Knowledge-based Context-aware Multi-turn Conversational Model with Hierarchical Attention

Chunquan Chen, Si Li *
School of Information and Communication Engineering
Beijing University of Post and Telecommunications
Beijing, China
{ccq1996, lisi}@bupt.edu.cn

***Abstract*—We study response generation in multi-turn open-domain dialogue systems. Background knowledge based response generation has been developed to make dialogue models generate more informative and appropriate responses. However, these knowledge-based dialogue models are limited to the domain of single round conversation, and fail to consider the role of dialogue context in the selection of relevant knowledge and response generation. As a result, these models might lose some useful information in the dialogue context and generate irrelevant responses. We argue that both dialogue context and relevant knowledge play important roles in the response generation of multi-turn open-domain dialogue systems. We propose a Knowledge-based Context-aware Multi-turn Conversational (*KCMC*) model to consider both dialogue context and relevant knowledge in a unified framework. The Knowledge Fusion module is designed to augment the semantic representation of dialogue context with associated knowledge triples. And we introduce hierarchical encoders to model the hierarchy of dialogue context and to capture important information in the dialogue context. Furthermore, a hierarchical attention mechanism attends to important parts of knowledge triples, which facilitates better knowledge selection and response generation. Through extensive experiments on two datasets, we demonstrate that the proposed model is capable of generating more informative and appropriate responses than baseline models.**

***Index Terms*—dialogue model, hierarchical attention, context, knowledge graph, memory**

I. Introduction

Recently, attention-based sequence-to-sequence models [1], [2] have been successfully applied to many natural language tasks including machine translation [1]–[4], text summarization [5]–[9], and reading comprehension [10], [11]. However, dialogues can have multiple valid responses with varying semantic content, different from aforementioned tasks where the generation is more uniquely constrained by the input source. Although lots of research efforts [12], [13] have been devoted to open-domain conversational model, achieving satisfactory performance on dialogue still remains a difficult problem. There have been several attempts to generate an output sentence for a given input sentence, but they tend to generate generic responses such as "I don't know", which are dull and meaningless in most cases. This is due to the fact that these models do not take into account the preceding context. It is quite tough to completely understand the meaning of a sentence without considering the content of preceding conversation, to say nothing of generating an appropriate and informative response.

In order to generate more appropriate response, researchers [14], [15] have taken conversational context into consideration and proposed response generation models for multi-turn conversation. These models produce relevant responses based on dialog context which refers to a message and several utterances in its previous turns. However, it is still quite challenging to generate a meaningful and informative response merely from dialogue context [16] without the help of relevant knowledge. This is due to the fact that socially shared knowledge is the background information that people intended to know and use during the conversation. We think that the background knowledge is essential to bridge the semantic gap of dialog context and response. Some studies [17], [18] have been conducted to introduce knowledge in conversation generation. The introduced knowledge is either unstructured texts [16] or domain-specific knowledge triples [18].

The dialogue context information is far from enough for generating an appropriate response, for the reason that in the real-world context, humans respond not only based on dialogue context, but also their knowledge in mind about the dialogue topic. Inspired by this, we argue that it is vital to jointly take into account dialogue context and associated knowledge in a unified framework for generating coherent and informative responses. On the one hand, a model can understand the dialogue context better and thus respond more properly with the help of external knowledge which facilitates semantic understanding. On the other hand, a model can select relevant knowledge more properly given the whole dialogue context which facilitates knowledge selection. Considering dialog context and knowledge together might yield mutually reinforcing advantages for generating more informative and coherent responses in open-domain conversation. However, based on our own knowledge of the task, there is less study on that.

In this work, we propose a novel knowledge-based context-aware multi-turn conversational (*KCMC*) model that jointly take into account dialogue context and associated knowledge

* Corresponding author

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

in a unified framework. The *KCMC* model is based on the effective encoder-decoder framework and is built in a hierarchical structure. We introduce hierarchical encoders to model the hierarchy and the important part of conversational context. Besides, two main components of *KCMC* model, Knowledge Fusion module and the knowledge-enhanced decoder, effectively introduce the relevant knowledge into the response generation model. Specially, before entering the encoder, the Knowledge Fusion module is firstly utilized to enhance the semantic representation of word with associated knowledge triples. In detail, the Knowledge Fusion module retrieves relevant knowledge triples for each word in dialog context and then encode the retrieved knowledge triples as a whole graph with static attention mechanism. The encoded knowledge is used to augment the semantic of word representation. Then the model uses a word level encoder and a word level attention to represent each utterance as an utterance vector. And then the utterance vectors are fed to an utterance level encoder, an utterance level attention mechanism is used to obtain the whole context vector. Furthermore, the knowledge-enhanced decoder attentively read knowledge graphs and the triples in each graph to improve the response generation process. Finally, during the decoding process, the decoder either generates a generic word from a fixed vocabulary or copies a token from dialog context or knowledge triples. We empirically demonstrate the effectiveness of the proposed *KCMC* model compared to several baselines [15], [16], [18] on two multi-turn dialogue dataset.

