UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases into End-to-End Task-Oriented Dialogue Systems

Shiquan Yang[1] , **Rui Zhang**[2*] , **Sarah Erfani**[1] and **Jey Han Lau**[1]
[1]The University of Melbourne, Australia
[2]Tsinghua University
shiquan@student.unimelb.edu.au, rayteam@yeah.net, {sarah.erfani, jeyhan.lau}@unimelb.edu.au

Abstract

Knowledge bases (KBs) are usually essential for building practical dialogue systems. Recently we have seen rapidly growing interest in integrating knowledge bases into dialogue systems. However, existing approaches mostly deal with knowledge bases of a single modality, typically textual information. As today's knowledge bases become abundant with multimodal information such as images, audios and videos, the limitation of existing approaches greatly hinders the development of dialogue systems. In this paper, we focus on task-oriented dialogue systems and address this limitation by proposing a novel model that integrate external multimodal KB reasoning with pre-trained language models. We further enhance the model via a novel multi-granularity fusion mechanism to capture multi-grained semantics in the dialogue history. To validate the effectiveness of the proposed model, we collect a new large-scale (14K) dialogue dataset *MMDialKB*, built upon multimodal KB. Both automatic and human evaluation results on *MMDialKB* demonstrate the superiority of our proposed framework over strong baselines.

1 Introduction

Incorporating knowledge bases (KBs) into dialogue systems has attracted increasing attentions over the past few years, particularly on the development of many end-to-end dialogue models such as Sequicity [Lei *et al.*, 2018] and GLMP [Wu *et al.*, 2019]. However, existing approaches can only deal with knowledge bases of a single modality, typically textual information. As today's knowledge bases become abundant with multimodal information such as images, audios and videos in addition to texts, the limitation of existing approaches greatly hinders the development of dialogue systems. In some scenarios, generating dialogue responses requires using knowledge from multiple modalities, and only using single modality is insufficient. For example in Figure 1, a user is asking for recommending a suitable restaurant for the anniversary.

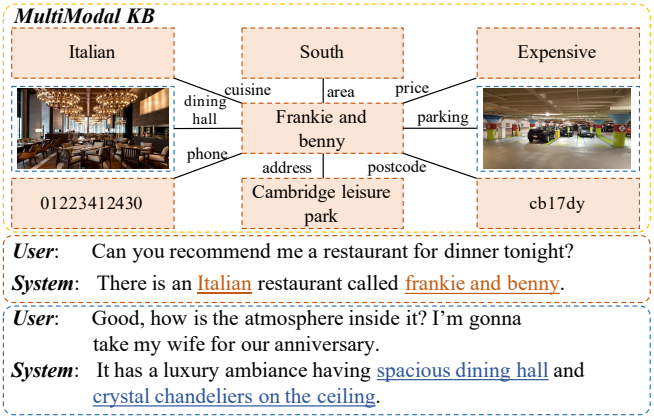

Figure 1: An example from *MMDialKB* dataset. The red square denotes the textual entities and the blue square denotes the visual entities in the multimodal KB. The goal is to generate the system responses provided the dialogue history and the multimodal KB.

The detailed information about the restaurant like the atmosphere is required in order to make recommendation for the user. However, this information is not available in the textual information provided by the knowledge base. The nodes in the knowledge base that contain the images of the restaurants are needed to answer questions about the restaurant's atmosphere. Without the ability to integrate multimodal entities in the KB, the system will not be able to answer such questions.

However, existing approaches cannot handle multimodal KBs due to the lack of appropriate mechanism to incorporate the multimodal information in the KBs. To address this challenge, in this paper, we focus on task-oriented dialogue systems and propose to incorporate multimodal KBs into task-oriented dialogue systems. We focus on two modalities, texts and images, although our model is straightforwardly generalizable to more modalities. We develop a novel model called **UniMF**, a **Uni**fied framework that integrates pre-trained language models with external knowledge retrievers, both visual and textual, to sustain reasoning over multimodal KBs and further enhanced by a **M**ulti-granularity **F**usion mechanism.

In order to train our model, we create a new large-scale dataset integrated with a multimodal KB called **MMDialKB**, short for "**M**ulti**m**odal **Dial**ogues with **KB**s". Specifically,

*Rui Zhang is the corresponding author.

