Knowledge-Aware Dialogue Generation via Hierarchical Infobox Accessing and Infobox-Dialogue Interaction Graph Network

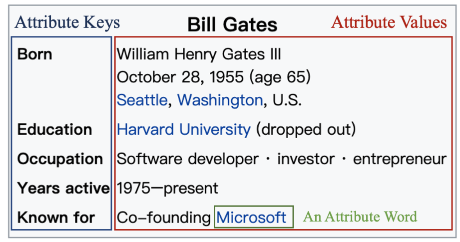
Sixing Wu[1], **Minghui Wang**[2], **Dawei Zhang**[1], **Yang Zhou**[3], **Ying Li**[4,5*] and **Zhonghai Wu**[4,5]
[1]School of Electronics Engineering and Computer Science, Peking University, Beijing, China
[2]School of Software and Microelectronics , Peking University, Beijing, China
[3]Auburn University, Auburn, Alabama, USA
[4]National Research Center of Software Engineering, Peking University, Beijing, China
[5]Key Lab of High Confidence Software Technologies (MOE), Peking University, Beijing, China
{wusixing, minghui_wang, daweizhang, li.ying, wuzh}@pku.edu.cn, yangzhou@auburn.edu

Abstract

Due to limited knowledge carried by queries, traditional dialogue systems often face the dilemma of generating boring responses, leading to poor user experience. To alleviate this issue, this paper proposes a novel infobox knowledge-aware dialogue generation approach, HITA-Graph, with three unique features. First, open-domain infobox tables that describe entities with relevant attributes are adopted as the knowledge source. An order-irrelevance Hierarchical Infobox Table Encoder is proposed to represent an infobox table at three levels of granularity. In addition, an Infobox-Dialogue Interaction Graph Network is built to effectively integrate the infobox context and the dialogue context into a unified infobox representation. Second, a Hierarchical Infobox Attribute Attention mechanism is developed to access the encoded infobox knowledge at different levels of granularity. Last but not least, a Dynamic Mode Fusion strategy is designed to allow the Decoder to select a vocabulary word or copy a word from the given infobox/query. We extract infobox tables from Chinese Wikipedia and construct an infobox knowledge base. Extensive evaluation on an open-released Chinese corpus demonstrates the superior performance of our approach against several representative methods.

1 Introduction

Open-domain dialogue systems aim to generate human-like responses; however, the inadequate knowledge carried by the queries dramatically constrains the ability of dialogue systems to understand the intrinsic semantics of the queries and generate user-friendly dialogues [Ghazvininejad *et al.*, 2018]. The frequently generated boring responses (e.g., " I don't know") often lead to frustrating user experience [Li *et al.*, 2016]. Recently, researchers have recognized that introducing external knowledge to dialogue generation systems has great potential for further improving performance. Knowledge is regarded as the awareness and understanding of the dialogue context [Yu *et al.*, 2020]; namely, it can effectively assist the dialogue systems in understanding the intrinsic semantics and fabricating informative responses.



Bill Gates	
Born	William Henry Gates III October 28, 1955 (age 65) Seattle, Washington, U.S.
Education	Harvard University (dropped out)
Occupation	Software developer · investor · entrepreneur
Years active	1975–present
Known for	Co–founding Microsoft

Figure 1: The 'Bill Gates' Infobox from Wikipedia.

Dialogue systems can gather knowledge from various information sources, which can be broadly classified into (1) Knowledge texts, which can provide rich semantic information, are easy to access on the Internet, such as online encyclopedia (e.g., Wikipedia and Answers.com), news websites, and search engines [Tam, 2020]; (2) Knowledge bases, such as knowledge graphs [Zhang *et al.*, 2020], and spread-sheet tables [Qin *et al.*, 2019]; their knowledge items are well-organized, and thus can be easily integrated into dialogue systems; and (3) Topics and keywords, which can promote the dialogue towards a specific and consistent direction [Wang *et al.*, 2018; Wu *et al.*, 2020b]. However, knowledge texts are unstructured, knowledge bases are hard to construct/collect, and topic words/keywords are not informative. Therefore, there is a question, can we utilize a new type of knowledge that is easy-collected, informative, structured, and consistent for further enhancing dialogue systems?

Recently, a new kind of knowledge, infobox tables, has attracted much attention in the data-to-text tasks [Bao *et al.*, 2018; Chen *et al.*, 2020]. Figure 1 provides an illustrating example of an infobox table that specifies an entity with multiple attribute key-values. Infobox knowledge integrates the advantages of the above three types of knowledge. First, massive infobox tables can be easily obtained from the Internet, such as Wikipedia articles or web pages of Google search results. Second, infobox tables are easy to use since infobox knowledge is well extracted and organized as informative attributes. Third, an infobox table focuses on one target en-

*Corresponding author.

tity without the interference of irrelevant entities, guaranteeing knowledge consistency. The infobox knowledge offers a great opportunity to employ one kind of reliable knowledge source to generate high-quality dialogue systems. Integrating the above three types of knowledge one by one often results in knowledge inconsistency and computational inefficiency.

To our best knowledge, using the infobox knowledge to improve the quality of open-domain dialogue generation has not been fully investigated yet. Compared to the usage of the infobox tables in the data-to-text tasks, conducting the infobox knowledge-aware dialogue generation is much more difficult to study. Unlike the data-to-text tasks that rephrase infobox attributes in an orderly fashion, dialogue generation tasks conduct attribute order-irrelevance knowledge selection, inference, and dialogue-infobox fusion. In addition, existing data-to-text studies usually only consider the infobox tables within limited domains. However, open-domain dialogue generation has to use infobox tables in a wide range of domains for maintaining enough knowledge coverage, which dramatically increases the difficulty of unambiguous knowledge representation. These challenges indicate that infobox knowledge-enhanced dialogue generation cannot directly utilize the techniques used in the existing data-to-text works.

