.1. Attention Mechanism and Multihead Attention. We take advantage of the attention mechanism to capture the interactions between the document and the dialogue, which allows the model to attend the useful information for response generation. We assume that there are n_1 queries and n_2 key-value pairs. Then, we use an attention function to obtain a weighted sum of the values for each query, where the query, key, and value are all vectors of dimension d_k. The weight α assigned to each value is computed by a scaled dot-product function of the query with the corresponding key, shown as the following equation:

$$\begin{aligned} \text{Attention}(Q, K, V) &= \alpha \cdot V, \\ \alpha &= \text{softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right), \end{aligned} \tag{1}$$

where Q is the matrix packing with a set of d_k-dimensional vectors of queries, $Q \in R^{n_1 \times d_k}$. K and V are also matrices, $K \in R^{n_2 \times d_k}$ and $V \in R^{n_2 \times d_k}$. The weight $\alpha \in R^{n_1 \times n_2}$.

Moreover, our model implements multihead attention [25] in all the attention computations to jointly collect information from different representation subspaces at different positions. We define the multihead attention function with M heads for projecting the queries, keys, and values M times with different learned linear projections. The result of the multihead attention function MultiHead is a vector that concatenates all the output vectors across M heads, shown as follows:

$$\text{MultiHead}(Q, K, V) = [\text{head}_1; \ldots; \text{head}_m; \ldots; \text{head}_M], \tag{2}$$

where head_m denotes the m-th weighted vector calculated as the following formula:

$$\begin{aligned} \text{head}_m &= \text{Attention}(Q_m, K_m, V_m), \\ Q_m &= \widetilde{Q} \cdot W_m^Q, \\ K_m &= \widetilde{K} \cdot W_m^K, \\ V_m &= \widetilde{V} \cdot W_m^V, \end{aligned} \tag{3}$$

where $\widetilde{Q}$, $\widetilde{K}$, and $\widetilde{V}$ indicate the vectors of the input query, key, and value with the same dimension d_{model}, respectively. $W_m^Q, W_m^K, W_m^V \in R^{d_{\text{model}} \times d_{\text{att}}}$ are trainable parameter matrices for the m-th head, and $d_{\text{att}} = d_{\text{model}}/M$.

3.2. Encoder. The encoder of the proposed model consists of an utterance encoder and a context encoder. The former aims to learn the representation of the current utterance, and the latter aims to learn the representation of the conversational context (document and dialogue utterances). The encoder is inspired by the reading behavior of human beings. Generally, a basic process of reading comprehension is that