In summary, our main contributions are illustrated below:

- We propose to jointly take into account dialogue context and associated knowledge in a unified framework in open-domain dialogue model. Considering dialog context and knowledge together might yield mutually reinforcing advantages, and our *KCMC* model thus can respond more appropriately and informatively.
- We present Knowledge Fusion module to augment the semantic of word representation. The module encodes the relevant knowledge triples as a whole graph rather than separately, from which the model can understand the semantic of a word from its neighboring entities and relations.
- Extensive experimental results on two datasets demonstrate that our proposed *KCMC* model outperforms various competitive baselines, and it is able to generate appropriate and informative responses.

II. Related Works

A. End-to-End open-domain conversation

Earlier work on open-domain conversation treated the response generation as statistical machine translation, where the goal of task is to generate a proper response given the previous dialogue turn [19]. Since sequence-to-sequence models [1] have been successfully applied to large-scale conversation generation, various models [12], [20], [21] under an encoder-decoder framework have been proposed to improve generation quality of response from different perspectives such as diversity and relevance. [12] proposed Neural Responding Machine (NRM) for one-round short-text conversation, and NRM formalizes the generation of response as a decoding process based on the latent representation of the input text. [20] presented a simple end-to-end approach for the conversational modeling task using the sequence to sequence framework, and they also found that the lack of consistency is a common failure of conversational models. While most effort of these studies is paid to single-turn conversation, they do not take into account that representing conversational context is vital to response generation.

B. Context-aware conversation

Some researchers have taken into account conversational context and proposed response generation models for generating more appropriate and consistent response. In order to model the dialogue context better, [15] proposed hierarchical recurrent encoder-decoder networks (HRED), which combines two level RNNs and employs hierarchical encoders to model the structural information of dialogue context. In order to improve the diversity of generated responses, HRED was extend with a latent variable in the VHRED approach [22]. [22] proposed a neural network-based generative architecture with latent stochastic variables that span a variable number of time steps and found that the latent variables facilitate the generation of long outputs and maintain the context. [23] proposed Hierarchical Recurrent Attention Network (HRAN) to simultaneously model the hierarchy of contexts and the importance of words and utterances in a unified framework.

C. Conversation with unstructured texts

With availability of a large amount of knowledge texts, the integration of unstructured knowledge text into the generated responses has become a research hotspot. [24] proposed a neural knowledge diffusion (NKD) model to introduce knowledge into dialogue generation. The NKD model not only matches the relevant facts for the input utterance but diffuses them to similar entities. [25] created a new dataset containing movie chats wherein each response is explicitly generated by copying or modifying sentences from unstructured background knowledge such as comments about the movie. [16] generalized the widely-used Seq2Seq approach by conditioning responses on both dialogue history and external knowledge facts. However, these models largely depend on the quality of unstructured knowledge text, which may introduce noise to the generated responses.

D. Conversation with knowledge graph

There are growing interests in leveraging factoid knowledge or structured knowledge. [26] incorporated background knowledge for conversational model through a specially designed Recall gate. [18] designed a dynamic knowledge enquirer which selects different answer entities as different positions in a single response according different local context. [27] integrated commonsense knowledge into the dialogue model

and investigate the impact of providing commonsense knowledge about the concepts covered in the dialog. [28] encoded the knowledge graphs with a static graph attention and facilitated better generation through a dynamic graph attention. In comparison with these methods, we select knowledge triples more properly and incorporate knowledge information more effectively through the Knowledge Fusion module and hierarchical attention mechanism. Moreover, we jointly take into account dialogue context and associated knowledge in a unified framework which enables better semantic understanding and responses generation.

III. Description of The Conversational Model

A. Problem Definition and Overview

Considering a dialogue as a sequence of M utterances $X = \{U_1, U_2, \cdots, U_M\}$ involving two interlocutors. Each utterance U_m contains a sequence of N_m tokens, i.e. $U_m = \{x_{m,1}, x_{m,2}, \cdots, x_{m,N_m}\}$, where $x_{m,n}$ represents the token at position n in utterance m. The dialogue is accompanied by a set of relevant knowledge graphs $G = \{G_1, G_2, \cdots, G_M\}$, each knowledge graph $G_m = \{g_{m,1}, g_{m,2}, \cdots, g_{m,N_m}\}$, where $g_{m,n}$ corresponding to the knowledge graph of $x_{m,n}$. Each graph $g_{m,n}$ consists of $N_{g_{m,n}}$ knowledge triples $g_{m,n} = \{\tau_1, \tau_2, \cdots, \tau_{N_{g_{m,n}}}\}$ and each triple containing head entity, relation and tail entity is denoted as $\tau = \{h, r, t\}$. The objective of the task is to build a dialog system that can generate informative and coherent response $Y = \{y_1, y_2, \cdots, y_T\}$ based on dialogue history X and relevant knowledge graphs G.

The overview of our proposed *KCMC* model is presented in Fig. 1. Before entering the hierarchical encoder, the Knowledge Fusion module is firstly utilized to enhance the semantic representation of dialogue context with associated knowledge triples. Then the model uses a hierarchical encoder to model the hierarchical structure of dialogue context, word level attention and sentence level attention are utilized to attend on important part of dialogue context. Furthermore, the knowledge-enhanced decoder attentively read knowledge graphs and the triples in each graph with graph level attention and triple level attention respectively, which improves the process of knowledge selection and response generation. Finally, during the decoding process, the decoder either generates a generic word from a fixed vocabulary or copies a token from dialog context or knowledge triples.

