Towards Complex Scenarios: Building End-to-End Task-Oriented Dialogue System across Multiple Knowledge Bases

Libo Qin[1*], **Zhouyang Li**[2*], **Qiying Yu**[2], **Lehan Wang**[2], **Wanxiang Che**[2†]
[1]School of Computer Science and Engineering, Central South University, China
[2]Research Center for Social Computing and Information Retrieval, Harbin Institute of Technology, China
lbqin@csu.edu.cn, {zhouyangli, car}@ir.hit.edu.cn, yuqiying@hit.edu.cn, lhwang@stu.hit.edu.cn

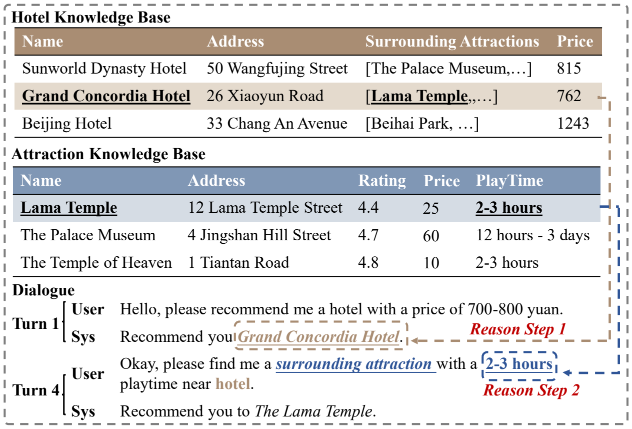
Abstract

With the success of the sequence-to-sequence model, end-to-end task-oriented dialogue systems (EToDs) have obtained remarkable progress. However, most existing EToDs are limited to single KB settings where dialogues can be supported by a single KB, which is still far from satisfying the requirements of some complex applications (multi-KBs setting). In this work, we first empirically show that the existing single-KB EToDs fail to work on multi-KB settings that require models to reason across various KBs. To solve this issue, we take the first step to consider the multi-KBs scenario in EToDs and introduce a **K**B-**o**ver-**KB** **H**eterogeneous Graph **A**ttention **N**etwork (KoK-HAN) to facilitate model to reason over multiple KBs. The core module is a triple-connection graph interaction layer that can model different granularity levels of interaction information across different KBs (*i.e.*, intra-KB connection, inter-KB connection and dialogue-KB connection). Experimental results confirm the superiority of our model for multiple KBs reasoning.

Introduction

Task-oriented dialogue systems (Young et al. 2013) aim to help user to complete hotel booking and restaurant reservations, which have attracted increasing attention. Recently, end-to-end task-oriented dialogue systems (EToDs) (Eric and Manning 2017; Wen et al. 2018; Lei et al. 2018; Wu, Socher, and Xiong 2019; Qin et al. 2020) have emerged to free manually designed pipeline modules.

Existing EToDs can be classified into two main categories. The first strand of work (Ham et al. 2020; Hosseini-Asl et al. 2020; Olabiyi et al. 2020; Peng et al. 2021; Kulhánek et al. 2021; Lee 2021; Yang, Li, and Quan 2021; Gao et al. 2021; Gu et al. 2021) treats all task-oriented dialogue pipeline tasks as a sequence prediction problem, using pre-trained models to predict dialog state, system action and system response in one sequence. These works can train all pipeline tasks in an end-to-end fashion but need annotations for intermediate results. The second strand of work (Eric and Manning 2017; Zhu et al. 2020) directly uses dialogue history and knowledge bases (KB) as input and optimizes neural encoder-decoder models to output system responses. In this paper, we focus on this line of work that does not need any intermediate results supervision.



Hotel Knowledge Base

Name	Address	Surrounding Attractions	Price
Sunworld Dynasty Hotel	50 Wangfujing Street	[The Palace Museum,…]	815
Grand Concordia Hotel	26 Xiaoyun Road	[**Lama Temple**,,…]	762
Beijing Hotel	33 Chang An Avenue	[Beihai Park, …]	1243

Attraction Knowledge Base

Name	Address	Rating	Price	PlayTime
Lama Temple	12 Lama Temple Street	4.4	25	**2-3 hours**
The Palace Museum	4 Jingshan Hill Street	4.7	60	12 hours - 3 days
The Temple of Heaven	1 Tiantan Road	4.8	10	2-3 hours

Dialogue

Turn 1 User: Hello, please recommend me a hotel with a price of 700-800 yuan.
Sys: Recommend you *Grand Concordia Hotel*.

Turn 4 User: Okay, please find me a ***surrounding attraction*** with a 2-3 hours playtime near hotel.
Sys: Recommend you to *The Lama Temple*.

Figure 1: Multi-KBs settings dialogue from the CrossWOZ dataset (Zhu et al. 2020) means that the dialogue needs to be grounded by multiple KBs (We translate the dialogue from Chinese to English for better illustration). Words with color refer to the queried knowledge entity from KB.

Successfully retrieving KB is the key to a task-oriented dialogue system (Qin et al. 2019b). To this end, some EToDs (Eric and Manning 2017; Wen et al. 2018; Madotto, Wu, and Fung 2018; Gangi Reddy et al. 2019; Wu, Socher, and Xiong 2019) perform attention mechanism to query KB. Qin et al. (2020) introduce a dynamic fusion network to consider domain features for better querying KB. Yang et al. (2022) propose an explicit and interpretable Neuro-Symbolic KB reasoning framework for EToDs. While current systems achieve near-human F1 scores on SMD (Eric and Manning 2017), it is questionable whether this can faithfully meet the real-world applications. Unfortunately, the answer is NOT and we argue that the current EToDs community is over-optimistic about the current progress, because we empirically observe that current researches on EToDs are mainly limited to the single-KB settings where dialogues can be supported by a single KB, **which is still far from satisfying the requirements of some complex applications (multiple KBs settings) in a real-world scenario.** For ex-

[*]These authors contributed equally.