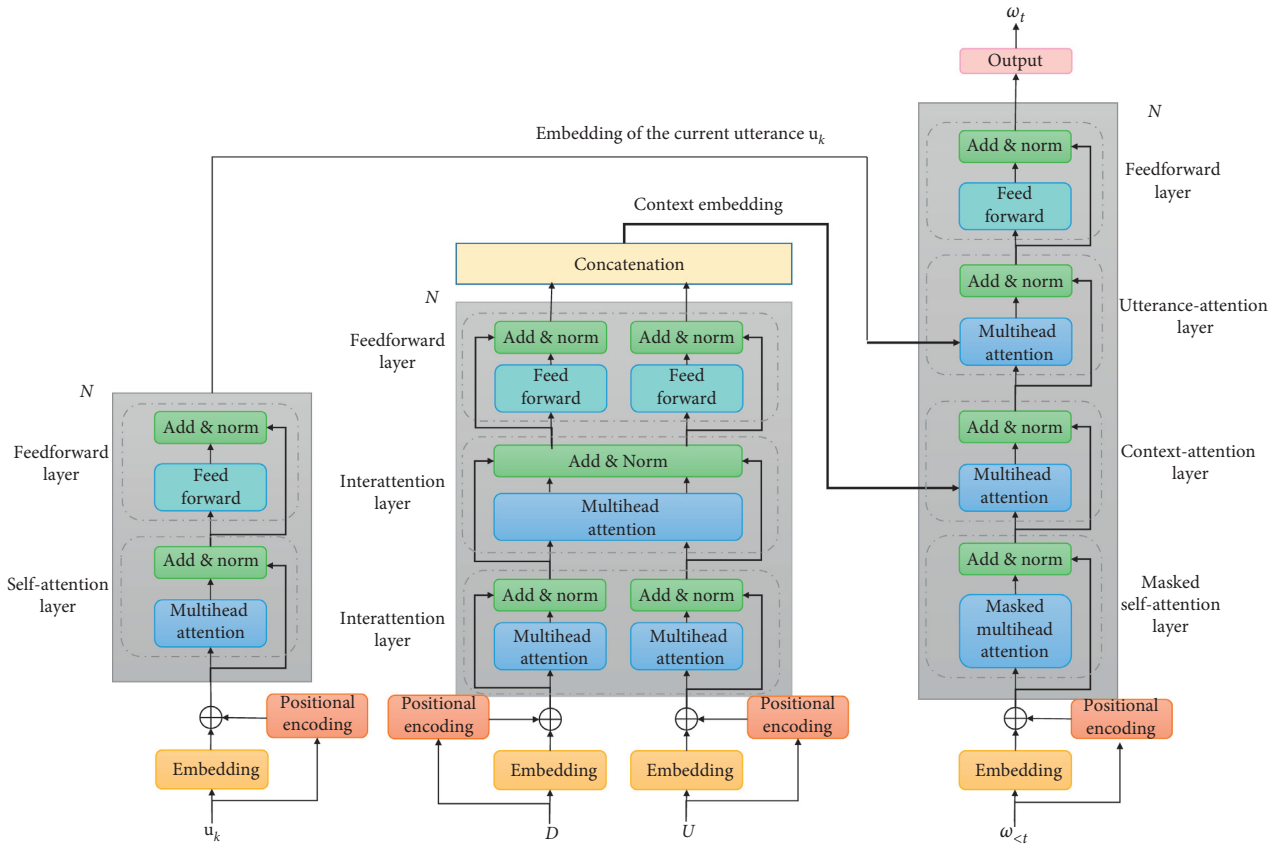

FIGURE 1: The framework of the BCTCE (bichannel transformer with context encoding). The left is the utterance encoder. The center is the context encoder which has binary channels. The right is the decoder for response generation.

firstly reading through the given document and dialogue to understand the theme and capture the key information from the context. Then, we focus on the current utterance for generating the answer. In the context encoder, the given document and dialogue utterances are encoded in parallel by intraattention interaction and interattention interaction. This parallel learning process aims to better represent the context and fuse the information of the document and utterances.

We first map the symbolic representations of input sequences to distribution representations. The tokens of the current utterances, document, and dialogue utterances are fed one by one into the encoder. Moreover, it has been widely accepted that position information is critical to indicate the order of the sequential input. However, the self-attention mechanism itself cannot distinguish between different positions. So, we introduce an additional position embedding to encode position information of the input into the word vectors, shown as the "positional encodings" module in Figure 1. The sum of the original word embedding and the position embedding is defined as the distribution representation $e(w)$ of word w:

$$e(w) = \text{embed}(w) + \text{PE}(w), \tag{4}$$

where $\text{embed}(\cdot)$ denotes an embedding lookup function; the position embedding $\text{PE}(\cdot)$ is defined as in [25]:

$$\text{PE}(w)_{\text{pos},d} = \begin{cases} \sin\left(\text{pos}/10000^{d/d_{\text{model}}}\right), & \text{if } d \text{ is even,} \\ \cos\left(\text{pos}/10000^{(d-1)/d_{\text{model}}}\right), & \text{otherwise,} \end{cases} \tag{5}$$

where pos is the position of w in the dialogue sequence or document sequence, d denotes the d-th dimension of the representation, and d_{model} is the dimension of the input embedding.

The utterance encoder is the same as that of the original transformer [25] with N stacks, shown in the left of Figure 1. It outputs the embedding of the current utterance and has two sublayers in each stack. The first sublayer is constructed by the multihead self-attention mechanism, and the second is a position-wise fully connected feedforward network.

The context encoder is a variant of the transformer encoder with N stacks, shown in the center of Figure 1. Differing from the original transformer that only deals with a single channel in the encoder, it builds binary channels to separately encode both dialogue and document in the first step. The context encoder is composed of a stack of N identical layers, and each stack contains three sublayers:

(i) Intra-attention layer: this layer is employed to encode two individual input sequences using the multihead

self-attention mechanism separately. The representation of dialogue utterances is

$$\overline{h}_n^U = \text{MultiHead}\left(h_{n-1}^U, h_{n-1}^U, h_{n-1}^U\right). \tag{6}$$

The representation of the document is

$$\overline{h}_n^D = \text{MultiHead}\left(h_{n-1}^D, h_{n-1}^D, h_{n-1}^D\right), \tag{7}$$

where h_{n-1}^U and h_{n-1}^D are the outputs of the last stack. h_0^U and h_0^D are the embedding representation of the given dialogues and document.

(ii) Interattention layer: this layer is proposed to fuse the semantic and syntactic features of the dialogue and document. It helps the model to focus on the relevant contents of the document and dialogue since the dialogue-related segment in the document may supply useful information for answering the current require. This layer is also implemented by a multihead attention function, and the query, key, and value are all $[\overline{h}_n^U; \overline{h}_n^D]$. The output of this layer is $\widetilde{h}_n^U$ and $\widetilde{h}_n^D$.

(iii) Feedforward layer: it uses a fully connected position-wise network to nonlinearly map tokens across different positions separately and identically. The nonlinear function used in this layer is

$$\begin{aligned} h_n^U &= \text{Relu}\left(\widetilde{h}_n^U \cdot W_{n1}^U + b_{n1}^U\right) \cdot W_{n2}^U + b_{n2}^U, \\ h_n^D &= \text{Relu}\left(\widetilde{h}_n^D \cdot W_{n1}^D + b_{n1}^D\right) \cdot W_{n2}^D + b_{n2}^D, \end{aligned} \tag{8}$$

where $W_{n1}^U, W_{n1}^D \in R^{d_{\text{model}} \times d_{\text{inner}}}$, $W_{n2}^U, W_{n2}^D \in R^{d_{\text{inner}} \times d_{\text{model}}}$, $b_{n1}^U, b_{n1}^D \in R^{d_{\text{inner}}}$, and $b_{n2}^U, b_{n2}^D \in R^{d_{\text{model}}}$ are trainable parameters. d_{inner} is the size of the hidden layer in the feedforward network.

In addition, each sublayer has an "Add & Norm" operation, which is defined in the original transformer framework. The output of the last stack is matric h_N^U and matric h_N^D, which are concatenated as the output of the context encoder.