B. Knowledge Fusion module

The relevant knowledge is important to the semantic understanding of the dialogue context. We thus base the response generation on associated knowledge to understand the dialogue context better. Before entering the encoder, the Knowledge Fusion module is designed to enhance the semantic representation of word with associated knowledge triples, as shown in Fig. 2.

To be more specific, we first retrieve the associated knowledge triples with each word in dialogue context from the knowledge base and then encode the retrieved knowledge

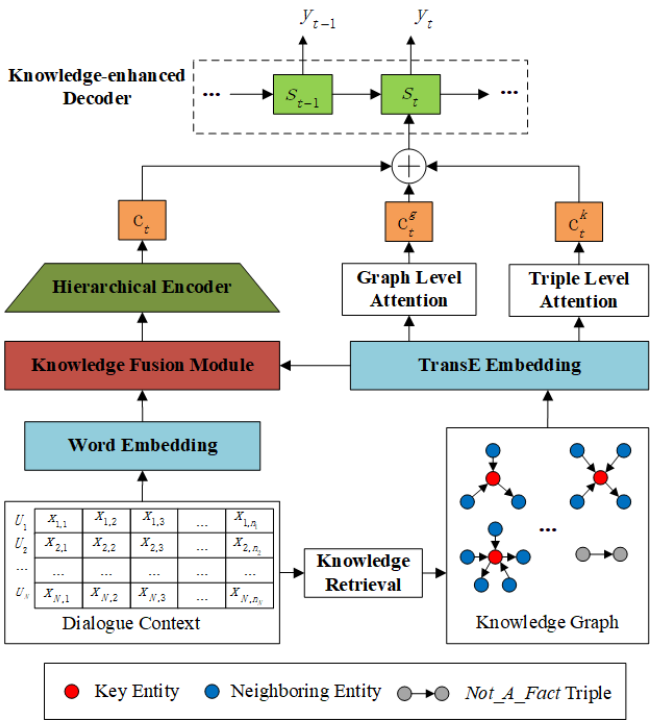

Fig. 1. Overview of *KCMC* model.

triples as a whole graph with static graph attention mechanism. Furthermore, the encoded knowledge representation is integrated with the word embedding to augment the semantic representation of word. We discuss the Knowledge Fusion module in detail in the following, including knowledge retrieval, knowledge encoding as well as knowledge combination.

Knowledge Retrieval - Knowledge Retrieval is responsible for retrieving associated knowledge triples for each word in dialogue context from the knowledge base ConceptNet [29], which is a large-scale structural knowledge graph in English. Specially, we use each word $x_{m,n}$ in dialogue context as the key entity to retrieve a knowledge graph $g_{m,n} = \{\tau_1, \tau_2, \cdots, \tau_{N_{g_{m,n}}}\}$ from the entire knowledge base. For common words which match no entity, a special knowledge graph *Not-A-Fact* is used.

Knowledge encoding - We adopt TransE [30] to represent the entities and relations in the knowledge triples. A full-connected layer is used to bridge the semantic gap between knowledge triple and unstructured knowledge text. A knowledge triple $\tau = \{h, r, t\}$ (head entity, relation, tail entity) is represented by the following formulation:

$$\boldsymbol{k} = (\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}) = MLP(TransE(h, r, t)) \tag{1}$$

where $\boldsymbol{h}, \boldsymbol{r}, \boldsymbol{t}$ are the transformed TransE embeddings for h, r, t respectively, and MLP stands for multilayer perceptron.

The static graph attention mechanism is designed to generate a static representation for a knowledge graph. Specially, for knowledge triples vectors $\boldsymbol{K}(g_{m,n}) = \{\boldsymbol{k_1}, \boldsymbol{k_2}, \cdots, \boldsymbol{k}_{N_{g_{m,n}}}\}$

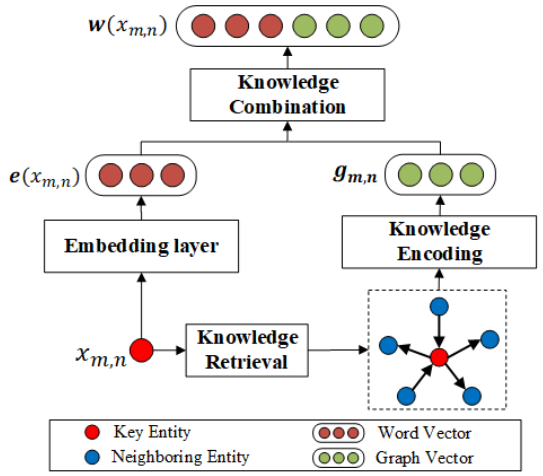

Fig. 2. The knowledge fusion module.

in the retrieved knowledge graph $g_{m,n}$, a graph representation $\boldsymbol{g_{m,n}}$ is calculated as follows:

$$\boldsymbol{g_{m,n}} = \sum_{i=1}^{N_{g_{m,n}}} \alpha_i[\boldsymbol{h_{m,n}; t_{m,n}}] \tag{2}$$

$$\alpha_i = softmax((W_r \boldsymbol{r_{m,n}})^\top tanh(W_e[\boldsymbol{h_{m,n}; t_{m,n}}])) \tag{3}$$

where $(\boldsymbol{h_{m,n}, r_{m,n}, t_{m,n}}) = \boldsymbol{k_{m,n}}$, and W_r, W_e are trainable weight matrices for relations and entities respectively.