With these challenges in mind, we propose **HITA-Graph**, a novel infobox knowledge-aware dialogue generation approach. (1) We propose an order-irrelevance Hierarchical Infobox Table Encoder (**HITE**) to represent an infobox table at three levels of granularity. To alleviate the sparsity of the word distribution of the open-domain infobox attributes, besides the usual word embedding, HITE further integrates the embedding learned from the char sequence of a word, the part-of-speech tag embedding, and the locally positional embedding into an intra-attribute level representation of infobox table. Subsequently, a multi-head attention network is adopted to compute an attribute-level representation. In the context-level, we propose an Infobox-Dialogue Interaction Graph Network (**IDCI-Graph**) to capture both the infobox and the dialogue context by building and inferring on an infobox-dialogue interaction graph. (2) We develop a Hierarchical Infobox Attribute Attention (**HIAA**), which allows the Decoder to access the encoded infobox knowledge in a coarse-to-fine manner. (3) When predicting the next target token, the Decoder is equipped with three modes: generating a vocabulary word, copying a word from the query, and copying an attribute word from the infobox. Such three modes are fused via the Dynamic Mode Fusion strategy (**DMF**).

We collected about 895K Chinese infobox tables from Wikipedia. Extensive experiments are conducted on a Chinese corpus [Cai *et al.*, 2019b]. Both the automatic and human evaluation results demonstrate that the proposed approach HITA-Graph outperforms the representative competitors in almost all experiments. Our contribution is three-fold: (1) To our best knowledge, this work is the first to study the infobox knowledge-aware dialogue generation; (2) A novel Hierarchical Infobox Table Encoder and a novel Infobox-Dialogue Interaction Graph Network are designed to integrate infobox knowledge. (3) A Hierarchical Infobox Attribute Attention mechanism is proposed to access the infobox knowledge at different levels of granularity.

2 Our Approach

2.1 Problem Formulation

Let $\mathcal{D} = \{\langle X, Y, T\rangle\}$ denote the corpus, where $X = (x_1, \cdots, x_n)$ is a query, $Y = (y_1, \cdots, y_m)$ is a response, and $T = \{\langle f^k, f^v\rangle\}^l$ is an infobox table that consists of a set of attribute key-values and can be retrieved from the infobox base $\mathcal{T}$. The goal of our task is: $Y^* = \arg\max_{Y'} P(Y'|X, T)$.

2.2 Context Encoder

A query X is firstly encoded into dialogue context states $\mathbf{H} = (\mathbf{h_1}, ..., \mathbf{h_n})$ by a bi-directional GRU [Cho *et al.*, 2014]:

$$\mathbf{h^f_t} = GRU(\mathbf{x_t}, \mathbf{h^f_{t-1}}); \mathbf{h^b_t} = GRU(\mathbf{x_{n-t+1}}, \mathbf{h^b_{t-1}}) \quad (1)$$

where $\mathbf{x}$ is the embedding of x, and $\mathbf{h_t}$ is the concatenation $[\mathbf{h^f_t}; \mathbf{h^b_t}]$. $\mathbf{h_n}$ is regarded as the context summary.

2.3 Hierarchical Infobox Table Encoder

For each attribute $f^{kv}_i = \langle f^k_i, f^v_i\rangle \in T$, its key f^k_i is an noun word, its value $f^v_i = (w_{i,1}, \cdots, w_{i,|f^v_i|})$ is a word sequence.

Intra-Attribute Encoding

For an attribute key f^k_i or a value word $w_{i,j}$, its word distribution is sparse; namely, there are rare words. If only represent them at the word-level, many words can not be recognized. Thus, we design a hybrid-level method. Taking f^k_i as an example, the corresponding embedding $\mathbf{f^{k,hybrid}_i}$ is given by:

$$\mathbf{f^{k,hybrid}_i} = MLP_{hf}(\mathbf{f^k_i}, \mathbf{f^{k,char}_i}) \quad (2)$$

where $\mathbf{f^k_i}$ is obtained by looking up the word-level embedding matrix, $\mathbf{f^{k,char}_i}$ is computed from the char sequence of f^k_i with the CharCNN Encoder [Kim *et al.*, 2016], and MLP_{hf} is a MLP network. The incorporation of chars can alleviate the issue of the sparse word distribution. Subsequently, each attribute f^{kv}_i is represented as a set of key-word embedding:

$$\mathbf{F^{kw}_i} = \{\mathbf{f^{kw}_{i,j}}\} = \{[\mathbf{f^{k,hybrid}_i}; \mathbf{w^{hybrid}_{i,j}}; \mathbf{w^{pos}_{i,j}}; \mathbf{p^f_{i,j}}; \mathbf{p^b_{i,j}}]\} \quad (3)$$

where $f^{kw}_{i,j}$ is the j-th key-word pair of f^{kv}_i, $w^{pos}_{i,j}$ is the part-of-speech tag of $w_{i,j}$, and $p^f_{i,j}$ and $p^b_{i,j}$ are local positions counted from the beginning (i.e., j) and the end (i.e., $|f^v_i| - j + 1$), respectively.