3.3. Decoder. The decoder also has N stacks and contains three layers per stack, as shown in the right of Figure 1. At time step t, the previous $t-1$ tokens and the output of the encoder are fed into the decoder to predict the t-th token in the response illustrated as the output of the N-th stack $h_t^R \in R^{d_{\text{model}}}$.

(i) Masked self-attention layer: this layer is similar to the intra-attention layer in the encoder. The difference is that we mask the subsequent positions of each token to ensure that the consequent utterance only depends on the previous tokens.

(ii) Context-attention layer: to enrich the context information in generated responses, the output of the context encoder is fed into the decoder and integrated with the previous response tokens by a multihead attention function, where the query of the function is the output of the masked self-attention layer, and the key and the value are the output of the context encoder.

(iii) Utterance-attention layer: generally, the generated response must be relevant to the current utterance. Thus, we also use a multihead attention function to introduce the key information of the current utterance for generating dialogue-related responses in this layer. The query of the function is the output of the context-attention layer, and the key and the value are the output of the utterance encoder.

(iv) Feedforward layer: this layer is the same as the feedforward layer in the encoder.

We select the token derived from an external vocabulary V_o; the probability of each candidate token being chosen is

$$P^v(w_t) = \text{softmax}\left(h_t^R \cdot e(w_t)\right). \tag{9}$$

Then, we define the probability of generating the t-th token as $P(w_t)$:

$$P(w_t) = P^v(w_t). \tag{10}$$

At each time step, we select the token that has the highest probability as the generated token:

$$\widehat{w}_t = \text{argmax}\left(P(w_t)\right). \tag{11}$$

During the training process, the loss for time step t is defined as the negative log-likelihood of the target word w_t^*:

$$l_t = -\log P(w_t^*). \tag{12}$$

The final loss is

$$l = \frac{1}{T}\sum_t l_t. \tag{13}$$

3.4. Copying Mechanism. In this work, we tend to generate more imaginative and context-aware responses. However, some tokens in the ground truth may not be included in the vocabulary (OOV, out of vocabulary). As such, we propose a variant of our transformer model that incorporates the copying mechanism [46, 47] into the decoder to generate tokens that appear in the document and dialogue in addition to the external vocabulary. The tokens in generated responses may be chosen from the input or an external vocabulary according to a computed probability.

At each time step t, according to the multihead attention weights resulted from the context-attention layer, we determine the probability that the generated tokens are derived from the input document and dialogue utterances as the average of all the attention weights $\alpha_t^1, \alpha_t^2, \ldots, \alpha_t^M$ derived from M heads:

$$\alpha_t = \frac{1}{M}\sum_{m=1}^{M}\alpha_t^m, \tag{14}$$

where $\alpha_t \in R^{l_u+l_d}$ and α_t^m indicates the attention weight of the m-th head.

According to the copying mechanism, the probability of tokens being chosen from the vocabulary is $p_t^g \in [0, 1]$, while the probability of tokens being chosen from the input sequences is $1 - p_t^g$:

$$p_t^g = \text{sigmoid}\left(W^T \cdot h_t^R\right), \tag{15}$$

where W is a trainable parameter.

The probability of the t-th generated token w_t is calculated according to the source it is derived from as follows:

$$P(w_t) = \begin{cases} p_t^g * P^v(w_t) + \left(1 - p_t^g\right) * \sum_{i:w_i=w_t}\alpha_t(i), & \text{if } \mathrm{w_t} \in \mathrm{V_o} \text{ and } \mathrm{w_t} \in \mathrm{U} \cup \mathrm{D}, \\ \left(1 - p_t^g\right) * \sum_{i:w_i=w_t}\alpha_t(i), & \text{if } \mathrm{w_t} \notin \mathrm{V_o} \text{ and } \mathrm{w_t} \in \mathrm{U} \cup \mathrm{D}, \\ p_t^g * P^v(w_t), & \text{if } \mathrm{w_t} \in \mathrm{V_o} \text{ and } \mathrm{w_t} \notin \mathrm{U} \cup \mathrm{D}, \\ 0, & \text{otherwise.} \end{cases} \tag{16}$$

4. Experiments

4.1. Experiment Settings. We conduct the experiments, and the stacks of both encoder and decoder are set to 4. The number of attention heads is set to 8. The dimension of input embedding d_{model} is set to 512, and the hidden size of the feedforward network d_{inner} is set to 2,048. In the process of encoding, we take the previous four utterances and the given document as the input. We use the Adam algorithm [48] with learning rate 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.

4.2. Dataset. We evaluate the proposed model on the dataset CMU_DoG (CMU document-grounded conversations) for document-driven conversations [24]. This dataset consists of a set of documents and a spectrum of dialogues corresponding to each document, which may contain movie names, ratings, introduction, and some other scenes. The documents present conversation-related information that may help generate context-aware responses in a multiturn conversation task.

The dataset has a total of 4,112 conversations with an average of 21.43 turns. The dialogue utterances are derived from two different scenarios, both of which involve two participants. In the first scenario, only one participant has access to the given document, while both participants have access to the same given document in the second scenario. The number of conversations for scenario ONE is 2,128, and for scenario TWO, it is 1,984. This high-quality dataset explicitly presents the corresponding relationship between each section of a document and the conversation turns. The average length of documents is approximately 200. There are 72,922 utterances for training, 3,626 utterances for validation, and 11,577 utterances for testing.