Knowledge combination - After computing the knowledge graph representation $\boldsymbol{g_{m,n}}$ using the static graph attention mechanism, we concatenate $\boldsymbol{g_{m,n}}$ with the word embedding $\boldsymbol{e}(x_{m,n})$ to augment the semantic of the word $x_{m,n}$:

$$\boldsymbol{w}(x_{m,n}) = [\boldsymbol{e}(x_{m,n}); \boldsymbol{g_{m,n}}] \tag{4}$$

where $[\cdot;\cdot]$ is vector concatenation operation. The concatenated vector $\boldsymbol{w}(x_{m,n})$ is then fed to the word level encoder.

C. Hierarchical Encoder

Word Level Encoder - We first employ a bidirectional recurrent neural network with Long-Short Term Memory (Bi-LSTM) [31] to encode words $\{\boldsymbol{w}(x_{m,n})\}_{n=1}^{N_m}$ in utterance U_m as word-level hidden states $\{h_{m,n}\}_{n=1}^{N_m}$ as follows:

$$h_{m,n} = BiLSTM(\boldsymbol{w}(x_{m,n})) = [h^f_{m,n}; h^b_{m,n}] \tag{5}$$

where $h^f_{m,n}$, $h^b_{m,n}$ are the hidden states of a forward LSTM [32] and a backward LSTM respectively.

Suppose that decoder has hidden state s_{t-1} at last time step $t-1$, word level attention takes as input the word level hidden states $\{h_{m,j}\}_{j=1}^{N_m}$ and represent utterance U_m as utterance vector $r_{m,t}$ as follows:

$$r_{m,t} = \sum_{j=1}^{N_m} \alpha^w_{j,t} h_{m,j} \tag{6}$$

$$\alpha^w_{j,t} = softmax(h^\top_{m,j} W_w s_{t-1}) \tag{7}$$

where W_w is a trainable weight matrix. The word level attention weights $\{\alpha^w_{j,t}\}_{j=1}^{N_m}$ measure the importance of words in utterance U_m.

Utterance Level Encoder - Utterance vectors $\{r_{m,t}\}_{m=1}^{M}$ are then fed to the utterance level encoder which adopt a unidirectional LSTM and transformed to $\{l_{1,t}, l_{2,t}, \cdots, l_{i,t}, \cdots, l_{M,t}\}$ as hidden vectors of the context. After that, utterance level attention is utilized to calculate a context vector c_t as follows:

$$c_t = \sum_{i=1}^{M} \alpha^u_{i,t} l_{i,t} \tag{8}$$

$$\alpha^u_{i,t} = softmax(l^\top_{i,t} W_u s_{t-1}) \tag{9}$$

where W_u is a trainable weight matrix. The utterance level attention weights $\{\alpha^u_{i,t}\}_{i=1}^{M}$ measures the importance of M utterances.

D. Knowledge-enhanced Decoder

The knowledge enhanced decoder adopts a unidirectional LSTM, and the decoder updates its hidden state based on context vector c_t and relevant knowledge graphs. Dynamic graph attention mechanism is designed to attentively read all the knowledge graphs and then attentively reads all the triples in each graph. Specially, we first attend on the knowledge graph vectors $\{\boldsymbol{g_{m,n}}\}_{m=1,n=1}^{M,N_m}$ computed in (2) to get the graphs context vector as follows:

$$c^g_t = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \alpha^u_{m,t} \alpha^g_{n,t} \boldsymbol{g_{m,n}} \tag{10}$$

$$\alpha^g_{n,t} = softmax(V_b^\top tanh(W_b s_{t-1} + U_b \boldsymbol{g_{m,n}})) \tag{11}$$

where V_b, W_b, U_b are trainable parameters. The utterance level attention weight $\alpha^u_{m,t}$ computed in (11) measures the importance of utterance U_m. And the graph level attention weight $\alpha^g_{n,t}$ measures the importance of knowledge graphs $\{g_{m,n}\}_{n=1}^{N_m}$ corresponding to the utterance U_m at step t. The graphs context vector c^g_t is the weighted sum of the graph vectors in each utterance.

The model then attends on the knowledge triple vectors $\boldsymbol{K}(g_{m,n}) = \{\boldsymbol{k}_{m,n,1}, \cdots, \boldsymbol{k}_{m,n,l}, \cdots, \boldsymbol{k}_{m,n,N_{g_{m,n}}}\}$ within each knowledge graph $g_{m,n}$ to get the triples context vector c^k_t as follows:

$$c^k_t = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \sum_{l=1}^{N_{g_{m,n}}} \alpha^u_{m,t} \alpha^g_{n,t} \alpha^k_{l,t} \boldsymbol{k}_{m,n,l} \tag{12}$$

$$\alpha^k_{l,t} = softmax(\boldsymbol{k}^\top_{m,n,l} W_k s_{t-1}) \tag{13}$$

Where W_k is a trainable weight matrix. The triple level attention weight $\alpha^k_{l,t}$ measures the importance of knowledge triples $\{\tau_{m,n,l}\}_{l=1}^{N_{g_{m,n}}}$ within graph $g_{m,n}$ at step t. The triples context vector c^k_t is the weighted sum of the triple vectors in each graph.