[†]Corresponding author.

ample, as shown in Figure 1, when responding to the user query "*please find me a surrounding attraction with a playtime of 2-3 hours around the hotel*" in Turn 4, the model not only needs to query *Hotel Knowledge Base* to find the hotel's `Surrounding Attractions` (***Reason Step 1***), but also retrieves *Attraction Knowledge Base* to search for qualified results (`2-3 hours`) (***Reason Step 2***). Such complex multiple KBs that require the ability to effectively reason across multiple KBs are practical and useful in real-world scenarios, which cannot be achieved by the previous single-KB dialogue models.

Motivated by these observations, we make the first attempt to explore the multi-KBs settings for EToDs, hoping to draw more attention to this complex real-world scenario. When extending the single-KB to multi-KBs settings, a natural approach is concatenating multiple KBs into a big single-KB for directly adopting the existing single-KB models. However, two new challenges arise when solely adopting concatenation approach. (1) Firstly, simply flattening all KBs makes it hard to capture high-order structure relationship information in KB, leading to imperfect KB representation learning; (2) Secondly, it fails to effectively reason across multiple KBs with simple KBs' concatenation.

To solve the aforementioned challenges, we propose a **KB**-**o**ver-**KB** **H**eterogeneous Graph **A**ttention **N**etwork (KoK-HAN) for multiple KBs EToDs. The core insight is a triple-connection heterogeneous graph interaction layer, which achieves to fully incorporates the high-order structure relationship and the multiple KBs interaction simultaneously. Specifically, to address the first challenge, we introduce an *intra-KB connection*, which connects each KB entity node with other nodes in the same KB row. With such consideration, the high-order structure relationship information within a KB can be effectively captured. To solve the second challenge, we propose an *inter-KB connection*, which connects all different KBs, to build the information flow across multiple KBs. Further, we introduce a *dialogue-KB connection*, where a co-occurrence edge is created to connect related nodes if they co-occur in the dialogue history and the corresponding KB, to establish the connection between dialogue and KB.

We conduct experiments on two datasets, including CrossWOZ (Zhu et al. 2020) and RiSAWOZ (Quan et al. 2020). Results show that KoK-HAN achieves superior performance. In addition, extensive analysis experiments show that KoK-HAN successfully captures intra-KB and inter-KB relationships.

The core contributions of this work are summarized as below:

- We empirically point out that the existing models are limited to single-KB settings that can not handle complex applications (multi-KBs settings), which motivates researchers to rethink the current progress of EToDs and shed a light on this direction.
- To the best of our knowledge, we are the first to consider the multi-KBs settings for EToDs, which is towards building a more practical and useful dialogue in real-world scenarios.
- We introduce a novel KB-over-KB heterogeneous graph network (KoK-HAN) for multi-KBs EToDs, which is able to facilitate model to effectively reason across multiple KBs. Besides, extensive analysis experiments show that our framework has successfully reasoned across multiple KBs.

To facilitate this research, all codes and datasets are publicly available at https://github.com/RaleLee/KoK-HAN.

Problem Formulation

We describe formulation definition for multi-KBs EToDs. The critical difference between single-KB EToDs and multi-KBs EToDs is that a single KB can support single-KB EToDs while multi-KBs EToDs are supported by multiple KBs that require reasoning ability across different KBs.

Specifically, following Wu, Socher, and Xiong (2019) and Qin et al. (2020), we define the EToD as predicting response $\mathbf{Y} = (y_1, y_2, \ldots, y_n)$ given the input dialogue history $\mathbf{X} = (x_1, x_2, \ldots, x_m)$ and multiple KBs $\mathcal{B} = \{\mathcal{B}^1, \ldots, \mathcal{B}^K\}$, where K denotes the number of KB; m and n are the length of dialog history and response, respectively. Formally, the probability of a response is defined as:

$$P(\mathbf{Y} \,|\, \mathbf{X}, \mathcal{B}) = \prod_{t=1}^{n} p(y_t \,|\, y_1, \ldots, y_{t-1}, \mathbf{X}, \mathcal{B}), \qquad (1)$$

where y_t denotes an output word at t timestep.

Model

This section describes the architecture of KB-over-KB heterogeneous graph attention network (KoK-HAN), which is illustrated in Figure 2. It mainly consists of three components: an encoder to encode dialogue history and multiple KBs, a KB-over-KB heterogeneous graph layer to model the interaction information across multiple KBs, and a multi-KBs aware decoder to generate dialogue response.

Encoder

Knowledge Base Representation For each $\mathcal{B}^i$ in $\mathcal{B}$, which consists of $|\mathcal{R}^i|$ rows and $|\mathcal{C}^i|$ columns, $b_{i,j}$ stands for the value of the entity in the i^{th} row and the j^{th} column.

Following Eric et al. (2017), each entity in $\mathcal{B}$ will be represented in triplet format as (subject, relation, object). Besides, we apply a word embedding function ϕ^{emb} to obtain the word embedding of the subject, relation, and object and sum them up to obtain the representation of each KB triplet. Finally, the knowledge base representations can be denoted as $\mathcal{B} = \{b^1_{\{1,1\}}, \ldots, b^1_{\{|\mathcal{R}^1|,|\mathcal{C}^1|\}}, \ldots, b^K_{\{1,1\}}, \ldots, b^K_{\{|\mathcal{R}^K|,|\mathcal{C}^K|\}}\}$.