Then, the multi-head self-attention (denoted as MHA, [Vaswani *et al.*, 2017]) is used to compute the intra-attribute level representations $\mathbf{F^{kw,ia}_i} = \{\mathbf{f^{kw,ia}_{i,j}}\}$:

$$\mathbf{F^{kw,ia}_i} = MHA(\mathbf{Q} = \mathbf{F^{kw}_i}, \mathbf{K} = \mathbf{F^{kw}_i}, \mathbf{V} = \mathbf{F^{kw}_i}) \quad (4)$$

Attribute-Level Encoding

For each key-value attribute f^{kv}_i, its attribute-level embedding $\mathbf{f^{kv,a}_i}$ is the weighted sum of $\mathbf{F^{kw,ia}_i} = \{\mathbf{f^{kw,ia}_{i,j}}\}$:

$$\mathbf{f^{kv,a}_i} = \sum_j \frac{\exp(\mathbf{f_s}^\mathbf{T} \mathbf{f^{kw,ia}_{i,j}})}{\sum_k \exp(\mathbf{f_s}^\mathbf{T} \mathbf{f^{kw,ia}_{i,k}})} \mathbf{f^{kw,ia}_{i,j}} \quad (5)$$

where $\mathbf{f_s}$ is a learn-able parameter, serving as a special query to compute the weight; thus, the infobox table T can be represented as a set of attribute embedding $\mathbf{F^{kv,a}} = \{\mathbf{f^{kv,a}_i}\}$.

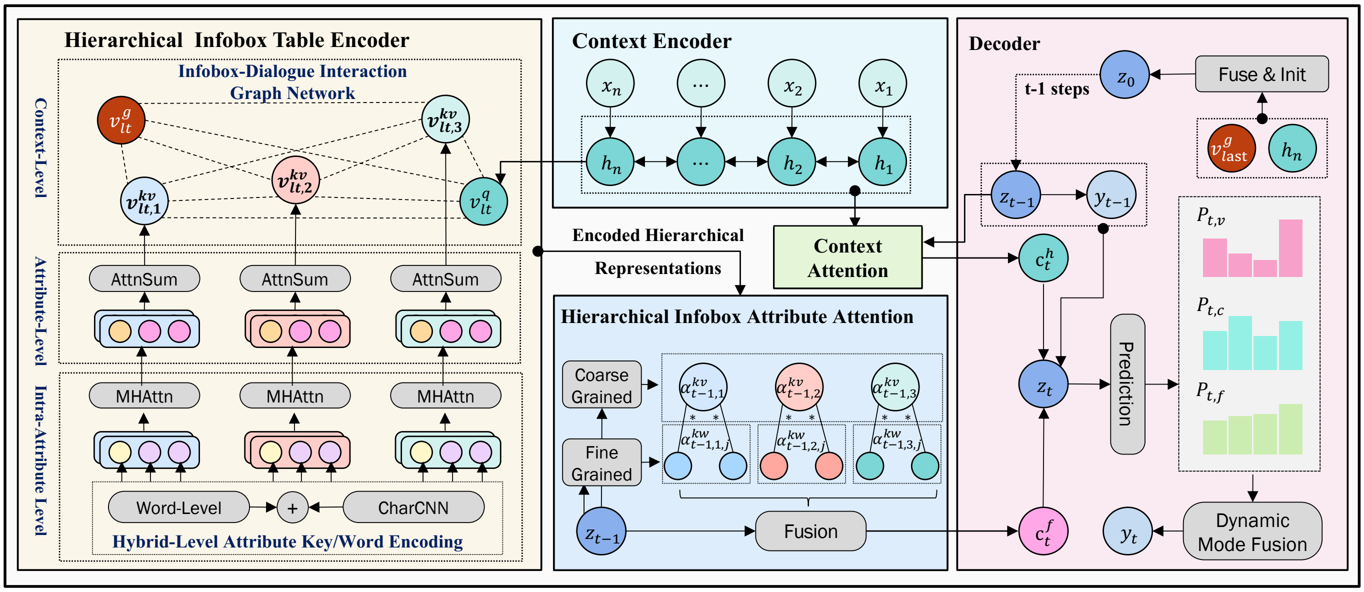


Figure 2: An overview of HITA-Graph. In the HITA-Graph, all operations would not be affected by the order of the given infobox attributes.

Infobox-Dialogue Interaction Graph Network

An attribute-level representation $\mathbf{f_i^{kv,a}}$ interacts with neither the dialogue context nor other attributes (i.e., the infobox context). Inspired by the success of GATs in knowledge-enhanced text generation [Zhang *et al.*, 2020], we design an Infobox-Dialogue Interaction Graph Network to capture the interaction information among attributes and the dialogue context. For each infobox-dialogue instance, we first construct a graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E}, \mathcal{R} \rangle$. The node set is $\mathcal{V} = \{v^{kv}\} \cup \{v^g\} \cup \{v^q\}$, where attribute nodes $\{v^{kv}\}$ correspond to the attributes $\{f^{kv}\}$ of T, v^g is a virtual node serving as the global node, and v^q is another virtual node representing the dialogue context. Meanwhile, $\mathcal{E} = \{(v_i, r_{i \to j}, v_j)\}$ is edge set, where $r_{i \to j} \in \mathcal{R}$ is the relational edge relation. For promoting the knowledge flow, $\mathcal{G}$ is fully-connected and directional in this paper. For each two nodes v_i, v_j, their directional edge $r_{i \to j}$ depends on the node types of v_i, v_j. Therefore, the relation set $\mathcal{R}$ consists of 3 self-connection types $\{self_{v^{kv}}, self_{v^g}, self_{v^q}\}$, 3 types starting from an attribute node $\{v_a^{kv} \to v_{b \neq a}^{kv}, v^{kv} \to v^g, v^{kv} \to v^q\}$, 2 types starting from the global node $\{v^g \to v^{kv}, v^g \to v^q\}$, and 2 types starting from the dialogue context node $\{v^q \to v^g, v^q \to v^{kv}\}$. Our graph network can stack multi-layers, assume $\mathbf{v_{lt,i}}$ is the representation of v_i at the layer lt:

$$\mathbf{v_{lt,i}} = \sum_{v_j \in \mathcal{V}} \alpha_{lt,j}^{g} \mathbf{W_a^g} \mathbf{v_{lt-1,j}} \; , \; \alpha_{lt,j}^{g} = \frac{\exp \gamma_{lt,j}^{g}}{\sum_{v_j \in \mathcal{V}} \exp \gamma_{lt,j}^{g}} \tag{6}$$

$$\gamma_{lt,j}^{g} = (\mathbf{r_{j \to i}})^T \, tanh(\mathbf{W_q^g} \mathbf{v_{lt-1,i}} + \mathbf{W_v^g} \mathbf{v_{lt-1,j}}) \tag{7}$$

where each $\mathbf{W}$ is a learn-able parameter, $\{\mathbf{v_{0,i}^{kv}}\}$ are initialized by $\{\mathbf{f_i^{kv,a}}\}$, $\mathbf{v_0^q}$ is initialized by $\mathbf{h_n}$, $\mathbf{v_0^g}$ is regarded as a learnable parameter. For simplicity, we use $\mathbf{v_{last,i}}$ to denote the representations outputted by the last layer.

2.4 Response Generation

Decoder State Initialization and Updating

Decoder is a GRU-based, whose initial state is given by $\mathbf{z_0} = \mathbf{W_{brg}}[\mathbf{v_{last}^g}; \mathbf{h_n}]$, and the following state $\mathbf{z_t}$ is updated as:

$$\mathbf{z_t} = GRU(\mathbf{z_{t-1}}, \mathbf{c_t^h}, \mathbf{c_t^f}, \mathbf{v_{last}^g}, \mathbf{y_{t-1}}) \tag{8}$$

$$\mathbf{c_t^h} = \sum_{i=1:n} \frac{\exp(\mathbf{h_i}^{\mathbf{T}} \mathbf{W_h} \mathbf{z_{t-1}})}{\sum_{j=1:n} \exp(\mathbf{h_j}^{\mathbf{T}} \mathbf{W_h} \mathbf{z_{t-1}})} \mathbf{h_i} \tag{9}$$

where $\mathbf{c_t^h}$ is the context attention,$\mathbf{c_t^f}$ is the infobox attention, and $\mathbf{y_{t-1}}$ is the embedding of the last predicted token.

Hierarchical Infobox Attribute Attention

We enable the Decoder to access the infobox knowledge in a coarse-to-fine manner. The hierarchical infobox attribute attention $\mathbf{c_t^f}$ is dynamically computed at each time step:

$$\mathbf{c_t^f} = \sum_i \alpha_{t-1,i}^{kv} \sum_j \alpha_{t-1,i,j}^{kw} \mathbf{W_{ia}}[\mathbf{v_{last,i}^{kv}}; \mathbf{f_{i,j}^{kw,ia}}] \tag{10}$$

The coarse-grained $\alpha_{t-1,i}^{kv}$ is of attribute level, which measures the relevance between the $\{\mathbf{v_{last}^{kv}}\}$ and $\mathbf{z_{t-1}}$:

$$\alpha_{t-1,i}^{kv} = \frac{\exp(MLP(\mathbf{v_{last,i}^{kv}}, \mathbf{z_{t-1}}))}{\sum_{v_{last,k}^{kv} \in T} \exp(MLP(\mathbf{v_{last,k}^{kv}}, \mathbf{z_{t-1}}))} \tag{11}$$

The fine-grained $\alpha_{t-1,i,j}^{kw}$ is of intra-attribute level; it considers the weight of each key-word pair in the attribute :

$$\alpha_{t-1,i,j}^{kw} = \frac{\exp(MLP(\mathbf{v_{last,i}^{kv}}, \mathbf{z_{t-1}}, \mathbf{f_{i,j}^{kw,ia}}))}{\sum_{f_{i,k'}^{kwc} \in f_i^{kv}} \exp(MLP(\mathbf{v_{last,i}^{kv}}, \mathbf{z_{t-1}}, \mathbf{f_{i,k'}^{kw,ia}}))} \tag{12}$$

Dynamic Mode Fusion

To minimize the impact of unknown words and diversify the generated responses, three modes are computed and fused when predicting the next token at time step t:

Vocab mode. The next token can be a word in the predefined vocab V, the probability distribution is given by:

$$P_{t,v} = softmax(\mathbf{W_{v2}}\, tanh(\mathbf{W_{v1}}[\mathbf{z_t}; \mathbf{c_t^h}; \mathbf{c_t^f}; \mathbf{v_{last}^g}; \mathbf{y_{t-1}}])) \tag{13}$$

Copy mode. The Decoder can copy a word from the query X. To obtain a more accurate copy probability distribution, we reuse the parameter $\mathbf{W_h}$ of the query attention Eq. 9.

$$P_{t,c} = softmax(\mathbf{H}\mathbf{W_h}\mathbf{z_t}) \tag{14}$$

Infobox mode. The Decoder can also select a relevant keyword pair, and extracts its word as the output. Here, we apply the same hierarchical technique as Eq. 10:

$$P_{t,f}(w_{i,j}) \propto \alpha_{t,i}^{kv} \cdot \alpha_{t,i,j}^{kw} \tag{15}$$

Mode fusion. subsequently, we fuse the above three distributions into one distribution P_t with a MLP network:

$$(\beta_{t,v}, \beta_{t,c}, \beta_{t,f}) = softmax(MLP([\mathbf{z_t}; \mathbf{c_t^h}; \mathbf{c_t^f}; \mathbf{v_{last}^g}; \mathbf{y_{t-1}}]))$$

$$P_t = \beta_{t,v} P_{t,v} + \beta_{t,c} P_{t,c} + \beta_{t,f} P_{t,f} \tag{16}$$

Training

The training follows the maximum likelihood estimation process, which minimizes the negative log likelihood:

$$\mathcal{L} = -\sum_t I(y_t) \log P_t(y_t | y_{1:t-1}, X, T) \tag{17}$$

where $I(\cdot)$ is an indicator function to alleviate the unknown words issue in the dialogue generation, it equals $0/1$ when the target token y_t is an unknown/known word, respectively.