MMDialKB has several distinguished features: 1) *Novelty*. To the best of our knowledge, this is the first task-oriented dialogue dataset that include multimodal KB and collect dialogues upon it. 2) *High-quality*. We have conducted careful quality control to ensure the consistency and reliability of the collected dialogue data. Also, all the images in the dataset are high-quality selected from the mainstream image search engines and checked manually by us. 3) *Challenging*. To generate appropriate system responses, the model needs to reason about the relations between textual concepts appeared in the dialogue and their image counterparts in the KBs, e.g. 'it' in the user utterance at the 2nd turn of the example in Figure 1. Also, co-reference resolution across modalities increases its complexity. These altogether make *MMDialKB* a challenging dataset for research and a reliable benchmark for assessment. We conduct systematical empirical studies over *MMDialKB* dataset. Both automatic and human evaluation results demonstrate the superiority of our proposed framework over strong baselines.

The contributions of this paper are summarized as follows:

- This is the first attempt to integrate multimodal knowledge bases into end-to-end task-oriented dialogue systems. This task is challenging since models are required to reason about the relations between dialogue context and the various visual and textual entities in KBs.
- We propose a novel framework to solve the above problem. Specifically, we propose to integrate external knowledge retrievers with pre-trained language model to extend its ability to incorporate external multimodal KBs. We further introduce a multi-granularity fusion mechanism to capture multi-grained semantics in the dialogue history.
- We create the first large-scale dataset for training task-oriented dialogue systems with multimodal KBs.
- We conduct systematical empirical studies over *MMDialKB* dataset. Both quantitative and qualitative results demonstrate the superiority of our proposed framework.

2 Related Work

Task-Oriented dialogue modeling has been one of the most popular topics over the past few years [Wu *et al.*, 2019; Huang *et al.*, 2020]. Bordes *et al.* early explored end-to-end memory networks [Sukhbaatar *et al.*, 2015] to handle KBs and shown promising results. To produce more flexible responses, generative dialogue model is proposed [Zhao *et al.*, 2017], which formulates the response generation problem as a translation task and employ the sequence-to-sequence (Seq2Seq) models to generate responses. Seq2Seq models have shown to be effective in language modeling but they struggle to incorporate external KB into responses. To mitigate this issue, Eric and Manning has enhanced the Seq2Seq model by integrating copy mechanism. Madotto *et al.* and Wu *et al.* combines the idea of pointer with memory networks and obtained improved performance.

On the other spectrum, there are also several attempts to incorporate multimodal information into dialogue systems [Wu *et al.*, 2016]. Specifically, Visual Dialogue [Das *et al.*, 2017] develops a task where an agent conducts a dialogue with a human about the visual content based on the given image. Image-grounded conversations [Mostafazadeh *et al.*, 2017] proposes to conduct a dialogue over a shared image between two humans. GuessWhat [De Vries *et al.*, 2017] introduces an image-grounded game where the goal is to locate the object by asking a sequence of questions. However, all these tasks are limited by the lack of external knowledge bases which is usually important to offer practical information to the users. Also, they are all intended to perform reasoning with a single image while *MMDialKB* is intended to perform visual reasoning across multiple images. A closely-related work to ours is SIMMC [Moon *et al.*, 2020] which mimic the shopping scenarios in a virtual reality environment and generate responses based on co-observed image and the user utterances. However, our work is different from theirs since we focuses more on the multimodal KBs which is usually an important infrastructure for practical dialogue systems.

3 Approach

Our framework consists of three components: Local Semantics Encoder, Knowledge Retriever, and Response Generator. Local Semantics Encoder takes the dialogue history as input and encode the conversation history into contextual representations based on RoBERTa. To incorporate both the textual and visual external knowledge, we design VisualKnowledgeRetriever and TextualKnowledgeRetriever to perform reasoning over both visual and textual entities in the KB. We further utilize a multi-granularity fusion layer to capture multi-grained semantics in the dialogue history. Finally, Response Generator takes the contextual representations, the extracted knowledge signals and the multi-grained semantics as inputs, and generates system response by finetuning over the pre-trained language model UniLM. Figure 2 illustrates the overall structure of the proposed framework. Next, we describe each component in details.