Dialogue History Representation For each token in dialogue history $\boldsymbol{X} = (x_1, x_2, \ldots, x_m)$, we follow Wu, Socher, and Xiong (2019) to store the speaker information and position encoding to capture the sequential dependencies. For example, the first utterance from the system in Figure 1 will be denoted as {($sys, turn1, Recommend),($sys, turn1, you),...}(Wu, Socher, and Xiong 2019). We first use a bag-of-word method to acquire each initial token embedding and

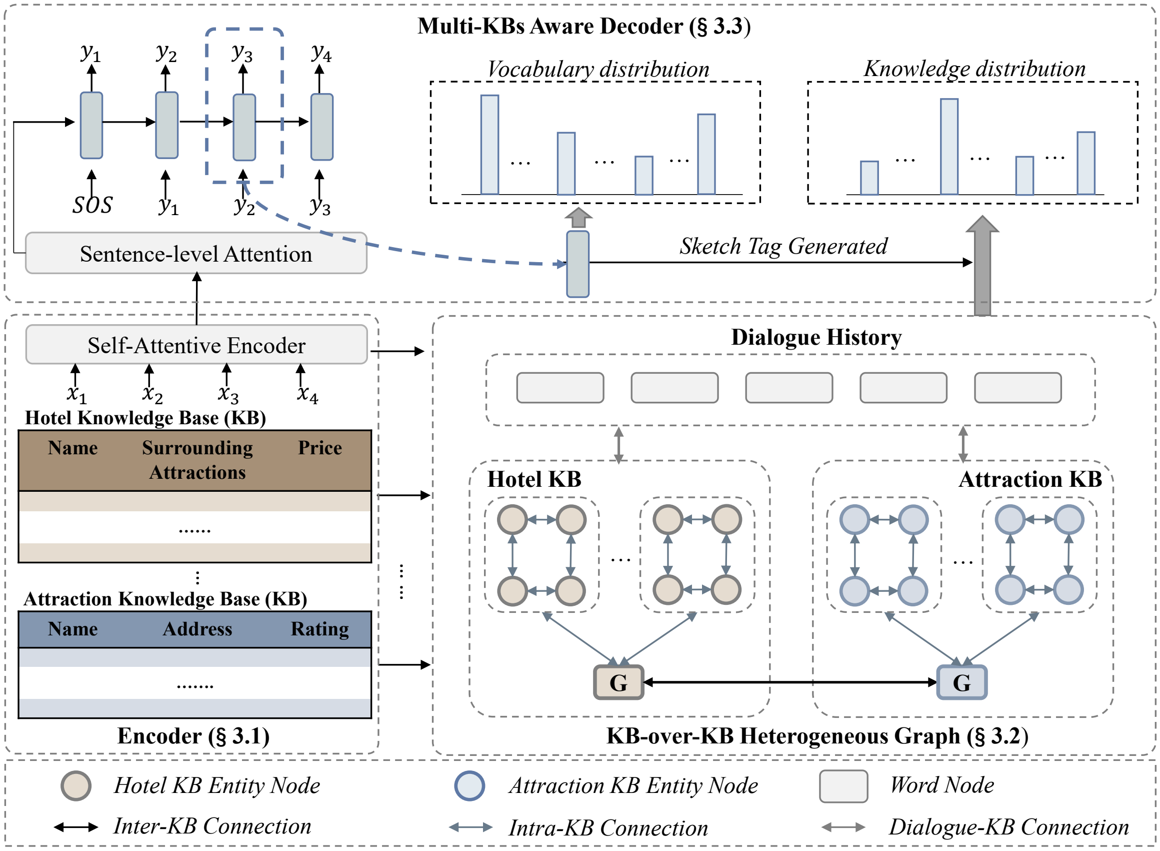


Figure 2: The illustration of KoK-HAN. To simplify, the *Intra-KB Connection*, *Inter-KB Connection*, and *Dialogue-KB Connection* are shown in high-level connection. G denotes the global Node. More details in Section 3.2.

obtain $\hat{\boldsymbol{X}}$. Then, following Qin et al. (2019a), we further adopt a self-attentive encoder to obtain the final dialogue history representation. Specifically, a Bi-GRU (Cho et al. 2014) is introduced to encode the sequential information while a self-attention mechanism (Vaswani et al. 2017) is applied to incorporate the contextual information.

Bi-GRU. The Bi-GRU reads the input $\hat{\boldsymbol{X}}$ to generate the corresponding hidden states $\boldsymbol{H} = (\boldsymbol{h}_1, \boldsymbol{h}_2, \ldots, \boldsymbol{h}_m)$.

Self-Attention. Self-attention is further introduced to capture context-aware features. Specifically, following Qin et al. (2019a), we first map the input matrix $\hat{\boldsymbol{X}}$ to queries ($\boldsymbol{Q}$), keys ($\boldsymbol{K}$) and values ($\boldsymbol{V}$) matrices and then the self-attention output is calculated by $\boldsymbol{S} = softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_k}})\boldsymbol{V}$ where d_k represents dimension of keys.

We concatenate the output of Bi-GRU and self-attention as final dialogue history representation:

$$\boldsymbol{C} = [\,\boldsymbol{H} \,||\, \boldsymbol{S}\,], \tag{2}$$

where $\boldsymbol{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_m\} \in \mathbb{R}^{m \times 2d_h}$ and $||$ is concatenation operation.

KB-over-KB Heterogeneous Graph Attention Network (KoK-HAN)

This section first introduces the graph building process, and then describes the heterogeneous message aggregation.

Graph building. Let a KoK-HAN graph be denoted as $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$, where $\mathcal{V}$ denotes node and $\mathcal{E}$ stands for edges.

Graph Node Building. In KoK-HAN, we treat knowledge entities from KB and words from dialogue history as nodes. Different node linear functions are adopted in order to distinguish different node features, which can be calculated as:

$$\tilde{\boldsymbol{C}} = f^{C}(\boldsymbol{C} + \phi^{emb}(\boldsymbol{X})), \tag{3}$$

$$\tilde{\mathcal{B}} = f^{\mathcal{B}}(\mathcal{B}), \tag{4}$$

where $\boldsymbol{Z} = [\tilde{\boldsymbol{C}}\,;\tilde{\mathcal{B}}] = [z_1, \ldots, z_{m+l}]$ can be directly employed to initialize the node representations ($l = \sum_{i=1}^{K}\{|\mathcal{R}^i| * |\mathcal{C}^i|\}$ denotes the number of entities in all multiple KBs); $f^C, f^{\mathcal{B}} = [f^{\mathcal{B}_1}, \ldots, f^{\mathcal{B}_K}]$ denote the different linear function for different KBs.

Graph Edge Building. For edge connections $\mathcal{E}$, as shown in Figure 2, we define the following types of edges to incorporate different granularity levels of information (i.e., high-order structure information, multiple KBs interaction information and dialogue-KB interaction information).