4.3. Evaluation

4.3.1. Baselines. We take the transformer [25], incremental transformer [44], D3G [26], SEQ, and SEQS [24] as the baseline models, which are proposed for document-grounded conversational generation.

4.3.2. Quantitative Evaluation. To measure the performance of the proposed model and the baselines, we take BLEU [27], METEOR [28], NW [24], and perplexity (PPL) [19] as evaluation criteria to perform automatic evaluation.

(1) BLEU: BLEU is known to correlate reasonably well with human evaluation on the task of conversational response generation. It measures n–gram overlap between generated responses and the ground truth, which is defined as BLEU–n. We calculate various BLEU scores between the golden responses and the generated responses. Moreover, we calculate the unigram overlap between the given document and the generated responses to further compare models in terms of the correlation between the responses and the document. Therefore, we only use the BLEU–1 score (called as Doc_BLEU) and ignore the brevity penalty factor in the BLEU computation.

(2) METEOR: we also compare our proposed model with state-of-the-art baselines in terms of the METEOR metric under the full mode (this mode contains the exact matching between words and phrase matching between stems, synonyms, and paraphrases). METEOR, which focuses on the recall rate, has more relevance with human judgment in comparison to BLEU.

(3) NW: we explore the set operation (NW) to evaluate the relevance between documents and the conversations generated by the models. Let the set of tokens in the generated response be N, the set of tokens in

the document be M, the set of tokens in the previous three utterances be H, and the set of stop words be S. We calculate the set operation (NW) as $|((N \cap M)/H)/S|$. A higher NW score indicates that more tokens that appear in the document are used to expand the information in responses.

(4) Perplexity: in addition to the previous three criteria, we use perplexity to automatically evaluate the fluency of the response. Lower perplexity indicates better performance of the models and higher quality of the generated sentences.

4.3.3. Human Judgment. Manual evaluations are essential for dialogue generation. So, we augment the automatic evaluation with the human judgment of fluency, dialogue coherence, and lexical diversity. All the three evaluation metrics are scored 0/1/2. We randomly sample multiple conversations containing 822 utterances from the test set. We used a crowdsourcing service that asks annotators to score these utterances given its previous utterances and related documents. The final score of each utterance is the average of the scores rated by three annotators.

(1) Fluency: whether the response is natural and fluent. Score 0 represents the response is not fluent and incomprehensible; 1 represents the response is partially fluent but still comprehensible; and 2 represents the response is sufficiently fluent.

(2) Dialogue coherence: whether the response is logically coherent with the dialogue. Score 0 represents the response is irrelevant with the previous utterances; 1 represents the response matches the topic of the previous utterances; and 2 represents the response is exactly coherent with the previous utterances.

(3) Lexical diversity: whether the response is vivid and diverse. Score 0 represents the safe response which is applicable to almost all conversations, e.g., "i think so" and "i agree with you"; 1 represents the response suitable to limited conversations but plain and uninformative; and 2 represents the response is evidently vivid, diverse, and informative.

4.4. Results and Discussion. Tables 1–3 show the comparison of our model with other models on the CMU_DoG dataset. As shown in Table 1, our model outperforms all baselines in terms of BLEU−n scores and METEOR score. Our model achieves a new state-of-the-art performance, which indicates that the responses generated by our model are more similar to the ground-truth responses. Our model significantly outperforms all the baselines by at least 1.43, 0.68, 0.39, 0.47, and 0.01 in terms of BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR, respectively.

Table 2 shows the comparison of document relevance and response quality in terms of Doc_BLEU score, NW, the average length of responses (avg_len), and the PPL score. Our model outperforms the baselines by 4.5%–11.9% in terms of BELU-1, while it has lower NW score than D3G and incremental transformer (our impl). These results indicate that our model can more effectively use the shared information between the document and the dialogue to produce responses than D3G and incremental transformer (our impl). Moreover, the average length of the responses generated by our model is higher than that of the baselines, which shows that our model may generate more informative responses. In addition, our model achieves a competitive PPL score with others.

As the results of human judgment shown in Table 3, our transformer-based model outperforms all the baselines in terms of the dialogue coherence and diversity. However, the performance of our model is slightly worse than incremental transformer (our impl) [44] on fluency.

4.5. Training Time. Figure 2 presents the training time for one epoch of our BCTCE model and some baselines (As the official source code of "Incremental Transformer," the process of training for each step is followed by the evaluation of the generated responses. Therefore, it is hard to get the actual training time for one epoch of "Incremental Transformer" model.). For a fair comparison, all models use the same batch size, max length of the document, max length of the dialogue sequence, and max length of the response.

As shown in the figure, the training time for our BCTCE model is much less than D3G, while it is higher than other models. The reasons are as follows:

(1) The SEQ model is a simple sequence-to-sequence RNN model with attention mechanism. It only uses the dialogue as the input of the encoder and discards the document, thus requiring considerably less time for model training than our model.

(2) The SEQS model feeds the concatenation of the encoded utterances and document into the decoder, and the original transformer directly takes the concatenation of the utterances and document as the input of the model. Although they have lower training time, our model extends the original transformer by introducing an extra context encoder and a "context-attention layer" in the decoder, where the interaction between utterances and document is time-consuming but effective.

(3) Our transformer-based model significantly reduces the training time in comparison with the D3G model which also uses the interaction between document and dialogue in its RNN structure.