3.1 Local Semantics Encoder

We use the large-scale pre-trained language model RoBERTa [Liu *et al.*, 2019] to serve as the backbone of our encoder in order to capture the semantics of the conversation history. Specifically, it takes the dialogue history as input, and we insert several special tokens at the start and the end of every turn of the dialogue history to indicate the spans of each individual dialogue turn. For example, we add <s> at the start of each turn and add </s> at the end of each turn. In this way, we can represent each turn of dialogue history by utilizing the representation of the corresponding <s> token which captures the aggregated semantics of the word tokens that follows it. We then concatenate all the individual dialogue turns together and utilize RoBERTa to encode the concatenated sequence. We obtain the local semantic representations for each dialogue turn $\mathbf{e}^{\mathrm{D}}_1$,$\mathbf{e}^{\mathrm{D}}_2$,...,$\mathbf{e}^{\mathrm{D}}_N$ by extracting the representation of each <s> token, where N is the total number of turns in the dialogue history. All these vectorized representations are of d dimensions (768 for RoBERTa-base). The obtained local semantic vectors are utilized to attend the external knowledge to incorporate multimodal knowledge into the response generation process.

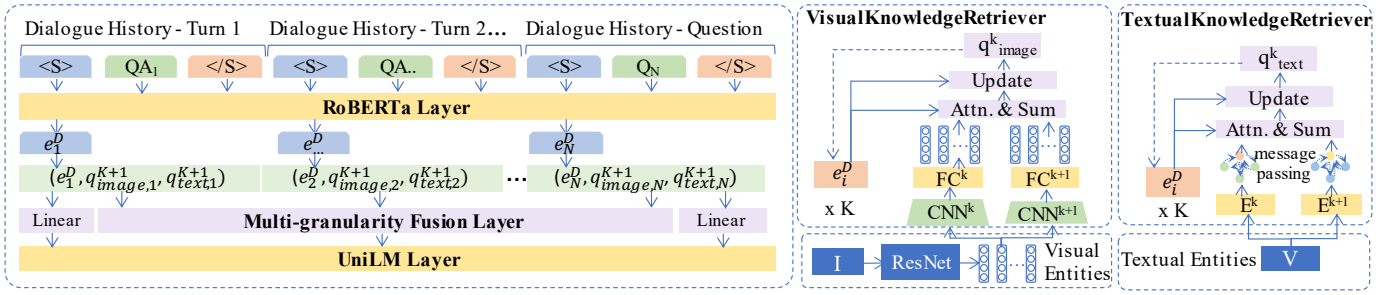

Figure 2: The proposed framework. Left part shows the unified model structure and right part shows the proposed knowledge retrievers.

3.2 Knowledge Retriever Over Multimodal KB

To effectively incorporate the multimodal external knowledge including both visual and textual knowledge in the external KB, we introduce two knowledge retrievers – i.e., VisualKnowledgeRetriever and TextualKnowledgeRetriever, aimed to incorporate different modalities into the reasoning process.

To incorporate the visual inputs from the image entities stored in the external KB, we design a novel module called VisualKnowledgeRetriever. This module takes a set of images $\mathbf{I} = \{\mathrm{I}_1,...,\mathrm{I}_{\mathrm{N_{image}}}\}$ as inputs and utilize pre-trained ResNet to transform these images into embedding space using:

$$\mathbf{e}^{\mathrm{I}}_i = \mathrm{ResNet}(\mathrm{I}_i),\ i = 1, 2, ..., \mathrm{N_{image}} \tag{1}$$

where $\mathrm{N_{image}}$ is the total number of visual entities in the external knowledge. Specifically, we use the activations of the last convolution layer of ResNet-101 to initialize the representation of these visual entities. We then employ a multi-hop structure to perform visual reasoning over the visual representations $\{\mathbf{e}^{\mathrm{I}}_i\}^{\mathrm{N_{image}}}_{i=1}$, which is inspired by the memory networks [Sukhbaatar *et al.*, 2015]. Existing memory-based architectures are powerful in reasoning due to its multi-hop nature. However, they struggle to perform reasoning over visual inputs such as images. To this end, we propose to extend the classic memory networks to a more broad setting which aims to handle multimodal inputs especially images.