- ***Intra-KB connection:*** Intra-KB connection is used for capturing the related KB information within a KB. It has two parts, including *in-rows connection* and *cross-rows connection*. More specifically, (1) *in-rows connection* means that each node connects other nodes in the same row of a KB; (2) *cross-rows connection* denotes that we add an additional global node for each KB where it connects to all nodes in the KB. The global node can contain the whole KB information.

- ***Inter-KB connection:*** Inter-KB connection denotes all global KB nodes link to each other to incorporate cross-KBs information, which enables model to reason across multiple KBs. For example, we should construct $n(n-1)/2$ inter-KB edges if there exists n KB.
- ***Dialogue-KB Connection:*** To build interaction between dialogue history and the corresponding KBs, we introduce a dialogue-KB connection where a co-occurrence edge is created to connect related nodes if they co-occur in both dialogue history and the corresponding KBs. The interaction between dialogue and KB can be established through the dialogue-KB connection edge message flow.

Heterogeneous Message Aggregation In this section, we illustrate how information propagates over the graph to do reasoning over the KoK-HAN. Here, we give the aggregation and combination formulation of the message passing over the proposed KoK-HAN. More specifically, the message aggregation can be formulated as:

$$\tilde{\mathbf{h}}'_{i,k} = \sum_{r\in\mathcal{R}} \frac{1}{|\mathcal{N}_i^r|} f_r\Big(\sum_{j\in\mathcal{N}_i^r} \alpha_{ij}^r \mathbf{W}_h^r \tilde{\mathbf{h}}_{j,k}\Big), \tag{5}$$

where $\mathcal{R}$ is the set of all edge types, $\mathcal{N}_i^r$ is the neighbors of node i with edge type r and $\tilde{\mathbf{h}}_{j,k}$ is the node representation of node j in layer k ($\tilde{\mathbf{h}}_{j,1}$ initialized with z_j); $|\mathcal{N}_i^r|$ indicates the size of the neighboring set. f_r defines a transformation on the neighboring node representations; $\tilde{\mathbf{h}}'_{i,k}$ represents the aggregated information in layer k for node i.

After L layer aggregation, we acquire the final updated output representation $\tilde{\mathbf{H}}'_L = \{\tilde{\mathbf{h}}'_{\{1,L\}}, \dots, \tilde{\mathbf{h}}'_{\{m+l,L\}}\}$.

Multi-KBs Aware Decoder

This section shows how the multi-KBs aware decoder generates dialogue response word by word.

Decoder Initialization Given the context-aware encoding representations $\mathbf{C} = \{\mathbf{c}_1, \dots, \mathbf{c}_m\} \in \mathbb{R}^{m\times 2d_h}$, we follow Zhong, Xiong, and Socher (2018) to utilize an attention mechanism to summarize the sentence representation:

$$\begin{aligned} \alpha_i &= \operatorname{softmax}(\mathbf{W}_c\mathbf{c}_i + \mathbf{b}), i \in [1, m], && (6)\\ \mathbf{s} &= \textstyle\sum_i \alpha_i \mathbf{c}_i, && (7) \end{aligned}$$

where $\mathbf{W}_c$ and b are trainable parameters.

The obtained sentence representation $\mathbf{s}$ is used for initializing the decoder.

Response Generation There are two types of generation words in the task-oriented dialogues system, including the common word and knowledge entity. More specifically, the common word is produced by *generation module* while knowledge entity is retrieved from the corresponding KB via *knowledge-retriever module*. In our framework, following Wu, Socher, and Xiong (2019), we use a sketch decoder with sketch tag to control whether model generates the common word or knowledge entity.

At each decoding time step, the hidden state of decoder is not only used to predict the next token in vocabulary but also serves as the query vector to query the graph node output. During the inference time, if a sketch tag is generated, the pointer network will lexicalize the tag by picking up the expected output from graph node output $\tilde{\mathbf{H}}'_L$. Otherwise, the generated word from sketch decoder will be the output word.

Generation Module. We use a unidirectional GRU as the sketch decoder. At each decoding step t, the process can be denoted as:

$$\begin{aligned} h_t &= \operatorname{GRU}(\phi^{emb}(y_{t-1}), h_{t-1}), && (8)\\ y_t &= \operatorname{softmax}(\mathbf{W}h_t), && (9) \end{aligned}$$

where y_t denotes the predicted token at t timestep; h_t is the current decoder state; y_{t-1} is the previous output.

Knowledge-retriever Module. After a sketch tag is generated by Equation 9 at t timestep, we query KB for retrieving relevant entity to lexicalize the sketch tag. More specifically, the hidden vector h_t is regarded as the query vector for retrieving the corresponding KB, which is computed as:

$$ptr_t = \operatorname{softmax}(h_t^\top \tilde{\mathbf{H}}'_L). \tag{10}$$

The ptr_t is the predicted pointer distribution at time step t, which is treated as the probabilities of queried knowledge.

We select the word with the highest probability as the generated word to replace sketch tag. Take the last response in Figure 1 for example, the generation module first generates the sketched response "*Recommend you to @address.*" while knowledge-retriever module aims to replace *@address* with the specific knowledge entity `The Lama Temple` to produce the final response "*Recommend you to The Lama Temple*".

Experiments

Datasets

We conduct experiments on two datasets, including CrossWOZ (Zhu et al. 2020) and RiSAWOZ (Quan et al. 2020). To simulate multi-KBs settings, we keep the dialogues with multiple KBs in CrossWOZ and RiSAWOZ. On CrossWOZ, the dataset contains 2,331 dialogues for training, 231 dialogues for validation, and 224 dialogues for testing. On RiSAWOZ, it includes 3,486 dialogues for training, 210 dialogues for validation, and 230 dialogues for testing. Since there is no corresponding KB for each conversation in the original dataset, we manually equip each conversation with the supported KBs. In addition, we fix some unalignment problems between dialogue and KB. The updated datasets will be available for future research. For pre-processing, we use jieba [1] for CrossWOZ.