Finally, the knowledge-enhanced decoder updates its hidden state based on both dialogue context information and relevant knowledge information as follows:

$$s_t = LSTM(s_{t-1}, [c_t; c^g_t; c^k_t; \boldsymbol{w}(y_{t-1})]) \tag{14}$$

where c_t, c_t^g, c_t^k are computed in (10),(13),(16) respectively.

E. Response Generation

During the decoding process, the model generates the response word by word. At time step t, the decoder either generates a word from the fixed vocabulary or copy a token from one of the knowledge triples memory and the dialogue context memory. Two soft gate mechanism are utilized to integrate generation mode and copy mode.

Generating words - Similar to [4], the decoder selects a generic word from fixed vocabulary at step t by the following probability produced by a softmax function:

$$P_g(y_t) = softmax(W_1[s_t; c_t; c_t^g; c_t^k] + b_1) \tag{15}$$

Copying words from dialogue context - The product of first utterance level attention weight α_m^u and second word level attention weight $\alpha_{m,n}^w$ gives the final attention scores of all tokens in dialogue context. Similar to [33], the final attention scores are used as the probability scores to form the copy distribution $P_{context}(y_t)$ over the dialogue context:

$$P_{context}(y_t) = \sum_m \sum_n \alpha_m^u \alpha_{m,n}^w \tag{16}$$

Copying words from knowledge triples - The product of first utterance level attention weight α_m^u, second graph level attention weight $\alpha_{m,n}^g$ and third triple level attention weight $\alpha_{m,n,l}^k$ forms the copy distribution $P_{kb}(y_t)$ over the tail entities in all knowledge triples:

$$P_{kb}(y_t) = \sum_m \sum_n \sum_l \alpha_m^u \alpha_{m,n}^g \alpha_{m,n,l}^k \tag{17}$$

Decoding - Similar to [34], soft gate mechanism is utilized to combine the generation and copy distributions. Specially, we use soft gate γ_1 to obtain the copy distribution $P_c(y_t)$ by combining $P_{kb}(y_t)$ and $P_{context}(y_t)$ as follows:

$$\gamma_1 = sigmoid(W_2[s_t; c_t; c_t^g; c_t^k] + b_2) \tag{18}$$

$$P_c(y_t) = \gamma_1 P_{kb}(y_t) + (1 - \gamma_1) P_{context}(y_t) \tag{19}$$

Finally, another soft gate γ_2 is utilized to obtain the final output distribution $P(y_t)$ by combining generation distribution $P_g(y_t)$ and copy distribution $P_c(y_t)$ as shown below:

$$\gamma_2 = sigmoid(W_3[s_t; c_t; c_t^g; c_t^k] + b_3) \tag{20}$$

$$P(y_t) = \gamma_2 P_g(y_t) + (1 - \gamma_2) P_c(y_t) \tag{21}$$

Where W_2, W_3, b_2, b_3 are trainable parameters and $\gamma_1 \in [0, 1]; \gamma_2 \in [0, 1]$.

We train the model by minimizing the cross entropy $-\sum_{t=1}^{T} P_t log(P(y_t))$ between the predicted distribution $P(y_t)$ and the reference distribution P_t.

TABLE I
STATISTICS OF TWO DATASETS AND THE KNOWLEDGE BASE

ConceptNet		**DailyDialog**		**CONVAI2**	
Entity	21,471	Training	11,118	Training	15,878
Relation	44	Validation	1,000	Validation	1,000
Triple	120,850	Test	1,000	Test	1,000

IV. EXPERIMENTS

In this section, we firstly describe the statistical details of two datasets, and we also introduce the details about the experiment setup and implementation. Then we make a brief introduction of compared baseline models for open-domain multi-turn conversation and the evaluation metrics. Finally, we conduct experiments on the aforementioned datasets to evaluate the performance of our proposed *KCMC* model. Experimental results demonstrate that our proposed *KCMC* model outperforms all the compared baselines and is able to generate both appropriate and informative response.

A. Dataset

Knowledge Base - ConceptNet [29] is a graph-structured knowledge base that connect words and phrases of natural language with labeled edges and is designed to represent the general knowledge involved in understanding language. ConceptNet concisely represents knowledge assertion as a triple containing head entity, relation label and tail entity. For example, the knowledge assertion that "a dog has a tail" can be represented as (dog, HasA, tail). For simplicity, after removing triples containing multi-word entities, 120,580 triples are retained with 21,471 entities and 44 relations.

Multi-turn Conversation Dataset - We present our experiments using two real-world publicly available multi-turn dialogue dataset DailyDialog [35] and the recently released CONVAI2 conversational AI challenge dataset [36]. DailyDialog is a high-quality open-domain dialog dataset which consists of dialogues that resemble day-to-day life. The dialogues in the dataset cover various topics about our daily life and the language in the dialogue is human-written and less noisy. It comprises of 13k dialogs with average 7.9 turns per dialog. CONVAI2 is an extended version of PERSONACHAT [37] which is an open domain dataset with multi-turn chit-chat conversations between turkers who are each assigned to a "persona" at random. It contains 17.8k dialogs with an average of 14.8 turns per dialog. The statistics for two datasets and the knowledge base can be seen in Table 1.