Formally, the VisualKnowledgeRetriever contains a set of trainable layers $\mathbf{L} = \{\{\mathbf{C}^1,\mathbf{W}^1\},...,\{\mathbf{C}^{K+1},\mathbf{W}^{K+1}\}\}$, where each layer $\mathbf{L}^k = \{\mathbf{C}^k,\mathbf{W}^k\}$ is composed of one convolution layer $\mathbf{C}^k$ followed by a linear projection $\mathbf{W}^k$ to transform the initialized visual representations $\mathbf{e}^{\mathrm{I}}$ into the same space with the representations of the encoded dialogue history (i.e., $\mathbf{e}^{\mathrm{D}}_1$ to $\mathbf{e}^{\mathrm{D}}_N$). K is the maximum number of reasoning hops. Our best results are achieved by using K=3 in our experiments.

During inference, the module loops over K hops on an input set of images. At each hop k, we first obtain two sets of transformed visual representations[1] $\mathbf{e}^{\mathrm{I},k}$ and $\mathbf{e}^{\mathrm{I},k+1}$ by applying layers $\mathbf{L}^k$ and $\mathbf{L}^{k+1}$ correspondingly to the original image embeddings $\mathbf{e}^{\mathrm{I}}$:

$$\mathbf{e}^{\mathrm{I},k}_i = \mathbf{W}^k(\mathbf{C}^k(\mathbf{e}^{\mathrm{I}}_i)),\ i = 1, 2, ..., \mathrm{N_{image}} \tag{2}$$

[1]We employ a weight-tying strategy between different hops of computations inspired by end-to-end memory networks [Sukhbaatar *et al.*, 2015]

$$\mathbf{e}^{\mathrm{I},k+1}_i = \mathbf{W}^{k+1}(\mathbf{C}^{k+1}(\mathbf{e}^{\mathrm{I}}_i)),\ i = 1, 2, ..., \mathrm{N_{image}} \tag{3}$$

The first set of the representations $\mathbf{e}^{\mathrm{I},k}$ is aimed to compute the correlations distribution between the query vector and the visual entities, while the second set of representations is utilized to summarize the important visual information for further reasoning. Specifically, we utilize a query vector $\mathrm{q}^k_{\mathrm{image}}$ ($\mathrm{q}^0_{\mathrm{image}} = \mathbf{e}^{\mathrm{D}}_i$) as input, and compute the attentions between query $\mathrm{q}^k_{\mathrm{image}}$ and visual representations $\mathbf{e}^{\mathrm{I},k}$ using:

$$\mathrm{p}^k_{i,\mathrm{image}} = \mathrm{Softmax}((\mathrm{q}^k_{\mathrm{image}})^{\mathsf{T}}\mathbf{e}^{\mathrm{I},k}_i)) \tag{4}$$

We then get the extracted visual information $\mathrm{o}^k_{\mathrm{image}}$ at hop k and the updated query for the next hop $\mathrm{q}^{k+1}_{\mathrm{image}}$ by combining the attentions $\mathrm{p}^k_{i,\mathrm{image}}$ with visual representations $\mathbf{e}^{\mathrm{I},k+1}$ using:

$$\mathrm{o}^k_{\mathrm{image}} = \sum_i \mathrm{p}^k_{i,\mathrm{image}}\mathbf{e}^{\mathrm{I},k+1}_i,\quad \mathrm{q}^{k+1}_{\mathrm{image}} = \mathrm{q}^k_{\mathrm{image}} + \mathrm{o}^k_{\mathrm{image}} \tag{5}$$

$\mathrm{q}^{K+1}_{\mathrm{image}}$ can be seen as the extracted semantically related visual information and is used as the inputs for finetuing UniLM to generate system responses.