3 Experiments

3.1 Dataset

We use a previous open-released Chinese corpus [Cai *et al.*, 2019b], and we have crawled about 895k infobox tables from Wikipedia. We use TF-IDF to rank the words of the query. According to the ranked order, we iteratively pick up a query word as the key to retrieve an infobox based on the pre-learned inverted index until a valid infobox has been found. If there is no matched infobox, a special blank infobox is adopted. Finally, the dataset is divided into train/validation/test sets where the size is 855K/30K/30K. https://github.com/pku-sixing/IJCAI2021-HITA-Graph.

3.2 Settings

Models. We first select 4 conversational models: **Seq2Seq**: The attentive Seq2Seq [Luong *et al.*, 2015]. **Pointer-Gen**: Based on the Seq2Seq, it allows the decoder to copy a word from the query [See *et al.*, 2017]. **CCM**: A commonsense knowledge-aware model that adopts graph attention [Zhou *et al.*, 2018]; **ConKADI**: One of the latest SOTA commonsense knowledge-aware generation models, which proposes multiple methods to better select the knowledge [Wu *et al.*, 2020a]. Then, we select 3 infobox-to-text baselines. **LSTMGate**: It proposes an LSTM-based Infobox Encoder [Liu *et al.*, 2018b], **HiLSTM**: It further proposes a Hierarchical LSTM Infobox Encoder [Liu *et al.*, 2019b]. **Trans**: The latest Transformer-based Infobox Encoder [Bai *et al.*, 2020]. Such three baselines similarly use the Dual Attention to access the infobox knowledge. To adapt to the dialogue generation, we integrate their infobox encoder and infobox attention modules into the Seq2Seq. For better comparison, we further integrate our proposed Dynamic Mode Fusion into such three infobox-to-text baselines, the variants are denoted as **'X+DMF'**.

Implementations. For CCM and ConKADI, we use their official codes and the commonsense knowledge released by ConKADI. For the others, we use our PyTorch implementations. Hyper-parameters are kept the same among models as possible. The word/char embedding dimension is 200; the part-of-speech tag embedding dimension is 10; the positional embedding dimension is 5; GRU's hidden size is 512. In our HITE, the multi-head attention has 4 heads and 2 layers; the interaction graph has 2 layers. Adam is used to optimizing parameters; the batch size is set to 50. The initial learning rate lr is 0.0001; after each epoch, lr will be halved if the perplexity on the validation set starts to increase. The training will be stopped if the perplexity on the validation set increases in two successive epochs. In the inference, beam search ($k = 10$ is adopted). The training of HITA-Graph consumes about 1 day on an Nvidia Titan-RTX GPU.

Metrics. For measuring the relevance to the ground-truth [Liu *et al.*, 2016], we employ 3 embedding-based metrics, Embedding-Average (**EmA**), Embedding-Greedy (**EmG**), Embedding-Extrema (**EmX**), and 5 overlapping-based metrics, **ROUGE-L**, **BLEU-1/2/3/4**. For measuring the diversity and the informativeness, we report the ratio of distinct 1/2-grams (**DIST1/2**) in generated words [Li *et al.*, 2016], and the 4-gram entropy (**Ent4**) [Zhang *et al.*, 2020].

3.3 Experimental Results and Analyses

Automatic evaluation. As shown in Table 1. HITA-Graph achieves the best overall performance; HITA-Graph wins first place in 9 metrics, and is comparable to first place in the remaining 2 metrics. Compared with the naive baseline Seq2Seq, knowledge-enhanced baselines have improvements more or less, indicating the necessity of incorporating knowledge. The improvements of three infobox-to-text baselines (i.e., LSTMGate, HiLSTM, Trans) are not notable. Unlike CCM, ConKADI, and our HITA-Graph, we find the reason is the inability to copy words from the query/infobox, so we subsequently equip our Dynamic Mode Fusion to such three baselines (i.e., '+DMF'). The enhanced variants achieve more notable improvements, especially in the aspect of diversity/informativeness. It shows the importance to design more generation modes for the Decoder. However, the enhanced variants are still behind our HITA-Graph, because our infobox-accessing solution is more suitable in the context of dialogue generation. The proposed HITA-Graph also outperforms two knowledge-enhanced conversational baselines, CCM and ConKADI, demonstrating: (1) Infobox can be a

Approach	EmA	EmG	EmX	ROUGE-L	BLEU1	BLEU2	BLEU3	BLEU4	DIST1	DIST2	Ent4
Seq2Seq	0.713	0.586	0.542	8.85	9.42	3.12	1.20	0.49	1.36	5.88	5.65
Pointer-Gen	0.755	0.618	0.577	10.62	10.51	3.94	1.67	0.78	5.49	19.78	7.73
CCM	0.822	**0.673**	0.621	10.61	11.02	3.52	1.30	0.51	1.66	8.86	8.33
ConKADI	0.824	0.653	0.617	11.82	11.02	3.95	1.65	0.75	6.49	28.28	**10.46**
LSTMGate	0.708	0.589	0.545	9.47	10.02	3.44	1.33	0.57	1.70	7.34	5.85
HiLSTM	0.731	0.604	0.561	9.94	10.35	3.58	1.38	0.55	1.91	8.73	6.64
Trans	0.730	0.601	0.560	9.73	10.22	3.50	1.33	0.54	1.78	8.11	6.33
LSTMGate+DMF	0.824	0.665	0.626	12.08	10.53	3.82	1.55	0.69	6.03	22.71	9.70
HiLSTM+DMF	0.823	0.665	0.625	12.07	10.70	3.84	1.50	0.64	6.24	22.65	9.64
Trans+DMF	0.820	0.663	0.624	12.16	10.23	3.75	1.51	0.67	6.37	23.95	9.75
HITA-Graph	**0.832**	0.668	**0.631**	**12.97**	**12.43**	**4.83**	**2.13**	**1.06**	**7.36**	**28.44**	10.43

Table 1: Automatic evaluation results. Scores in **bold** stand for the leadership among models.

