Knowledge-Aware Dialogue Generation via Hierarchical Infobox Accessing and Infobox-Discourse 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, dawei_zhang, 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-Discourse Interaction Graph Network is built to effectively integrate the infobox context and the discourse 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. Lastly 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.

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 spreadsheet 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-

![Figure 1: The ‘Bill Gates’ Infobox from Wikipedia.](image)
2 Our Approach

2.1 Problem Formulation

Let \( D = \{ (X, Y, T) \} \) denote the corpus, where \( X = (x_1, \cdots, x_n) \) is a query, \( Y = (y_1, \cdots, y_m) \) is a response, and \( T = \{ (f^k, f^v) \}^l \) is an infobox table that consists of a set of attribute key-values and can be retrieved from the infobox base \( T \). The goal of our task is: \( Y^* = \arg \max_{y \in Y} P(Y^* | X, T) \).

2.2 Context Encoder

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

\[
h_t^a = GRU(x_t, h_{t-1}^a); h_t^b = GRU(x_{n-t+1}, h_{t-1}^b)
\]

where \( x \) is the embedding of \( x \), and \( h_t \) is the concatenation \( [h_t^a; h_t^b] \). \( h_n \) is regarded as the context summary.

2.3 Hierarchical Infobox Table Encoder

For each attribute \( f_i^k \) 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 cannot be recognized.

Thus, we design a hybrid-level method. Taking \( f_i^k \) as an example, the corresponding embedding \( \hat{f}_{i,j,k}^{hybrid} \) is given by:

\[
\hat{f}_{i,j,k}^{hybrid} = MLP_h(f_{i,j,k}^{char})
\]

where \( \hat{f}_{i,j,k}^{hybrid} \) is obtained by looking up the word-level embedding matrix, \( f_{i,j,k}^{char} \) is computed from the char sequence of \( f_i^k \) with the CharCNN Encoder [Kim et al., 2016], and \( MLP_h \) is a MLP network. The incorporation of chars can alleviate the issue of the sparse word distribution. Subsequently, each attribute \( f_{i,j,k}^{hybrid} \) is represented as a set of key-word embedding:

\[
F_{i,j,k}^{kw} = \{ f_{i,j,k}^{kw} \} = \{ [f_{i,j,k}^{hybrid}; w_{i,j}^{pos}; p_{i,j}^f; p_{i,j}^b]\}
\]

where \( f_{i,j,k}^{hybrid} \) is the \( j \)-th key-word pair of \( f_{i,j,k}^{kw} \), \( w_{i,j}^{pos} \) is the part-of-speech tag of \( w_{i,j} \), and \( p_{i,j}^f \) and \( p_{i,j}^b \) are local positions counted from the beginning (i.e., \( j \)) and the end (i.e., \( |f_{i,j}^k| - j + 1 \)), respectively.

Then, the multi-head self-attention (denoted as \( MHA \), [Vaswani et al., 2017]) is used to compute the intra-attribute level representations \( F_{i,j,k}^{kw,ia} = \{ f_{i,j,k}^{kw,ia}\} \):

\[
F_{i,j,k}^{kw,ia} = MHA(Q = F_{i,j,k}^{kw}, K = F_{i,j,k}^{kw}, V = F_{i,j,k}^{kw})
\]

Attribute-Level Encoding

For each key-value attribute \( f_{i,j,k}^{kw} \), its attribute-level embedding \( f_{i,j,k}^{kw,a} \) is the weighted sum of \( F_{i,j,k}^{kw,ia} = \{ f_{i,j,k}^{kw,ia}\} \):

\[
f_{i,j,k}^{kw,a} = \sum_j \sum_k \exp(f_{i,j,k}^{kw,ia}) f_{i,j,k}^{kw,ia}
\]

where \( f_s \) is a learnable parameter, serving as a special query to compute the weight; thus, the infobox table \( T \) can be represented as a set of attribute embedding \( F_{i,j,k}^{kw,a} = \{ f_{i,j,k}^{kw,a}\} \).
Infobox-Discourse Interaction Graph Network