To support reasoning over the textual knowledge in KBs, we present TextualKnowledgeRetriever based on graph neural networks. This component takes a set of textual entities $\mathrm{V} = \{\mathrm{v}_1,...,\mathrm{v}_{\mathrm{N_{text}}}\}$ as inputs. $\mathrm{N_{text}}$ is the total number of textual entities in the KB. Like VisualKnowledgeRetriever, we perform multi-hop reasoning over all the textual entities to extract the contextual relevant textual information. Formally, we first define a set of embedding layers $\mathbf{E} = \{\mathbf{E}^1,...,\mathbf{E}^{K+1}\}$ in order to transform all the textual entities into embedding space. At each hop k, we encode all the textual entities by applying $\mathbf{E}^k$ and get the representation of the i-th textual entity v_i using $\mathbf{e}^{\mathrm{T},k}_i = \mathbf{E}^k(v_i)$. We then perform self-attention between those entities to capture the interrelationship information among them. Following this, we update the representation of each entity via weighted sum of its first-order neighbors using the following equations:

$$e_{ij} = \varphi\left((\mathrm{Q})^{\mathrm{T}}[\mathbf{e}^{\mathrm{T},k}_i, \mathbf{e}^{\mathrm{T},k}_j]\right) \tag{6}$$

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{\mathrm{k}\in\mathrm{N}_i}\exp(e_{ik})},\ \left(\mathbf{e}^{\mathrm{T},k}_i\right)' = \sum_{\mathrm{j}\in\mathrm{N}_i}\alpha_{ij}\mathbf{e}^{\mathrm{T},k}_j \tag{7}$$

where φ is LeakyReLU activation function, Q is parametrized weight vector of the attention mechanism, N_i is the first-order neighbors of entity *i* (including *i*), exp is exponential function. We then utilize a query vector q^k_{text} ($q^0_{\text{text}} = \mathbf{e}^{\text{D}}_i$) to attend the updated textual entities and obtain the attentions using:

$$p^k_{i,\text{text}} = \text{Softmax}((q^k_{\text{text}})^{\top}(\mathbf{e}^{\text{T},k}_i)')) \quad (8)$$

We extract the relevant textual knowledge and update the query for next hop reasoning using:

$$o^k_{\text{text}} = \sum_i p^k_{i,\text{text}}(\mathbf{e}^{\text{T},k+1}_i)', \quad q^{k+1}_{\text{text}} = q^k_{\text{text}} + o^k_{\text{text}} \quad (9)$$

We use q^{K+1}_{text} as the extracted context-aware textual knowledge and use it as another input for UniLM finetuning.

3.3 Multi-Granularity Fusion Layer

Since the overall semantics for every turn of the dialogue history should not only consider the utterances themselves, but also include the background knowledge which is related to the context determined by the utterances. To incorporate such knowledge into the representations for the dialogue history, we utilize the local semantic vectors $\mathbf{e}^{\text{D}}_1$,$\mathbf{e}^{\text{D}}_2$,...,$\mathbf{e}^{\text{D}}_N$ as query vectors to attend both knowledge retrievers.

Specifically, we first use $\mathbf{e}^{\text{D}}_i$ as the query vector to attend the VisualKnowledgeRetriever and acquire q^{K+1}_{image} which represents the relevant visual knowledge to the context at turn *i*. Then $\mathbf{e}^{\text{D}}_i$ again is utilized to query the TextualKnowledgeRetriever and obtain the related textual knowledge q^{K+1}_{text}. Finally, we concatenate the local semantic vector $\mathbf{e}^{\text{D}}_i$, the related visual information q^{K+1}_{image} and the textual information q^{K+1}_{text} to get a combined representation R_i for the semantics of the dialogue history at turn *i*.

To encourage richer semantic representations for the dialogue history, we further design a cross-turn multi-granularity fusion layer to aggregate the semantics for the dialogue history at multiple scales, as dialogue topics usually flow over time. Formally, we take the set of combined representations $R = \{R_1,...,R_N\}$ as input. We then define M which is a variable size block taking sizes from $S = \{S_1,...,S_t\}$, where S is the set of all valid block sizes and each $\{S_i\}^t_{i=1}$ is a positive integer. Next, we loop over the *t* block sizes and utilize them to split representations R into different semantic blocks with various sizes. Specifically, for size S_i, we first obtain the semantic blocks set B_{S_i} from the input representations R using:

$$B_{S_i} = \{\{R_1, ..., R_{S_i}\}, ..., \{R_{N-S_i+1}, ..., R_N\}\} \quad (10)$$