[1] https://github.com/fxsjy/jieba

	CrossWOZ		RiSAWOZ	
Model	BLEU	F1	BLEU	F1
Seq2Seq+Attn	6.27	7.53	6.41	10.30
Mem2Seq	8.22	11.33	10.43	21.83
GLMP	14.49	19.12	11.67	30.56
DDMN	14.45	29.23	14.01	29.71
DF-Net	15.76	29.57	14.96	34.51
Neuro-Symbolic RF	16.23	30.25	16.12	40.76
KoK-HAN	**21.26***	**38.07***	**20.61***	**50.87***

Table 1: Main results. The bolded number indicates the best performance. The numbers with * indicate that the improvement of our framework is statistically significant with $p < 0.05$ under t-test.

Experimental Settings

The dimensionality of all hidden units and embedding size of two datasets are 128. The batch size we use is selected from $\{16, 32\}$ and the dropout ratio is 0.2. We use AdamW (Loshchilov and Hutter 2019) to optimize the parameters in our model. All hyper-parameters are selected according to the validation set. The epoch for training is 50. All experiments are conducted at Tesla V100.

Baselines

We adopt the following state-of-the-art end-to-end task-oriented dialogue models to multi-KBs settings: (1) Seq2Seq+Attn (Eric et al. 2017): the model adopts an attention mechanism to query KB; (2) Mem2Seq (Madotto, Wu, and Fung 2018): the model adopts a memory network to encode knowledge entities; (3) GLMP (Wu, Socher, and Xiong 2019): the framework introduces a global-to-local pointer mechanism to query the KB; (4) DDMN (Wang et al. 2020): the models use a dual memory network for better selecting knowledge; (5) DF-Net (Qin et al. 2020): the model considers domain features to promote the multi-domain EToDs; (6) Neuro-Symbolic RF (Yang et al. 2022): the framework proposes a neuro-symbolic to perform explicit reasoning for EToDs, which achieves the promising performance.

For Seq2Seq+Attn, we re-implement their model to obtain the results. For other baselines, we retrain their open-source code to acquire performance. To facilitate the model with multi-KBs ability, we concatenate multiple KBs to an extended single-KB.

Automatic Evaluation

Following Wu, Socher, and Xiong (2019), we adopt the *Entity F1* and *BLEU* (Papineni et al. 2002) to evaluate the knowledge querying and fluent response generation ability.

Results are shown in Table 1. We have the following observations:

- The performance of prior best single-KB EToDs drop a lot from 64.5% to 40.76% (64.5% is the performance on single-KB setting (Yang et al. 2022) while 40.76% is the result attained by Neuro-Symbolic). This indicates the difficulty of multi-KBs settings requiring ability to reason across multiple KBs to generate the final response, which cannot be achieved by prior single-KB dialogue models;
- KoK-HAN achieves the state-of-the-art performance compared with all baselines, especially on F1 scores. Specifically, our framework outperforms Neuro-Symbolic RF 7.8% and 10.1% on F1 scores on CrossWOZ and RiSAWOZ, respectively. This demonstrates that KoK-HAN has successfully captured the relationship between multiple KBs, which is beneficial to multiple KBs reasoning.

	CrossWOZ		RiSAWOZ	
Model	BLEU	F1	BLEU	F1
w/o Intra-KB	19.34	27.49	18.34	47.35
w/o Inter-KB	17.18	29.60	17.76	45.82
w/o Dialogue-KB	17.91	23.87	17.12	34.75
Homogeneous GAT	17.27	21.55	15.85	26.81
KoK-HAN	**21.26**	**38.07**	**20.61**	**50.87**

Table 2: Ablation study.

Analysis

To analyze KoK-HAN in more depth, we perform comprehensive studies to answer the following research questions (RQs): (1) Does the *Intra-KB Connection* benefit to capture the high-order structure information? (2) Can the *Inter-KB Connection* achieve to reason across different KBs? (3) Can the *Dialogue-KB Connection* capture relationship between dialogue and KB? (4) Does heterogeneous graph attention network work better than homogeneous graph attention network? (5) Is KoK-HAN able to successfully work well in multi-KBs settings? (6) How the number of graph layers affects the final performance?

Answer1: Intra-KB Connection Helps to Capture the High-order Structure Information To verify the effectiveness of the *intra-KB connection*, we conduct comparison experiment by removing the *intra-KB connection* while the remained components are unchanged. Result is shown in Table 2 (*w/o intra-KB*). We can see 1.9-point and 2.3-point drops in terms of BLEU while 10.6% and 3.5% drop on F1 scores on CrossWOZ and RiSAWOZ, respectively. This is because that *Intra-KB* connection can incorporate the high-order structure information of KB, which enables model to reason within a KB.

Answer2: Inter-KB Connection Performs the Cross-KB Reasoning To investigate the impacts of *inter-KB connection*, we remove the *inter-KB connection* and keep other components unchanged. We refer it to *w/o inter-KB*. Table 2 shows the results. We observe that performances drop 8.5% and 5.1% on F1 scores on two datasets. This is because without the *inter-KB connection*, the model fails to perform the cross-KB knowledge reasoning, which harms the performances, especially for the dialogues requiring multi-KBs reasoning.

Answer3: Dialogue-KB Connection Builds the Relationship between Dialogue and KB We further investigate

(a) Dialogue for visualization

User	你好，我想吃美食街，帮我推荐一个人均消费在50-100元的餐馆，谢谢。 Hi, I want to eat Food Street, and recommend a restaurant with an average consumption of 50-100 yuan per capita, thank you.
Sys	为您推荐鲜鱼口老字号美食街，人均消费75元，有您想吃的美食街哦。 I recommend the **Xianyu Food Street** for you, with an average consumption of 75 yuan per capita.
	......
User	哦，我想在它附近的景点里找一个4.5分以上的，有吗？ Oh, I want to find one with **a score of 4.5** or higher among **the nearby attractions**, is there?
Sys	故宫和天安门广场都是不错的选择。 **The Forbidden City** and **Tiananmen Square** are both good choices.