An attribute-level representation $v_{t_i}^{kv,a}$ interacts with neither the discourse 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-Discourse Interaction Graph Network to capture the interaction information among attributes and the discourse context. For each infobox-discourse instance, we first construct a graph $G = (V, E, R)$. The node set is $V = \{v^{kv}\} \cup \{v^g\} \cup \{v^q\}$, 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 discourse context. Meanwhile, $E = \{(v_i, r_{i,j} \rightarrow v_j)\}$ is edge set, where $r_{i,j} \in R$ is the relational edge relation. For promoting the knowledge flow, $G$ is fully-connected and directional in this paper. For each two nodes $v_i, v_j$, their directional edge $r_{i,j}$ depends on the node types of $v_i, v_j$. Therefore, the relation set $R$ consists of 3 self-connection types $\{sel_{f^{kv}} = sel_{f^g}, sel_{f^q}\}$, 3 types starting from an attribute node $\{v_i^{kv} \rightarrow v_{i,j}^{g}, v_{i,j}^{kv} \rightarrow v_{i,j}^g, v_{i,j}^{kv} \rightarrow v_{i,j}^g\}$, 2 types starting from the global node $\{u^g \rightarrow v^{kv}, v^g \rightarrow v^q\}$, and 2 types starting from the discourse context $\{v^q \rightarrow v^g, v^q \rightarrow v^{kv}\}$. Our graph network can stack multi-layers, assume $v_{l.t,i}$ is the representation of $v_i$ at the layer $lt:$

$$v_{l.t,i} = \sum_{v_j \in V} \alpha_{l.t,j}^g W^g_n v_{l-1.j} + \alpha_{l.t,j} = \exp(\gamma_{l.t,j})$$

$$\gamma_{l,t,j} = (r_{j-i})^T \tanh(W^g_n v_{l-1.j} + W^g v_{l-1.j})$$

where each $W$ is a learnable parameter, $\{v_{0}^{\gamma}\}$ are initialized by $\{f^{kv,a}\}$, $v_{0}^{\gamma}$ is initialized by $h_n$, $v_{0}^{g}$ is regarded as a learnable parameter. For simplicity, we use $v_{l-1,i}$ to denote the representations output by the last layer.

2.4 Response Generation

Decoder State Initialization and Updating

Decoder is a GRU-based, whose initial state is given by $z_{0} = W_{brg}[v_{l-1}, h_n]$, and the following state $z_{t}$ is updated as:

$$z_{t} = GRU(z_{t-1}, c_{t}^{h}, c_{t}^{f}, v_{l-1}^{g}, y_{t-1})$$

$$c_{t}^{h} = \sum_{i=1:n} \sum_{j=1:n} \exp(h_{i}^T W_{h} z_{t-1}) h_{i}$$

where $c_{t}^{h}$ is the context attention, $c_{t}^{f}$ is the infobox attention, and $y_{t-1}$ 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 attention $c_{t}^{f}$ is dynamically computed at each time step:

$$c_{t}^{f} = \sum_{i=1:n} \alpha_{t-1,i}^{kw} \sum_{j=1:n} \exp(MLP(v_{l-1,i}^{kw}, v_{l-1,j}^{kw, in}))$$

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

$$\alpha_{t-1,i}^{kw} = \frac{\exp(MLP(v_{l-1,i}^{kw}, z_{t-1}))}{\sum_{i=1:n} \exp(MLP(v_{l-1,i}^{kw}, z_{t-1}))}$$

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

$$\alpha_{t-1,i}^{kw, in} = \frac{\exp(MLP(v_{l-1,i}^{kw, in}, z_{t-1}, e_{i,j}^{kw, in}))}{\sum_{i=1:n} \sum_{k' \in f_{i,j}} \exp(MLP(v_{l-1,i}^{kw, in}, z_{t-1}, e_{i,j}^{kw, in}))}$$

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.
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:

$$P_{t,v} = \text{softmax}(W_{v2} \tanh(W_{v1}[z_t; c^f_t; c^i_t; v^E_{\text{inst}}; y_{t-1}]))$$

(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 $W_h$ of the query attention Eq. 9.

$$P_{t,c} = \text{softmax}(HW_hz_t)$$

(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:

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

(15)

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

$$\beta_{t,v}, \beta_{t,c}, \beta_{t,f} = \text{softmax}(\text{MLP}([z_t; c^f_t; c^i_t; v^E_{\text{inst}}; y_{t-1}]))$$

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

(16)

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

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

(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. LSTM-MGate: 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., LSTM-MGate, 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
Table 1: Automatic evaluation results. Scores in bold stand for the leadership among models.

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

Table 2: Human evaluation results, where R/I indicates Rationality/Informativeness. \( r_{win/ie/loss} \) 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.)

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

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.

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
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 spreadsheet 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