B. Implementation Details

We implement our knowledge-based context-aware multi-turn conversational model based on the code released by [28] and that by [34]. To be more specific, the word encoder, utterance encoder and decoder are 2-layer LSTMs with 512 hidden units for each layer. We use TransE [30] to initialize

https://github.com/tuxchow/ccm
https://github.com/DineshRaghu/multi-level-memory-network

TABLE II
COMPARISON OF OUR MODEL WITH BASELINES

	DailyDialog dataset			**CONVAI2 dataset**		
Model	perplexity	BLEU	entity score	perplexity	BLEU	entity score
Attn seq2seq [1]	48.19	2.18	0.717	48.23	2.21	0.738
HRED [15]	45.43	2.76	0.839	45.32	2.80	0.847
MemNet [16]	43.78	3.07	0.913	42.19	3.23	0.994
CopyNet [18]	39.54	3.61	1.181	40.06	3.56	1.138
KCMC model	**38.35**	**3.92**	**1.270**	**39.14**	**3.89**	**1.172**

entity and relation representations in knowledge triples, and the embedding size for TransE is set as 100. We set the dimension of word embeddings to 300. The word embeddings are randomly initialized and updated during the training process. We build our vocabulary by keeping words that appear more than three times in the dataset, and replacing words that appear less than three times with a special 'UNK' token. What's more, the vocabulary size is limited to 30,000. The vocabulary and word embeddings are shared by the encoder and decoder. In order to avoid over-fitting, we employ a dropout of 0.1 to each RNN cell during the training process. We apply gradient clipping to 5.0 when its norm exceeds this value. And we adopt simple greedy search without any re-scoring techniques. The Adam optimizer is used to train our model with a mini-batch size of 32 and a learning rate of 0.001. We implement the model with an open source deep learning tool Tensorflow.

C. Baselines

We compare our system with a set of carefully selected baselines, shown as follows:

- **Attn seq2seq:** A sequence-to-sequence model [1] with simple attention [2] over the input context, which is widely used in open domain dialogue system.
- **HRED:** It is a hierarchical recurrent encoder-decoder model proposed by [15], which employs hierarchical encoders to model the structural information of dialogue context.
- **MemNet:** It is an end-to-end knowledge-grounded generative model [16], where the TranE embeddings of knowledge triples are stored into the memory units.
- **CopyNet:** The model [18] augments the sequence-to-sequence architecture with attention-based copy mechanism over the input context which generates a generic word from the vocabulary or copies an entity from knowledge triples.

D. Evaluation Metrics

- ***perplexity*:** The perplexity metric [14] is used to evaluate our model at the content level which measures how well a model predicts human responses. Smaller perplexity scores indicate that the model can generate more grammatical and fluent responses.
- ***BLEU*:** BLEU [38] analyzes the co-occurrences of n-grams in the ground truth and the generated response. The BLEU metric measures how similar the candidate text is to the ground truth, with larger values representing more similar to the ground truth.
- ***entity score*:** Following [28], we also employ entity score as an evaluation metric. This metric denotes the number of entities per response which measures the model's ability to select concepts from the associated knowledge triples when generating response.

https://github.com/tensorflow/tensorflow

E. Results

Table 2 shows the performance comparison of each model on the unbiased test set. As can be seen, the *KCMC* model outperforms other baseline models on perplexity, BLEU and *entity score*. Note that the performance of the Attn Seq2Seq model is extremely low, since the model does not take into account the structure of dialogue context and produces only short and repetitive responses. The HRED model can generate more fluent responses compared to the Attn Seq2Seq model. The HRED model employ hierarchical encoders to model the hierarchical information of dialogue context, demonstrating the importance of dialogue context to response generation. And the MemNet model combines knowledge information through memory network, which can generate more relevant responses. Furthermore, the CopyNet model is able to copy entity from the knowledge triple via the pointer network, and the model can generate more informative responses. Compared to aforementioned models, the *KCMC* model obtains the best performance on all metrics on both DailyDialog and CONVAI2 dataset. The *KCMC* model not only captures the hierarchical structure information of dialogue context through hierarchical encoders, but also effectively incorporates relevant knowledge information through the Knowledge Fusion module and Knowledge-enhanced decoder.