For each semantic block in B_{S_i}, we utilize average pooling operation followed by a linear projection W_1 to aggregate the information across turns and obtain the block representations $\{F^{S_i}_1,...,F^{S_i}_{N-S_i+1}\}$, where each $\{F^{S_i}_k\}^{N-S_i+1}_{k=1}$ is computed using:

$$F^{S_i}_k = W_1(\text{Pooling}(R_k, ..., R_{k+S_i-1})) \quad (11)$$

Following the same procedure, we iterate over the *t* granularities corresponding to the block sizes taking values in S, and get the corresponding semantic representations about the dialogue history at every granularity. The final semantic representations for all the granularities are utilized as the inputs of the response generation module. Intuitively, different utterance blocks can vary in importance to the overall semantics of the dialogue context. We utilize multi-granularity fusion scheme to allow capturing multi-scale fine-grained semantics in the dialogue history at block-level of various sizes, which enriches the feature representations of the dialogue history.

3.4 Response Generation

To generate the system response conditioned both on the dialogue history and the relevant external knowledge, we utilize the pre-trained language model UniLM [Dong *et al.*, 2019] and finetune to generate the responses. Specifically, we employ the sequence-to-sequence mode of the UniLM which is finetuned to perform language generation tasks conditioned on the source inputs. Under this mode, all the source inputs are allowed to interact with each other and the generated token at the current timestep can interact with both all the source inputs and the previously generated tokens by using masked attention mechanism over transformer layers. In this way, it encourages sufficient interactions between encoding and decoding procedures.

We construct the input sequence for UniLM as follows. Firstly, we concatenate the local semantic vectors $\mathbf{e}^{\text{D}}_1$,$\mathbf{e}^{\text{D}}_2$,...,$\mathbf{e}^{\text{D}}_N$ with their corresponding visual and textual knowledge signals followed by linear layer to form the per-turn representations of the the dialogue history. We then concatenate the outputs of the multi-granularity fusion layer with the per-turn representations as the input sequence, and finetune UniLM over the target responses with cross-entropy losses between the predicted tokens and the ground-truth tokens.

4 MMDialKB Dataset

To verify the effectiveness of the proposed model, we create a new large-scale (14K) dialogue dataset called *MMDialKB* [2] which provides human–human dialogues upon a multimodal KB in the restaurant domain and is used for empirical studies in this paper. Next we introduce our data collection procedure and the dataset statistics in detail.

We first utilize the KB data provided in the *MultiWOZ 2.1* dataset to construct the textual part of the multimodal KB. We then collect images for the textual entities in the knowledge base from mainstream image search engines, e.g., Google, Bing. To ensure the quality of the images, we manually check each one of them and discard those that don't meet our standards, e.g. blurred images, images with words and watermarks. Overall, we have collected 6589 images to build the KB. Finally, we extend the textual KB by manually matching the images with the textual entities to obtain a multimodal KB. We then collect dialogue utterances via AMT. Specifically, we utilize the open-source library ParlAI[3] to collect the dialogue data since it provides several useful functionalities (e.g. backend messaging, data storage, pool of workers maintaining) for conveniently deploying live chats on AMT.

[2] https://github.com/ruizhang-ai/MMDialKB

[3] https://github.com/facebookresearch/ParlAI

Specifically, we match two workers and ask them to conduct a dialogue where the user role is instructed to ask questions and the agent role needs to answer them based on the provided KB information. See more details about the data collection setup, interface, and quality control in Appendix E.

4.1 Dataset Statistics

We have conducted data statistics of the collected dataset and make comparisons with popular datasets such as *SMD* and *MultiWOZ 2.1*. Table 1 has shown the comparison results. Compared to existing popular datasets such as *SMD* and *MultiWOZ 2.1*, *MMDialKB* is the only dataset that has multimodal KB and have dialogues involving visual information. Particularly, there are 68% of the dialogue turns related to the visual entities in the KB while none of the other two datasets contains such data. Among the 68% dialogue turns that require visual knowledge, about 59% require only visual knowledge while about 41% require both visual and textual knowledge. There is a good mix and balance of different types of questions and they are not significantly biased. Therefore, the dataset is well suited to verify the proposed problem. More details about the question and answer distributions and comparisons are included in Appendix F. In total, *MMDialKB* contains 14420 dialogues, and we split 10000 for training, 1420 for validation and 3000 as test.