Ours vs.	$\mathbf{R}_{win}$	$\mathbf{R}_{tie}$	$\mathbf{R}_{loss}$	$\mathbf{I}_{win}$	$\mathbf{I}_{tie}$	$\mathbf{I}_{loss}$
Seq2Seq	**75.5%**	11.7%	12.8%	**71.3%**	12.5%	16.2%
Pointer-Gen	**64.5%**	16.0%	19.5%	**63.0%**	12.7%	24.3%
Trans	**75.7%**	12.0%	12.3%	**71.3%**	10.5%	18.2%
Trans+DMF	**62.2%**	30.0%	17.8%	**61.3%**	15.7%	23.0%
CCM	**70.3%**	10.7%	19.0%	**62.8%**	9.2%	28.0%
ConKADI	**47.2%**	17.2%	35.6%	43.2%	14.5%	42.3%

Table 2: Human evaluation results, where **R/I** indicates Rationality/Informativeness. $_{win/tie/loss}$ means the ratio that our approach wins/ties/loses compared to the baseline. Scores in **bold** indicate our approach is significantly better (sign test, p-value < 0.005.)

#	Settings	EmA	BLEU3	DIST2	Ent4
0	Full Model	0.832	2.13	28.44	10.43
1	- Hybrid Embedding (Eq. 2)	0.833	2.13	27.03	10.21
2	- UNK Indicator (Eq. 17)	0.825	1.91	26.34	9.95
3	- HIAA	0.832	2.06	28.29	10.18
4	- DMF	0.825	1.88	14.66	9.98
5	- HITE (i.e., -HIAA&DMF)	0.830	1.63	15.66	9.92

Table 3: Ablation study. HITE is the precondition of both HIAA and DMF, and only servers for them. Therefore, there is no setting that only ablates the HITE, or only uses the HIAA+DMF.

knowledge source in dialogue generation; (2) The proposed HITA-Graph is effective.

Human evaluation. Three volunteers are invited to annotate 200 sampled cases (1,200 pairs in total). The judgment is pair-wise, and follows two criteria: (1) Rationality: measuring the fluency and the relevance); (2) Informativeness: checking how much relevant knowledge is provided. As reported in Table 2, HITA-Graph outperforms all baselines. Compared to the commonsense-based ConKADI, HITA-Graph has a notable advantage in terms of rationality, and a slightly better performance in terms of informativeness. The agreement among annotators is highly consistent: (1) for the rationality, 94%/62% cases are given the same label by at least 2/3 volunteers; (2) for the informativeness, 94%/58% cases are given the same label by at least 2/3 volunteers.

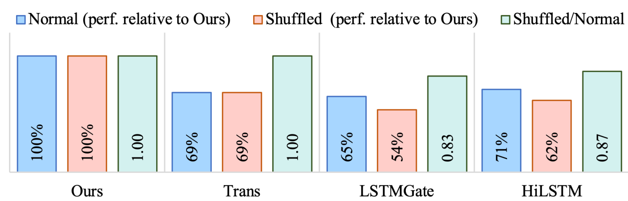


Figure 3: Average performance (perf.) on 4 metrics (see Table 3). All reported scores are based on the percentage of the scores of all methods to the un-shuffled normal scores of our HITA-Graph model. To make a fair comparison, all approaches are not equipped with DMF. 'Normal'/'Shuffled's show the relative performance on the normal/shuffled test sets; the performance of our approach is defined as 100%. 'Shuffled/Normal' shows the performance difference between the normal test set and the shuffled test set.

Ablation study. As reported in Table 3, we designed and evaluated 5 HITA-Graph variants: (1) In case #1 and #2, we remove two methods that can alleviate the issue of unknown words that appears in the infobox and the dialogues, respectively. The decreased performance indicates the importance of handling unknown words in HITA-Graph explicitly; (2) In the next case #3-5, we evaluate the performance of HITA-Graph if it incorporates less or no infobox knowledge. DMF brings more notable improvement than HIAA, especially in the aspect of diversity (i.e., DIST2). We find HIAA mainly improves the word overlap-based relevance (BLEU3) if we compare #3 and #0, or #5 with #4.

Attribute-order sensitivity study. Previous infobox-to-text works show the order of infobox attributes can be a crucial factor to the performance [Bai *et al.*, 2020]. However, unlike them, dialogue generation does not access the infobox in a specific order. Therefore, in dialogue generation, the infobox encoder should not be affected by order of attributes, and we subsequently propose an order-irrelevance infobox encoder, HITE. In this study, we demonstrate this feature by checking the performance (perf.) on the shuffled test set, which has the same attributes as the original test set, but with attribute orders randomly shuffled. As reported in Figure 3: (1) 'LSTMGate' and 'HiLSTM' employ the sequential LSTMs to encode attributes, and thus their performance

Query: 最近发现林宥嘉的'突然想起你'。 Recently, I discovered **Yoga Lin**'s '**Suddenly Thinking of You**'.
Seq2Seq: 我也喜欢林宥嘉。 I like Yoga Lin, too.
Pointer-Gen: 我也是突然想起了 我也是突然想起我了。 I also suddenly think of me, I also suddenly think of me.
Trans: 我也喜欢林宥嘉。I like Yoga Lin, too.
Trans+DMF: 我也喜欢林宥嘉。 I like Yoga Lin, too.
CCM: 我也想看我的偶像。 I also want to see my idol.
ConKADI: 林宥嘉是我最喜欢的歌了。 Yoga Lin is my favorite song.
HITA-Graph: 我也很喜欢林宥嘉的歌。 I also like Yoga Lin's songs.
Related Attributes: (表: 林宥嘉)(Infobox: Yoga Lin), (职业: 歌手)(Job: Singer), (演奏乐器: 歌唱)(Musical Instruments: Singing)...