4.6. Ablation Study. To validate the effectiveness of each module of the BCTCE, we conduct ablation experiments on the CMU_DoG dataset:

(1) +copy: we introduce the copy mechanism into the BCTCE model to generate the response from the document and dialogue utterances in addition to external vocabulary.

Table 1: Quantitative evaluation results for baselines and the proposed models in terms of BLEU scores and METEOR score.

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
SEQ [24]§	6.12	1.52	0.59	0.30	4.18
SEQS [24]§	6.57	1.65	0.67	0.35	4.30
D3G [26]§	6.32	1.71	0.71	0.41	4.17
Transformer [25]¶	8.55	2.49	1.12	0.60	4.53
Incremental transformer [44]†	—	—	—	0.95	—
Incremental transformer (our impl)‡	8.19	2.88	1.66	0.85	5.21
BCTCE	9.98	3.56	2.05	1.42	5.22

Results marked with § are trained and evaluated with the source code from [26], results marked with ¶ are trained and evaluated with our implemented code, results marked with † are from [44], and results marked with ‡ are trained and evaluated with the code published by Li et al. [44].

Table 2: Quantitative evaluation results of document relevance and response quality.

Models	Doc_BLEU	NW	avg_len	PPL
SEQ [24]§	24.88	0.23	7.31	15.62
SEQS [24]§	27.96	0.34	7.21	19.53
D3G [26]§	26.76	0.39	6.83	18.40
Transformer [25]¶	27.55	0.33	7.91	13.70
Incremental transformer [44]†	—	—	—	15.11
Incremental transformer (our impl)‡	26.96	0.42	8.52	11.01
BCTCE	28.23	0.36	9.16	17.80

The marks in this table are the same as those in Table 1.

Table 3: Human evaluation of baselines and our proposed model.

Models	Fluency	Dialogue coherence	Diversity
SEQ [24]§	1.27	0.81	0.42
SEQS [24]§	1.13	0.96	0.71
D3G [26]§	1.29	1.12	0.84
Transformer [25]¶	1.34	1.17	0.90
Incremental transformer (our impl)‡	1.35	1.27	0.93
BCTCE	1.34	1.29	0.95

The marks in this table are the same as those in Table 1.

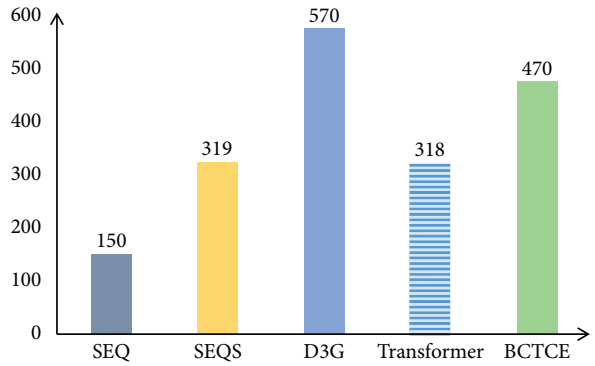

Figure 2: Training time of various models.

(2) -document: we replace the context encoder with the original transformer and take only dialogue utterances as its input

(3) -history: we remove the dialogue history from the inputs, remaining only the current dialogue utterance and document

(4) -context encoder: we discard the context encoder and the context-attention layer in the decoder

(5) -utterance encoder: we discard the utterance encoder and the utterance-attention layer in the decoder

(6) -bi_channel: we replace the context encoder with the original transformer and take the concatenation of the dialogue utterances and document as its input

The automatic evaluation results are shown in Tables 4 and 5, respectively. The human evaluation results are shown in Table 6. As shown in Table 4, the ablation models, which remove some modules of BCTCE, perform worse than the basic BCTCE model on the similarity between generated responses and ground truth (BLEU-n scores and METEOR score). The results of "-bi_channel" indicate that the interaction between the document and the dialogue in bichannel encoding is effective for generating responses. The results of "-context encoder" and "-utterance encoder" show that the context encoder and the utterance encoder are beneficial for response generation. The results of "--document" and "-history" represent that the multiturn dialogue and the document knowledge are important as they contain some vital information useful for generating reasonable response.

Table 5 shows that removing the document or introducing copy mechanism reduces the Doc_BLEU and NW scores. The results indicate that the BCTCE may pay more

Table 4: Ablation study in terms of BLEU scores and METEOR score.

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
BCTCE	9.98	3.56	2.05	1.42	5.22
+copy	9.53	3.21	1.70	1.08	4.95
-document	8.55	2.33	0.95	0.46	4.62
-history	8.50	2.84	1.51	1.01	4.54
-context encoder	7.68	1.96	0.74	0.34	4.06
-utterance encoder	9.87	3.34	1.89	1.30	5.13
-bi_channel	9.40	3.05	1.58	1.02	4.80

Table 5: Ablation study on document relevance and response quality.

Models	Doc_BLEU	NW	avg_len	PPL
BCTCE	28.23	0.36	9.16	17.80
+copy	27.38	0.34	8.63	16.33
-document	26.25	0.23	8.42	12.97
-history	28.47	**0.65**	8.05	15.97
-context encoder	23.66	0.39	6.72	**11.23**
-utterance encoder	28.25	0.41	**9.40**	16.88
-bi_channel	**29.15**	0.33	7.94	14.26

Table 6: Human evaluation for ablation study. All metrics are scored 0/1/2.