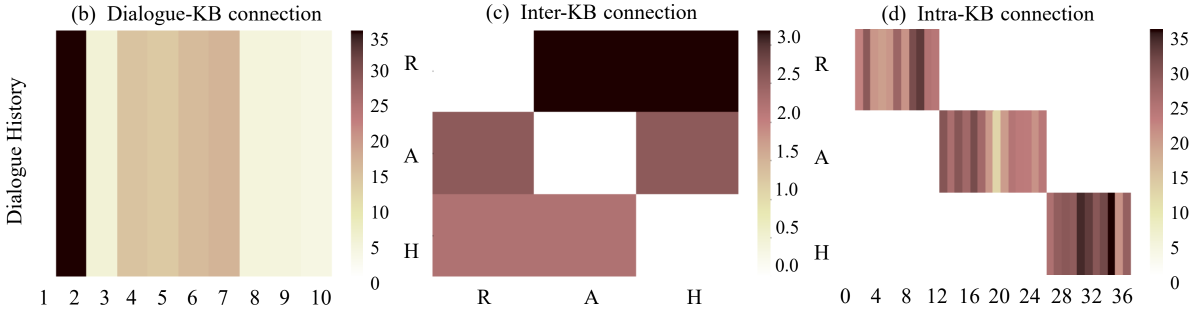


Figure 3: Visualization. Subplot (a) denotes dialogue history while (b), (c) and (d) show Heatmap for the three kinds of connection's weights, respectively. The axis on the right in each figure represents the color corresponding to the heat value. In subplot (b), it shows the attention weight between the dialogue history node itself and the entity nodes in KBs. We selected the top 10 with the highest weights for display. The 2nd node is `Xianyu Food Street`. As for subplot (c), it shows the weights between the global KB nodes in *Inter-KB connection*. `R`, `A` and `H` represent `Restaurant`,`Attraction` and `Hotel`. Subplot (d) shows the cross-rows weight in *Intra-KB connection*. There are a total of 40 nodes, among which 0, 1 and 2 are three global KB nodes. Remaining 37 nodes are from three KBs. In Attraction row, the 15th node is `The Forbidden City` and the 17th node is `Tiananmen Square`.

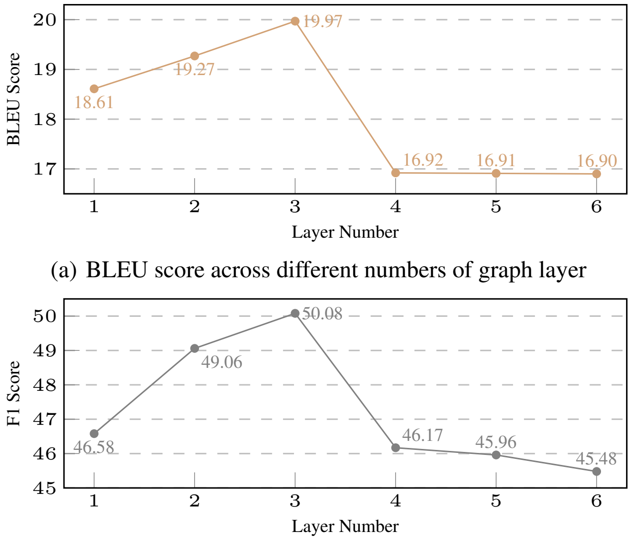
the effectiveness of *dialogue-KB connection*. We remove the *dialogue-KB connection* and refer it to *w/o Dialogue-KB*. Table 2 shows the results. We find that performances drop significantly on both BLEU and F1 scores on two datasets. We attribute it to the fact that the connection between dialogue history and KBs is breaking, resulting in optimizing KB and dialogue history separately, which seriously breaks the connection between dialogue and KB.

Answer4: Heterogeneous Graph Attention Network vs. Homogeneous Graph Attention Network Instead of adopting the heterogeneous GAT to model the *inter-KB* and *intra-KB* connection, we utilize the homogeneous GAT to model the interaction. This means that all edges are the same. We name it *Homogeneous GAT*, which is shown in Table 2. It can be seen that performance drop 16.5% and 24.1% on F1 scores on two datasets. We think that adopting homogeneous GAT cannot model different information, which may confuse the model to integrate information interaction and thus make it hard to retrieve relevant knowledge across KBs.

Answer5: KoK-HAN Achieves Better Performance over Multiple KBs Settings We further explore whether `KoK-HAN` can perform better reasoning across multiple KBs. We divided the RiSAWOZ dataset according to the number of KBs, then utilized different numbers of KBs' data to train `KoK-HAN` and `Neuro-Symbolic RF`. Results are presented in Table 3. As the number of KBs increases, reasoning in them is more difficult. It can be observed that at this time the performance gap between `KoK-HAN` and `Neuro-Symbolic RF` becomes larger. This demonstrates that our model can handle multiple KBs interactions better than single-KB model.

In addition, to provide an intuitive analysis for understanding why our framework works well in multi-KB settings, we conduct a visualization study, which is shown in Figure 3. From the dialogue in (a), the user asks for specific attractions near the restaurant `Xianyu Food Street` and our model tries to generate a response. Subplot (b) shows the attention weights between entity nodes in KBs and the dialogue history in *Dialogue-KB* connection. Since there are many words in the dialogue history, we integrate all the words in the dialogue history into one node for observation. We observe the model's attention on the `Xianyu Food Street` surpasses other nodes, representing our model localizes `Xianyu Food Street` in the dialogue history successfully. Next, to meet the user's request, the model needs to perform reasoning between *Attraction* KB and *Restaurant* KB to find surrounding attractions of `Xianyu Food Street`.

Subplot (c) shows weights between global KB nodes in *Inter-KB* connection. Higher weights can be seen between Attraction node and Restaurant node, representing our model successfully captures the interaction. Finally, in



(a) BLEU score across different numbers of graph layer

(b) F1 Score across different numbers of graph layer

Figure 4: Performance across layer numbers.

subplot (d), we show cross-rows weights of *Intra-KB* connection. We observe node The Forbidden City and Tiananmen Square has the highest weights, which indicates our model queries Attraction KB correctly.

Answer6: The Impact of the Number of Graph Layers We further investigate the effectiveness of the number of layers in KoK-HAN. Experiment results on RiSAWOZ are shown in Figure 4. We observe that more layers bring better performance when the number of graph layers is less than four. This is because the stacked interaction layer can help model to better query cross-KB knowledge querying.