Table 4: Case Study. '林宥嘉' ('Yoga Lin') is a pop singer, who has sang a song '突然想起你' ('Suddenly Thinking of You').

can be seriously impacted by the order. They only have 83-87% of performance on the shuffled test set compared with the normal test set. (2) Our approach would not be impacted by order of attributes, and has a notable leadership compared with baselines; (3) The transformer-based 'Trans' is similarly irrelevant to the order because its attribute re-ordering module is inactivated in the dialogue [Bai *et al.*, 2020]. However, the performance is worse than ours. In summary, this study proves our approach is much robust and powerful than baselines, demonstrating our approach can achieve better performance on the more complex real scenarios.

Case study. We report a real case in Figure 4. In the query, the user says he/she found a song, 'Suddenly Thinking of You ', performed by the singer 'Yoga Lin'. Except for the ConKADI and our HITA-Graph, other approaches did not realize 'Yoga Lin' is a singer. Although ConKADI realized this, it failed to generate a fluent response. This case indicates the proposed approach HITA-Graph is able to generate a high-quality and informative response with the incorporation of the infobox knowledge.

4 Related Work

Knowledge-aware dialogue generation. Suffering from outputting dull responses (such as 'I don't know') [Li *et al.*, 2016], many efforts have been made to diversify the generations; for example, new objective [Li *et al.*, 2016], latent variable [Gao *et al.*, 2019], back-translation [Su *et al.*, 2020], etc. Unlike human beings who can enhance their dialogue understanding and inference with various learned knowledge, machines only receive a query with limited knowledge [Ghazvininejad *et al.*, 2018; Yu *et al.*, 2020]. Therefore, the generation quality is always far from satisfactory. To bridge the gap of accessing knowledge, various types of knowledge are explored: 1) Knowledge texts, such as encyclopedia texts [Dinan *et al.*, 2019], documents [Meng *et al.*, 2019], prototype dialogues [Cai *et al.*, 2019a], and web pages [Tam, 2020]; 2) Structured knowledge bases, such as commonsense knowledge [Liu *et al.*, 2018a; Wu *et al.*, 2020a; Zhang *et al.*, 2020] and spread-sheet tables [Qin *et al.*, 2019]; 3) Topic words and keywords [Wang *et al.*, 2018], which are also regarded as knowledge guidance knowledge. Different from such approaches, the proposed HITA-Graph uses the infobox knowledge, which is seldom used in the context of open-domain dialogue response generation.

Data-to-text. Infobox-to-text has been well studied [Chen *et al.*, 2019; Chen *et al.*, 2020]. The basic paradigm is similar to the Seq2Seq-based dialogue generation: An infobox table is first encoded into hidden states, and then a decoder generates a text by attentively accessing the hidden states [Nema *et al.*, 2018; Liu *et al.*, 2019a]. The representative approach LSTMGate [Liu *et al.*, 2018b] proposes an LSTM-based encoder to encode an infobox table, and a dual-attention mechanism to access the infobox knowledge during the decoding. Subsequently, to better encode an infobox, pretrained language models (such as GPT-2 [Chen *et al.*, 2020]), hierarchical LSTM encoders [Liu *et al.*, 2019b], and the transformer-based encoder [Bai *et al.*, 2020]) are successively proposed. Compared to them, HITA-Graph is notably different in encoding/accessing the infobox knowledge: (1) The encoding process of HITE is independent of the attribute order. The attribute order is a crucial factor to the generated text in the infobox-to-text task, and many efforts are devoted to it [Puduppully *et al.*, 2019; Bai *et al.*, 2020]. However, unless dialogue generation rephrases the infobox in an orderly fashion, the attribute order is meaningless. (2) Infobox-to-text approaches usually consider only one or two vertical domains, but HITA-Graph is an open-domain approach. (3) HITA-Graph proposes a novel Infobox-Dialogue Interaction Graph Network to conduct the context-level infobox encoding, which not only promotes the data flow, but also interacts with the dialogue context more easily and effectively.

5 Conclusion

This paper proposes an infobox knowledge-aware dialogue generation approach, HITA-Graph. An infobox table specifies an entity with multiple attributes. The advantages of the infobox tables include: (1) easy to be collected from the Internet; (2) knowledge has been well-organized as attributes; (3) each infobox focuses on only one entity, bringing high knowledge consistency. In HITA-Graph, we propose a novel order-irrelevance Hierarchical Infobox Table Encoder and a novel Infobox-Dialogue Interaction Graph Network to encode the infobox knowledge better. We also propose a Hierarchical Infobox Attribute Attention mechanism to support the access of the encoded knowledge at different levels of granularity, and a Dynamic Mode Fusion strategy to generate more informative responses. Both the automatic and human evaluation demonstrate the proposed HITA-Graph outperforms representative competitors in almost all experiments.

Acknowledgments

This work is partly supported by ICBC Technology.