Models	Fluency	Dialogue coherence	Diversity
BCTCE	1.34	**1.29**	0.95
+copy	1.33	1.26	**1.01**
-document	1.30	1.27	0.80
-history	1.31	1.06	0.88
-context encoder	**1.36**	0.81	0.53
-utterance encoder	1.34	1.27	0.87
-bi_channel	1.29	1.26	0.85

attention to the dialogue utterances after removing the document information, and the copy mechanism has less influence on the generated response than expected since the BCTCE has sufficient capability to learn the document knowledge for response generation. The Doc_BLEU and NW scores increase when removing the dialogue history or utterance encoder from the basic BCTCE model as the lack of sufficient dialogue information makes the model to be more focused on the document. The ablation model "--bi_channel" increases the Doc_BLEU score and reduces the NW score, which indicate its generated responses pay a little more attention to the shared information between the document and the dialogue. It is worth noting that the ablation model "-context encoder" significantly outperforms the BCTCE model on PPL and the fluency shown in Table 6. A possible reason is that it tends to generate safe and unremarkable responses (e.g., I don't know). The decrease of its avg_len and the diversity shown in Table 6 also supports our argument. Moreover, the copy mechanism effectively reduces the PPL from 17.80 to 16.33 and increases the diversity from 0.95 to 1.01, which indicate that it can improve the response quality of the basic BCTCE model.

4.7. Case Study. In this section, we demonstrate three conversation cases and show the responses generated by our BCTCE model and several ablation models (as shown in Table 7). The first case shows that our BCTCE model produces a response that contains an "OOV," while the BCTCE with a copy mechanism extracts the token "post" and "times" from the document. It indicates that the copy mechanism eases the "OOV" problem, although it reduces the performance of the basic BCTCE model on many evaluation metrics.

In the second case, the BCTCE without the document or context encoder produces a completely incorrect answer, while the basic BCTCE and other ablation models can produce at least one correct answer "jime carrey" contained in the ground truth. These results show that the document is necessary for generating answers, and the BCTCE can take full advantage of the valid information in the document.

The third case indicates the importance of the dialogue history. For the conversation whose current utterance is an uninformative interjection "wow," the BCTCE model produces a response "all around, it has a lot of great stories," referring to the "story" in the previous utterances, while the

Table 7: Sample responses of multiple variants.

Document	Dialogue	Responses
. . .the white house retaliates, and in short order the post and times appear together before the supreme court to plead their first amendment argument for the right to publish the material. meanwhile, newspapers across the country pick up the story in solidarity with the post and times. the court rules 6 3 in the newspapers' favor, vindicating graham's decision. nixon demands that the post should be barred from the white house. one year later, a security guard discovers a break-in in progress at the watergate complex . . .	. . . the supreme court ruled 6-3 in favor of the news papers, which vindicated graham's decision. Now i really want to see it. it sounds really juicy, but it will be a little difficult because i feel like the situation will anger me.	Ground truth: but they won. nixon god mad and barred the post from the white house BTMPE: OOV -document: i think the post won the oscar for the OOV -history: i like the OOV of the white house, but they OOV the post to OOV the house. -context encoder: i think that is a good way to end the movie. -utterance encoder: it sounds like a good movie to Watch. -bi_channel: and then the supreme court 6-3 in the supreme court, the court rules. +copy: i wonder if the post joined the times or not.
. . .bruce nolan (jim carrey) is a television field reporter for eyewitness news at wkbw-tv in buffalo, new york, but desires to be the news anchorman . he is in a relationship with his girlfriend grace connelly (jennifer aniston), but also has a mild crush on his co-worker, susan ortega (catherine bell) , who barely seems to notice him. bruce, however, suffers from constant bad luck and reaches breaking point when he is passed over for promotion by his rival, evan baxter (steve carell) . . .	. . . it did not get a good rating on rotten tomatoes. i liked it though that happens a lot with rotten tomato who is in the movie?	Ground truth: jim carrey and jennifer aniston BTMPE: it is a comedy film starring jim carrey and jennifer aniston as the main character -document: it's got 92% on rotten tomatoes -history: jim carrey plays bruce nolan. -context encoder: the movie is about a man named michael oher who is a stockbroker. -utterance encoder: jim carrey plays the lead -bi_channel: jim carrey is a funny actor. +copy: jim carrey
. . .the duo tries to make it to the moving truck, but sid's dog, scud, sees them, and gives chase. buzz gets left behind while saving woody from scud, and woody tries rescuing him with andy's rc car, but the other toys, thinking woody eliminated rc as well, attack and toss him off the truck. having evaded scud, buzz and rc retrieve woody, and continue after the truck. upon seeing woody and buzz together on rc, the other toys realize their mistake, and . . .	. . . so tell me the good things you heard about it heard that it had a cute story line/plot and that there are three movies total, toy story 1, 2, and 3 wow	Ground truth: that's about it BTMPE: all around, it has a lot of great stories -document: i'm -history: yeah he did a great job. -context encoder: OOV -utterance encoder: do you know what kind of movie it is? -bi_channel: have a great day +copy: i like the first pixar movies

BCTCE without the dialogue history outputs a dialogue-independent response "yeah he did a great job" that contains an uncertain person pronoun. And the BCTCE without the context encoder produces an uninformative "OOV." This case shows that the BCTCE without the dialogue history or context encoder generates dialogue-inconsistent responses, while the responses generated by others which use historical dialogue information are more reasonable and consistent with the current utterance.