Human Evaluation

In this section, we provide human evaluation of KoK-HAN and Neuro-Symbolic RF. We randomly generated 100 responses from the CrossWOZ test data. We hired 3 human experts and asked them to judge the quality of the responses. Following Qin et al. (2020), we evaluate the correctness, fluency, and humanlikeness metrics on a scale from 1 to 5.

As shown in Table 4, we observe that our framework outperforms Neuro-Symbolic RF on all metrics. This is consistent with the automatic evaluation, which further verifies the effectiveness of our proposed framework.

Related Work

End-to-End Task-Oriented Dialogue System

End-to-end Task-oriented Dialog systems (EToDs) have attracted more and more attention since they can be easily adapted to a new domain. With the success of the sequence-to-sequence (Seq2Seq) models in text generation, some EToDs use Seq2Seq-based model with attention mechanism (Eric et al. 2017; Lei et al. 2018; Wen et al. 2018) to implicitly query the knowledge from the corresponding KB. Another series of work (Madotto, Wu, and Fung 2018; Gangi Reddy et al. 2019; Wu, Socher, and Xiong 2019; He et al. 2020a; Yang, Zhang, and Erfani 2020; Wang et al. 2020; He et al. 2020b; Madotto et al. 2020) introduce the memory network (Sukhbaatar et al. 2015) to encode KB for better querying, which obtains promising performance. Qin et al. (2019b) propose a KB-retriever module to improve the entity consistency in system response. Recently, Qin et al. (2020) consider the different characteristics of different domains for querying KB in multi-domain EToDs. Yang et al. (2022) introduce neuro-symbolic to perform explicit reasoning for EToDs, which achieves the state-of-the-art performance. In contrast to their work, we consider the multi-KBs dialogues settings that are able to handle the need of a complex application in real-world scenarios while the above work mainly focuses on the single-KB grounded settings. To our knowledge, this is the first work to explore the multi-KBs settings for EToDs.

RiSAWOZ	2 KBs		3 KBs	
Model	BLEU	F1	BLEU	F1
KoK-HAN	20.54	52.66	11.48	40.75
Neuro-Symbolic RF	16.22	36.68	6.82	17.91
Δ	-4.32	-15.98	-4.66	-22.84

Table 3: Performance across different numbers of KB. Red and BLEU numbers denote gap performance between KoK-HAN and Neuro-Symbolic RF, respectively.

Model	Correct	Fluent	Humanlike
Neuro-Symbolic RF	2.1	3.9	3.8
Our framework	3.3	4.3	4.4

Table 4: Human evaluation.

Graph Neural Network for NLP

Recent years have witnessed remarkable success in graph neural network (GNN) for natural language understanding tasks. Specifically, GNN shows superior performance on text summarization task (Feng et al. 2021; Jing et al. 2021), sentiment analysis task (Huang et al. 2019; Liang et al. 2022), dialogue understanding task (Qin et al. 2021a,b) and language modeling (Meng et al. 2022). Inspired by the above work, we explore GNN for better capturing relationship across multiple KBs in EToDs.

Conclusion

We first pointed out the limitations of the existing single-KB end-to-end task-oriented dialogue systems (EToDs) and further explored the multi-KBs EToDs. In addition, we proposed a novel KB-over-KB heterogeneous graph network, which enables the model to reason across multiple KBs and capture high-order structure relationship information of KBs. To our knowledge, this is the first work to consider multiple KBs settings in EToDs. We hope this work can draw more attention to this complex real-world scenario.

Acknowledgements

We thank all anonymous reviewers for their constructive comments. We also thank Jieming Lou for his discussion and help. This work was supported via the grant 2020AAA0106501 and NSFC grants 62236004 and 61976072

References

Cho, K.; van Merriënboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1724–1734.

Eric, M.; Krishnan, L.; Charette, F.; and Manning, C. D. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 37–49. Saarbrücken, Germany: Association for Computational Linguistics.

Eric, M.; and Manning, C. 2017. A Copy-Augmented Sequence-to-Sequence Architecture Gives Good Performance on Task-Oriented Dialogue. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 468–473. Valencia, Spain: Association for Computational Linguistics.

Feng, X.; Feng, X.; Qin, B.; and Geng, X. 2021. Dialogue Discourse-Aware Graph Model and Data Augmentation for Meeting Summarization. In Zhou, Z.-H., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 3808–3814. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Gangi Reddy, R.; Contractor, D.; Raghu, D.; and Joshi, S. 2019. Multi-Level Memory for Task Oriented Dialogs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 3744–3754. Minneapolis, Minnesota: Association for Computational Linguistics.

Gao, S.; Takanobu, R.; Peng, W.; Liu, Q.; and Huang, M. 2021. HyKnow: End-to-End Task-Oriented Dialog Modeling with Hybrid Knowledge Management. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 1591–1602. Online: Association for Computational Linguistics.

Gu, J.; Wu, Q.; Wu, C.; Shi, W.; and Yu, Z. 2021. PRAL: A Tailored Pre-Training Model for Task-Oriented Dialog Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, 305–313. Online: Association for Computational Linguistics.

Ham, D.; Lee, J.-G.; Jang, Y.; and Kim, K.-E. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 583–592.

He, W.; Yang, M.; Yan, R.; Li, C.; Shen, Y.; and Xu, R. 2020a. Amalgamating Knowledge from Two Teachers for Task-oriented Dialogue System with Adversarial Training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3498–3507. Online: Association for Computational Linguistics.

He, Z.; He, Y.; Wu, Q.; and Chen, J. 2020b. Fg2seq: Effectively Encoding Knowledge for End-To-End Task-Oriented Dialog. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE.

Hosseini-Asl, E.; McCann, B.; Wu, C.; Yavuz, S.; and Socher, R. 2020. A Simple Language Model for Task-Oriented Dialogue. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Huang, L.; Ma, D.; Li, S.; Zhang, X.; and Wang, H. 2019. Text Level Graph Neural Network for Text Classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3444–3450. Hong Kong, China: Association for Computational Linguistics.