References

[Bai *et al.*, 2020] Yang Bai, Ziran Li, Ning Ding, Ying Shen, and Hai-Tao Zheng. Infobox-to-text generation with tree-like planning based attention network. In *IJCAI*, 2020.

[Bao *et al.*, 2018] Junwei Bao, Duyu Tang, Nan Duan, Zhao Yan, and Yuanhua Lv et al. Table-to-text: Describing table region with natural language. In *AAAI*, 2018.

[Cai *et al.*, 2019a] Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, and et al. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *NAACL*, 2019.

[Cai *et al.*, 2019b] Deng Cai, Yan Wang, and et al. Retrieval-guided dialogue response generation via a matching-to-generation framework. In *EMNLP-IJCNLP*, 2019.

[Chen *et al.*, 2019] Shuang Chen, Jinpeng Wang, Xiaocheng Feng, Feng Jiang, Bing Qin, and Chin-Yew Lin. Enhancing neural data-to-text generation models with external background knowledge. In *EMNLP-IJCNLP*, 2019.

[Chen *et al.*, 2020] Wenhu Chen, Jianshu Chen, Yu Su, Zhiyu Chen, and William Yang Wang. Logical natural language generation from open-domain tables. In *ACL*, 2020.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, and et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *EMNLP*, 2014.

[Dinan *et al.*, 2019] Emily Dinan, Stephen Roller, Kurt Shuster, and et al. Wizard of wikipedia: Knowledge-powered conversational agents. In *ICLR*, 2019.

[Gao *et al.*, 2019] Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and et al. A discrete CVAE for response generation on short-text conversation. In *EMNLP*, 2019.

[Ghazvininejad *et al.*, 2018] Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, and et al. A knowledge-grounded neural conversation model. In *AAAI*, 2018.

[Kim *et al.*, 2016] Yoon Kim, Yacine Jernite, and et al. Sontag. Character-aware neural language models. *AAAI*, 2016.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, and et al. A diversity-promoting objective function for neural conversation models. In *NAACL*, 2016.

[Liu *et al.*, 2016] Chia-Wei Liu, Ryan Lowe, Iulian Serban, and et al. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2016.

[Liu *et al.*, 2018a] Shuman Liu, Hongshen Chen, Zhaochun Ren, Yang Feng, Qun Liu, and Dawei Yin. Knowledge diffusion for neural dialogue generation. In *ACL*, 2018.

[Liu *et al.*, 2018b] Tianyu Liu, Kexiang Wang, Lei Sha, Baobao Chang, and Zhifang Sui. Table-to-text generation by structure-aware seq2seq learning. In *AAAI*, 2018.

[Liu *et al.*, 2019a] Tianyu Liu, Fuli Luo, Baobao Chang, and et al. Towards comprehensive description generation from factual attribute-value tables. In *ACL*, 2019.

[Liu *et al.*, 2019b] Tianyu Liu, Fuli Luo, Qiaolin Xia, Shuming Ma, and et al. Hierarchical encoder with auxiliary supervision for neural table-to-text generation: Learning better representation for tables. In *AAAI*, 2019.

[Luong *et al.*, 2015] Thang Luong, Hieu Pham, and Christopher D. Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[Meng *et al.*, 2019] Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. Refnet: A reference-aware network for background based conversation. *CoRR*, abs/1908.06449, 2019.

[Nema *et al.*, 2018] Preksha Nema, Shreyas Shetty, Parag Jain, Anirban Laha, and et al. Generating descriptions from structured data using a bifocal attention mechanism and gated orthogonalization. In *NAACL*, 2018.

[Puduppully *et al.*, 2019] Ratish Puduppully, Li Dong, and Mirella Lapata. Data-to-text generation with content selection and planning. In *AAAI*, 2019.

[Qin *et al.*, 2019] Libo Qin, Yijia Liu, Wanxiang Che, Haoyang Wen, Yangming Li, and Ting Liu. Entity-consistent end-to-end task-oriented dialogue system with KB retriever. In *EMNLP-IJCNLP*, 2019.

[See *et al.*, 2017] Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, 2017.

[Su *et al.*, 2020] Hui Su, Xiaoyu Shen, Sanqiang Zhao, Xiao Zhou, Pengwei Hu, and et al. Diversifying dialogue generation with non-conversational text. In *ACL*, 2020.

[Tam, 2020] Yik-Cheung Tam. Cluster-based beam search for pointer-generator chatbot grounded by knowledge. *Comput. Speech Lang.*, 64:101094, 2020.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, and et al. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[Wang *et al.*, 2018] Wenjie Wang, Minlie Huang, Xin-Shun Xu, and et al. Chat More: Deepening and Widening the Chatting Topic via A Deep Model. In *SIGIR*, 2018.

[Wu *et al.*, 2020a] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Diverse and informative dialogue generation with context-specific commonsense knowledge awareness. In *ACL*, 2020.

[Wu *et al.*, 2020b] Sixing Wu, Ying Li, Dawei Zhang, Yang Zhou, and Zhonghai Wu. Topicka: Generating commonsense knowledge-aware dialogue responses towards the recommended topic fact. In *IJCAI*, 2020.

[Yu *et al.*, 2020] Wenhao Yu, Chenguang Zhu, Zaitang Li, and et al. A survey of knowledge-enhanced text generation. *CoRR*, abs/2010.04389, 2020.

[Zhang *et al.*, 2020] Houyu Zhang, Zhenghao Liu, Chenyan Xiong, and Zhiyuan Liu. Grounded conversation generation as guided traverses in commonsense knowledge graphs. In *ACL*, 2020.

[Zhou *et al.*, 2018] Hao Zhou, Tom Young, Minlie Huang, and et al. Commonsense Knowledge Aware Conversation Generation with Graph Attention. In *IJCAI*, 2018.