To be more specific, the model obtains the lowest perplexity on both DailyDialog and CONVAI2 dataset, demonstrating that the model can understand the semantic of dialogue context better and generate more appropriate and grammatical responses. Owing to the Knowledge Fusion module augmenting the semantic representation of word with associated knowledge triples, the model can understand the dialogue context better and thus respond more properly. What's more, the model generates the most entities from the graph-structured knowledge base among all the models, indicating that the model can utilize the external knowledge more effectively. Owing to the hierarchical attention mechanism and the Knowledge-enhanced decoder, the model can select relevant knowledge

TABLE III
COMPARING THE RESPONSES GENERATED BY VARIOUS MODELS ON AN EXAMPLE IN DAILYDIALOG

Dialog context	A: You look so tan and healthy! B: Thanks. I just got back from summer camp. A: How was it? B: Great. I got to try so many things for the first time. A: Like what?
Knowledge triples	(boys, Desires, camp); (camp, RelatedTo, experience); (camp, RelatedTo, campfire);(camp, RelatedTo, fishing); (camp, RelatedTo, forest); (camp, RelatedTo, gathering); (camp, RelatedTo, hiking); (camp, RelatedTo, kids); (camp, RelatedTo, lakes); (camping, IsA, activity)
Attn seq2seq	I do not know.
HRED	Yes, I'm sure it is.
MemNet	I like this summer camp.
CopyNet	I enjoyed fishing this summer camp.
KCMC	I went fishing and hiking. It was a great experience.

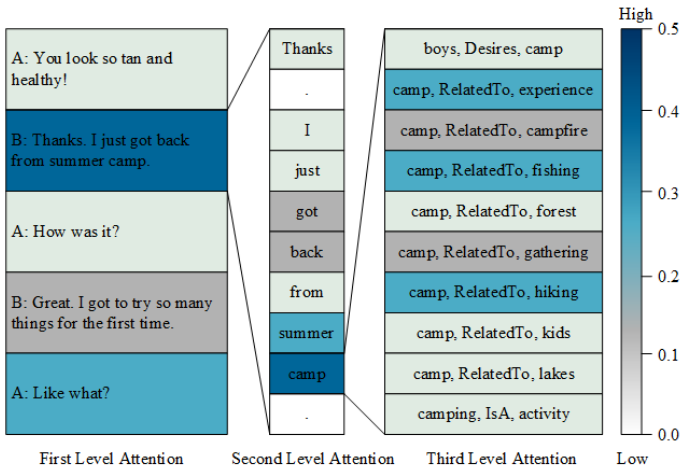

Fig. 3. Visualization of hierarchical attention over the knowledge triples

entities more properly and incorporate knowledge information more effectively.

F. Case Study

Table 3 lists a dialogue sample from the DailyDialog dataset to compare *KCMC* with other baselines. Without modeling the structural information of dialogue context, the Attn Seq2Seq model is unable to understand the semantic of dialog context well and only generate generic response, which is short and dull. The HRED model also generate a short response without the help of external relevant knowledge, which demonstrates the importance of relevant knowledge to generate a meaningful and informative response. The MemNet model can generate some meaningful words as it reads relevant knowledge triple embeddings in its memory, which demonstrates that relevant knowledge can facilitate the understanding of dialogue context. The CopyNet model can read and copy words from knowledge triples, but it generates fewer entity words than the *KCMC* model. The *KCMC* model can not only understand semantic of the dialogue context better with the Knowledge Fusion module, but also select knowledge triples more properly and copy more entities from knowledge triples with hierarchical attention. This dialogue sample shows that *KCMC* model can generate more appropriate and informative responses than other baselines.

G. Attention Visualization

Analyzing the hierarchical attention weights can help us better understand how the model generates a response. The visualization of hierarchical attention over the knowledge triples while generating a response for the dialogue example in Table 3 is shown in Fig. 3. For the first level attention, the second utterance gets the highest attention weight among the dialogue context. This indicates that the model figures out that the last utterance was talking about the 'summer camp' and the generated response should be related to the 'summer camp'. And for the second level attention, the word 'camp' gets the highest attention weight among all words in the second utterance. Then the model will copy entities from the knowledge triples related to the word 'camp'. Specially, for the third level attention, three knowledge triples ("camp, RelatedTo, experience"; "camp, RelatedTo, fishing"; "camp, RelatedTo, hiking") get the first three highest attention weights among all triples. The *KCMC* model copies three tail entities (experience, fishing, hiking) from the three knowledge triples while generating a response. This suggests that in this case the *KCMC* model attends to the right utterances and words and knowledge triples, and the model works well as we have expected.

The visualization of hierarchical attention of this dialogue example demonstrates that the *KCMC* model not only selects relevant knowledge triples more properly given the whole dialogue context, but copies more entities and incorporate knowledge more effectively. The *KCMC* model is capable of generating informative and coherent response through better use of knowledge.

CONCLUSIONS

In this paper, we propose a *KCMC* model with hierarchical attention mechanism for open-domain multi-turn dialogue response generation. Our approach jointly takes into account dialogue context and relevant structured knowledge in a unified framework. The *KCMC* model can understand the dialogue context better and thus respond more properly with the help of Knowledge Fusion module which augment the semantic representation of word. The proposed model selects relevant knowledge more properly given the whole dialogue context owing to the hierarchical attention mechanism. And the *KCMC* model can copy entity from knowledge triples and incorporate knowledge information effectively with knowledge-enhanced decoder. We demonstrate the effectiveness of our approach through experiments in comparison with several baselines on DailyDialog dataset and CONVAI2 dataset. Extensive experimental results demonstrate that our model can generate more appropriate and informative responses than state-of-the-art baselines.

ACKNOWLEDGE

This work was supported by National Natural Science Foundation of China (61702047) .

REFERENCES

[1] I. Sutskever, O. Vinyals, and Q. Le, "Sequence to sequence learning with neural networks," *Advances in NIPS*, 2014.