Jing, B.; You, Z.; Yang, T.; Fan, W.; and Tong, H. 2021. Multiplex Graph Neural Network for Extractive Text Summarization. In *Proc. of EMNLP*.

Kulhánek, J.; Hudeček, V.; Nekvinda, T.; and Dušek, O. 2021. Augpt: Dialogue with pre-trained language models and data augmentation. *arXiv preprint arXiv:2102.05126*.

Lee, Y. 2021. Improving End-to-End Task-Oriented Dialog System with A Simple Auxiliary Task. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1296–1303.

Lei, W.; Jin, X.; Kan, M.-Y.; Ren, Z.; He, X.; and Yin, D. 2018. Sequicity: Simplifying Task-oriented Dialogue Systems with Single Sequence-to-Sequence Architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1437–1447. Melbourne, Australia: Association for Computational Linguistics.

Liang, B.; Su, H.; Gui, L.; Cambria, E.; and Xu, R. 2022. Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks. *Knowledge-Based Systems*, 235: 107643.

Loshchilov, I.; and Hutter, F. 2019. Decoupled Weight Decay Regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Madotto, A.; Lin, Z.; Wu, C.-S.; Shin, J.; and Fung, P. 2020. Attention over parameters for dialogue systems. *arXiv preprint arXiv:2001.01871*.

Madotto, A.; Wu, C.-S.; and Fung, P. 2018. Mem2Seq: Effectively Incorporating Knowledge Bases into End-to-End Task-Oriented Dialog Systems. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 1468–1478. Melbourne, Australia: Association for Computational Linguistics.

Meng, Y.; Zong, S.; Li, X.; Sun, X.; Zhang, T.; Wu, F.; and Li, J. 2022. GNN-LM: Language Modeling based on Global Contexts via GNN. In *International Conference on Learning Representations*.

Olabiyi, O. O.; Bhattarai, P.; Bruss, C. B.; and Kulis, Z. 2020. DLGNet-Task: An End-to-end Neural Network Framework for Modeling Multi-turn Multi-domain Task-Oriented Dialogue. *arXiv preprint arXiv:2010.01693*.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 311–318. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Peng, B.; Li, C.; Li, J.; Shayandeh, S.; Liden, L.; and Gao, J. 2021. SOLOIST: Building Task Bots at Scale with Transfer Learning and Machine Teaching. In *Transactions of the Association for Computational Linguistics*.

Qin, L.; Che, W.; Li, Y.; Wen, H.; and Liu, T. 2019a. A Stack-Propagation Framework with Token-Level Intent Detection for Spoken Language Understanding. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2078–2087. Hong Kong, China: Association for Computational Linguistics.

Qin, L.; Li, Z.; Che, W.; Ni, M.; and Liu, T. 2021a. Co-GAT: A Co-Interactive Graph Attention Network for Joint Dialog Act Recognition and Sentiment Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15): 13709–13717.

Qin, L.; Liu, Y.; Che, W.; Wen, H.; Li, Y.; and Liu, T. 2019b. Entity-Consistent End-to-end Task-Oriented Dialogue System with KB Retriever. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 133–142. Hong Kong, China: Association for Computational Linguistics.

Qin, L.; Wei, F.; Xie, T.; Xu, X.; Che, W.; and Liu, T. 2021b. GL-GIN: Fast and Accurate Non-Autoregressive Model for Joint Multiple Intent Detection and Slot Filling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 178–188. Online: Association for Computational Linguistics.

Qin, L.; Xu, X.; Che, W.; Zhang, Y.; and Liu, T. 2020. Dynamic Fusion Network for Multi-Domain End-to-end Task-Oriented Dialog. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 6344–6354. Online: Association for Computational Linguistics.

Quan, J.; Zhang, S.; Cao, Q.; Li, Z.; and Xiong, D. 2020. RiSAWOZ: A Large-Scale Multi-Domain Wizard-of-Oz Dataset with Rich Semantic Annotations for Task-Oriented Dialogue Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 930–940. Online: Association for Computational Linguistics.

Sukhbaatar, S.; Szlam, A.; Weston, J.; and Fergus, R. 2015. End-To-End Memory Networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NIPS*.

Wang, J.; Liu, J.; Bi, W.; Liu, X.; He, K.; Xu, R.; and Yang, M. 2020. Dual Dynamic Memory Network for End-to-End Multi-turn Task-oriented Dialog Systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4100–4110. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Wen, H.; Liu, Y.; Che, W.; Qin, L.; and Liu, T. 2018. Sequence-to-Sequence Learning for Task-oriented Dialogue with Dialogue State Representation. In *Proceedings of the 27th International Conference on Computational Linguistics*, 3781–3792. Santa Fe, New Mexico, USA: Association for Computational Linguistics.

Wu, C.; Socher, R.; and Xiong, C. 2019. Global-to-local Memory Pointer Networks for Task-Oriented Dialogue. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Yang, S.; Zhang, R.; and Erfani, S. 2020. GraphDialog: Integrating Graph Knowledge into End-to-End Task-Oriented Dialogue Systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1878–1888. Online: Association for Computational Linguistics.

Yang, S.; Zhang, R.; Erfani, S.; and Lau, J. H. 2022. An Interpretable Neuro-Symbolic Reasoning Framework for Task-Oriented Dialogue Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4918–4935. Dublin, Ireland: Association for Computational Linguistics.

Yang, Y.; Li, Y.; and Quan, X. 2021. UBAR: Towards Fully End-to-End Task-Oriented Dialog System with GPT-2. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 14230–14238.

Young, S. J.; Gasic, M.; Thomson, B.; and Williams, J. D. 2013. POMDP-Based Statistical Spoken Dialog Systems: A Review. *Proceedings of the IEEE*, 101(5): 1160–1179.

Zhong, V.; Xiong, C.; and Socher, R. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1458–1467.

Zhu, Q.; Huang, K.; Zhang, Z.; Zhu, X.; and Huang, M. 2020. CrossWOZ: A Large-Scale Chinese Cross-Domain Task-Oriented Dialogue Dataset. *Transactions of the Association for Computational Linguistics*, 8: 281–295.