5. Conclusions

In this paper, we propose the BCTCE model based on the transformer framework for document-driven conversational

response generation. This model consists of a context encoder, an utterance encoder, and a decoder, aiming to generate context-aware responses. The context encoder is used to learn the distributed representations of the input document and dialogue utterances in parallel and then fuse them by the multihead self-attention function. The utterance encoder aims to represent the current utterance as distributed embedding. The decoder focuses on the document knowledge and dialogue coherence for predicting the next response. The competitive models are evaluated by comparing the generated responses with the ground truth. Empirical results show that the BCTCE outperforms state-of-the-art baselines in terms of various BLEU scores, METEOR, and NW. The effectiveness of the modules in the BCTCE is indicated by the ablation study. And the manual evaluation and case study show that our model can capture the useful information contained in the document and dialogue, which helps to generate diverse and reasonable responses with much more relevance with the context. In the future work, we will try to build various encoders and concatenate the output from the encoders to integrate the input sequences and generate reasonable context representation.

Data Availability

The textual data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Authors' Contributions

All authors have contributed to the creation of this manuscript for important intellectual content and approved the final manuscript.

Acknowledgments

This work was supported in part by the R&D Program of Beijing Municipal Commission of Education (no. KM202010011011), the Beijing Municipal Natural Science Foundation (nos. 4202014 and 4184084), National Key R&D Program of China (nos. 2019YFC1606401, 2018YFC0831605, and 2016YFD0401205), Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan (no. CIT&TCD201804031), the Social Science Foundation of Beijing (no. 19GLB036), and the Humanity and Social Science Youth Foundation of Ministry of Education of China (no. 17YJCZH007).

References

[1] L. Shang, Z. Lu, and H. Li, "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 Language Processing (Volume 1: Long Papers)*, pp. 1577–1586, Beijing, China, July 2015.

[2] I. V. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pp. 3776–3784, Phoenix, AZ, USA, February 2016.

[3] A. Sordoni, M. Galley, M. Auli et al., "A neural network approach to context-sensitive generation of conversational responses," *Transactions of the Royal Society of Tropical Medicine & Hygiene*, vol. 51, no. 6, pp. 502–504, 2015.

[4] O. Vinyals and Q. Le, "A neural conversational model," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015*, Lille, France, July 2015.

[5] 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 *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Volume 1: Long Papers*, pp. 3721–3730, Florence, Italy, July 2019.

[6] Z. Li, F. Xiong, X. Wang, H. Chen, and X. Xiong, "Topological Influence-Aware Recommendation on Social Networks," *Complexity*, vol. 2019, Article ID 6325654, 12 pages, 2019.

[7] Y. Wu, W. Wu, C. Xing, C. Xu, Z. Li, and M. Zhou, "A sequential matching framework for multi-turn response selection in retrieval-based chatbots," *Computational Linguistics*, vol. 45, no. 1, pp. 163–197, 2019.

[8] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Proceedings of the Annual Conference on Neural Information Processing Systems 2014*, pp. 3104–3112, Montreal, Canada, December 2014.

[9] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 2015.

[10] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pp. 1412–1421, Lisbon, Portugal, September 2015.

[11] Z. Lin, M. Feng, C. N. dos Santos et al., "A structured self-attentive sentence embedding," in *5th International Conference on Learning Representations, ICLR 2017*, Toulon, France, April 2017.

[12] C. Peng, Z. Sun, L. Bing, and Y. Wei, "Recurrent attention network on memory for aspect sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, September 2017.

[13] F. Xiong, X. Wang, S. Pan, H. Yang, H. Wang, and C. Zhang, "Social recommendation with evolutionary opinion dynamics," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018.

[14] F. Xiong, W. Shen, H. Chen, S. Pan, X. Wang, and Z. Yan, "Exploiting implicit influence from information propagation for social recommendation," *IEEE Transactions on Cybernetics*, 2019.

[15] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversitypromoting objective function for neural conversation models," *Computer Science*, 2015.

[16] T. Mikolov, M. Karafiát, L. Burget, J. Cernocký, and S. Khudanpur, "Recurrent neural network based language model," in *Proceedings of the INTERSPEECH 2010, 11th*

Annual Conference of the International Speech Communication Association, pp. 1045–1048, Makuhari, Japan, September 2010.

[17] H. Wen, M. Gasic, N. Mrksic, P. H. Su, D. Vandyke, and S. Young, "Semantically conditioned LSTM-based natural language generation for spoken dialogue systems," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pp. 1711–1721, Lisbon, Portugal, September 2015.

[18] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018*, pp. 4623–4629, Stockholm, Sweden, July 2018.

[19] E. Dinan, S. Roller, K. Shuster, A. Fan, M. Auli, and J. Weston, "Wizard of wikipedia: knowledge-powered conversational agents," in *Proceedings of the 7th International Conference on Learning Representations, ICLR 2019*, New Orleans, LA, USA, May 2019.

[20] W. Cui, Y. Xiao, H. Wang, Y. Song, S. Hwang, and W. Wang, "KBQA: learning question answering over QA corpora and knowledge bases," in *Proceedings of the 43rd International Conference on Very Large Data Bases, VLDB 2017*, pp. 565–576, Munich, Germany, August 2017.