[2] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[3] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," *arXiv preprint arXiv:1406.1078*, 2014.

[4] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[5] A. M. Rush, S. Chopra, and J. Weston, "A neural attention model for abstractive sentence summarization," *arXiv preprint arXiv:1509.00685*, 2015.

[6] R. Nallapati, B. Zhou, C. Gulcehre, B. Xiang *et al.*, "Abstractive text summarization using sequence-to-sequence rnns and beyond," *arXiv preprint arXiv:1602.06023*, 2016.

[7] A. See, P. J. Liu, and C. D. Manning, "Get to the point: Summarization with pointer-generator networks," *arXiv preprint arXiv:1704.04368*, 2017.

[8] J. Tan, X. Wan, and J. Xiao, "Abstractive document summarization with a graph-based attentional neural model," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1171–1181.

[9] B. Kim, H. Kim, and G. Kim, "Abstractive summarization of reddit posts with multi-level memory networks," *arXiv preprint arXiv:1811.00783*, 2018.

[10] C. Xiong, V. Zhong, and R. Socher, "Dynamic coattention networks for question answering," *arXiv preprint arXiv:1611.01604*, 2016.

[11] W. Wang, N. Yang, F. Wei, B. Chang, and M. Zhou, "Gated self-matching networks for reading comprehension and question answering," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 189–198.

[12] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," *arXiv preprint arXiv:1503.02364*, 2015.

[13] A. Baheti, A. Ritter, J. Li, and B. Dolan, "Generating more interesting responses in neural conversation models with distributional constraints," *arXiv preprint arXiv:1809.01215*, 2018.

[14] I. V. Serban, A. Sordoni, Y. Bengio, A. Courville, and J. Pineau, "Hierarchical neural network generative models for movie dialogues," *arXiv preprint arXiv:1507.04808*, vol. 7, no. 8, 2015.

[15] ——, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[16] M. Ghazvininejad, C. Brockett, M.-W. Chang, B. Dolan, J. Gao, W.-t. Yih, and M. Galley, "A knowledge-grounded neural conversation model," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[17] S. Han, J. Bang, S. Ryu, and G. G. Lee, "Exploiting knowledge base to generate responses for natural language dialog listening agents," in *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 2015, pp. 129–133.

[18] W. Zhu, K. Mo, Y. Zhang, Z. Zhu, X. Peng, and Q. Yang, "Flexible end-to-end dialogue system for knowledge grounded conversation," *arXiv preprint arXiv:1709.04264*, 2017.

[19] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics, 2011, pp. 583–593.

[20] O. Vinyals and Q. Le, "A neural conversational model," *arXiv preprint arXiv:1506.05869*, 2015.

[21] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan, "A diversity-promoting objective function for neural conversation models," *arXiv preprint arXiv:1510.03055*, 2015.

[22] I. V. Serban, A. Sordoni, R. Lowe, L. Charlin, J. Pineau, A. Courville, and Y. Bengio, "A hierarchical latent variable encoder-decoder model for generating dialogues," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[23] C. Xing, Y. Wu, W. Wu, Y. Huang, and M. Zhou, "Hierarchical recurrent attention network for response generation," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[24] 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 (Volume 1: Long Papers)*, 2018, pp. 1489–1498.

[25] N. Moghe, S. Arora, S. Banerjee, and M. M. Khapra, "Towards exploiting background knowledge for building conversation systems," *arXiv preprint arXiv:1809.08205*, 2018.

[26] Z. Xu, B. Liu, B. Wang, C. Sun, and X. Wang, "Incorporating loose-structured knowledge into lstm with recall gate for conversation modeling," *arXiv preprint arXiv:1605.05110*, vol. 3, 2016.

[27] T. Young, E. Cambria, I. Chaturvedi, H. Zhou, S. Biswas, and M. Huang, "Augmenting end-to-end dialogue systems with commonsense knowledge," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[28] H. Zhou, T. Young, M. Huang, H. Zhao, J. Xu, and X. Zhu, "Commonsense knowledge aware conversation generation with graph attention." in *IJCAI*, 2018, pp. 4623–4629.

[29] R. Speer, J. Chin, and C. Havasi, "Conceptnet 5.5: An open multilingual graph of general knowledge," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[30] A. Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Advances in neural information processing systems*, 2013, pp. 2787–2795.

[31] Z. Huang, W. Xu, and K. Yu, "Bidirectional lstm-crf models for sequence tagging," *arXiv preprint arXiv:1508.01991*, 2015.

[32] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[33] C. Gulcehre, S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio, "Pointing the unknown words," *arXiv preprint arXiv:1603.08148*, 2016.

[34] R. Reddy, D. Contractor, D. Raghu, and S. Joshi, "Multi-level memory for task oriented dialogs," *arXiv preprint arXiv:1810.10647*, 2018.

[35] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," *arXiv preprint arXiv:1710.03957*, 2017.

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

[37] S. Zhang, E. Dinan, J. Urbanek, A. Szlam, D. Kiela, and J. Weston, "Personalizing dialogue agents: I have a dog, do you have pets too?" *arXiv preprint arXiv:1801.07243*, 2018.

[38] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, 2002, pp. 311–318.