[21] R. Lian, M. Xie, F. Wang, J. Peng, and H. Wu, "Learning to select knowledge for response generation in dialog systems," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019*, pp. 5081–5087, Macao, China, August 2019.

[22] M. Ghazvininejad, C. Brockett, M. Chang et al., "A knowledge-grounded neural conversation model," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 5110–5117, New Orleans, LA, USA, February 2018.

[23] H. Le, D. Sahoo, N. Chen, and S. Hoi, "Multimodal transformer networks for end-to-end video-grounded dialogue systems," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019 Volume 1: Long Papers*, pp. 5612–5623, Florence, Italy, July 2019.

[24] K. Zhou, S. Prabhumoye, and A. W. Black, "A dataset for document grounded conversations," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 708–713, Brussels, Belgium, October 2018.

[25] A. Vaswani, N. Shazeer, N. Parmar et al., "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pp. 5998–6008, Long Beach, CA, USA, December 2017.

[26] K. Li, Z. Bai, X. Wang, and C. Yuan, "A document driven dialogue generation model," in *Chinese Computational Linguistics—18th China National Conference, CCL 2019*, pp. 508–520, Kunming, China, October 2019.

[27] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a method for automatic evaluation of machine translationk," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics—ACL '02*, pp. 311–318, Stroudsburg, PA, USA, July 2002.

[28] M. Denkowski and A. Lavie, "Meteor universal: language specific translation evaluation for any target language," in *Proceedings of the EACL 2014 Workshop on Statistical Machine Translation*, Baltimore, MD, USA, 2014.

[29] X. Zhou, L. Li, D. Dong et al., "Multi-turn response selection for chatbots with deep attention matching network," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018 Volume 1: Long Papers*, pp. 1118–1127, Melbourne, Australia, July 2018.

[30] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialog systems with commonsense knowledge," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, USA, February 2018.

[31] L. Nio, S. Sakti, G. Neubig, K. Yoshino, and S. Nakamura, "Neural network approaches to dialog response retrieval and generation," *IEICE Transactions on Information and Systems*, vol. E99.D, no. 10, pp. 2508–2517, 2016.

[32] K. Cho, B. van Merrienboer, C. Gulcehre et al., "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pp. 1724–1734, Doha, Qatar, October 2014.

[33] J. Chung, Ç. Gülçehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on sequence modeling," 2014, https://arxiv.org/abs/1412.3555.

[34] H. Chen, Z. Ren, J. Tang, Y. E. Zhao, and D. Yin, "Hierarchical variational memory network for dialogue generation," in *Proceedings of the 2018 World Wide Web Conference on World Wide Web—WWW'18*, pp. 1653–1662, Lyon, France, Apirl 2018.

[35] Y. Kim, C. Denton, L. Hoang, and A. M. Rush, "Structured attention networks," in *Proceedings of the International Conference on Learning Representations*, Toulon, France, April 2017.

[36] A. P. Parikh, O. Täckström, D. Das, and J. Uszkoreit, "A decomposable attention model for natural language inference," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016*, pp. 2249–2255, Austin, TX, USA, November 2016.

[37] X. Wu, A. Martinez, and M. Klyen, "Dialog generation using multi-turn reasoning neural networks," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pp. 2049–2059, New Orleans, LA, USA, June 2018.

[38] W. Wu, C. Xing, Y. Wu, M. Zhou, Y. Huang, and W. Y. Ma, "Hierarchical recurrent attention network for response generation," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, LA, USA, February 2018.

[39] T. H. Wen, D. Vandyke, N. Mrkšić et al., "A network-based end-to-end trainable task-oriented dialogue system," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017 Volume 1: Long Papers*, pp. 438–449, Valencia, Spain, April 2017.

[40] C. Xing, W. Wu, Y. Wu et al., "Topic aware neural response generation," in *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, pp. 3351–3357, San Francisco, CA, USA, February 2017.

[41] S. Liu, H. Chen, Z. Ren, Y. Feng, Q. Liu, and D. Yin, "Knowledge diffusion for neural dialogue generation," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018 (Volume 1: Long Papers)*, pp. 1489–1498, Melbourne, Australia, July 2018.

[42] A. Madotto, C.-S. Wu, and P. Fung, "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 (Volume 1: Long Papers), pp. 1468–1478, Melbourne, Australia, July 2018.

[43] S. Moon, P. Shah, A. Kumar, and R. Subba, "OpenDialKG: explainable conversational reasoning with attention-based walks over knowledge graphs," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 845–854, Florence, Italy, July 2019.

[44] Z. Li, C. Niu, F. Meng, Y. Feng, Q. Li, and J. Zhou, "Incremental transformer with deliberation decoder for document grounded conversations," in *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019 Volume 1: Long Papers*, pp. 12–21, Florence, Italy, July 2019.

[45] V.-K. Tran and L.-M. Nguyen, "Natural language generation for spoken dialogue system using RNN encoder-decoder network," in *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, Vancouver, Canada, August 2017.

[46] A. See, P. J. Liu, and C. D. Manning, "Get to the point: summarization with pointer-generator networks," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017 (Volume 1: Long Papers)*, pp. 1073–1083, Vancouver, Canada, July 2017.

[47] J. Gu, Z. Lu, H. Li, and V. O. K. Li, "Incorporating copying mechanism in sequence-to-sequence learning," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1631–1640, Berlin, Germany, August 2016.

[48] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," in *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 2015.