Metrics	**SMD**	**MultiWOZ**	**MMDialKB**
With KB?	✓	✓	✓
With Multimodal KB?	×	×	✓
# turns per dialog	5.25	13.46	4.83
# tokens per turn	8.02	13.13	10.74
# of dialogue turns	12,732	113,556	69,648
# instances in total	3,031	10,438	14,420

Table 1: Statistics for *SMD*, *MultiWOZ* and *MMDialKB*.

5 Experiments

5.1 Baselines and Metrics

We compare our model with the following state-of-the-art models in task-oriented dialogue systems, including: (1) **Mem2Seq** [Madotto *et al.*, 2018]: the model takes dialogue history and textual KB as inputs, and generates system responses either by selecting an input token or by generating tokens from vocabulary; (2) **GLMP** [Wu *et al.*, 2019]: the model utilizes a global pointer mechanism over textual KB entities to improve copy accuracy and uses standard GRU with template-based method to generate system responses; (3) **DF-Net** [Qin *et al.*, 2020]: the model uses a shared-private structure to exploit cross-domain knowledge and favors transferability across domains; (4) **GraphDialog** [Yang *et al.*, 2020]: the framework incorporate graph knowledge both in dialogue history and the textual KB to improve the quality of generated responses. We use three common evaluation metrics including BLEU [Papineni *et al.*, 2002], Entity F1 [Eric *et al.*, 2017] and Perplexity for evaluations.

5.2 Implementation Details

We finetune RoBERTa-base model [Wolf *et al.*, 2019] with Adam [Kingma and Ba, 2014] optimizer with a learning rate of 5e-5, batch size of 16, drop out rate of 0.2. We try 1,2,3,4,5,6 for the maximum number of hops K and try all the combinations selected from sizes [2,3,4] for block sizes S, and find the best combination is K = 3 and S = [2,3] based on the validation set results. For the response generation module, we finetune UniLM1.2-based-uncased model [Dong *et al.*, 2019] 20 epochs with a batch size of 16, a learning rate of 2e-5, and a beam size of 5 for decoding during inference. We repeat all the experiments 10 times with different random seeds and report the average results. The model is trained on a 8-core server with 64 GB memory and an NVIDIA GeForce RTX 2080 Ti GPU. All the training can be done in one day.

5.3 Results

Table 2 has shown the results for both baselines and ours on *MMDialKB* dataset. We evaluate all the models using different combinations of the inputs to ablate the effects of every type of inputs. For all the baselines, we report the results using dialogue history with and without textual KBs as inputs since they can't handle multimodal inputs as ours. It can be seen that our model consistently outperforms all the baselines under all settings, i.e., only use dialogue history as inputs and plus textual KB as inputs. Specifically, our best performing version with K = 3 has achieved an average about 3% absolute improvement over all the baselines including the state-of-the-art. This verifies the effectiveness of our proposed model. When incorporating the visual knowledge, the performance gain has become even larger for our model compared to the baselines. This verifies that the external visual knowledge actually bring benefits to the model performance. Interestingly, we also find that although by adding textual knowledge or visual knowledge can both be useful to the model performance, the gain is much more significant ($\sim$1% vs. $\sim$3%) when adding visual knowledge than textual knowledge. This indicates that visual knowledge is more important to achieve better performance. We also observe that comparing to purely utilize dialogue history as inputs, adding external knowledge inputs (e.g., textual knowledge or visual knowledge) can remarkably improve the performance of both baseline models and ours. This shows that the effective use of the external knowledge data is critical to the model performance in knowledge-intensive tasks such as ours.

5.4 Ablation Study

Table 3 has shown the ablated results. We have ablated each component in our model including both knowledge retrievers, multi-granularity fusion layer, RoBERTa encodings and UniLM finetuning. As we can see from the table, all the individual components have notably contributed to the full model performance. Specifically, when removing both knowledge retrievers, the performance has decreased significantly (an average about 2% absolute drop). This confirms the effectiveness of the knowledge retrievers especially the novel VisualKnowledgeRetriever component. After removing the multi-granularity fusion layer in the encoder, the performance has dropped remarkably which confirms that by incorporating multi-granularity semantics we can further improve the system performance. When replacing the pre-trained RoBERTa

	Dialogue History			*Dialogue History + Textual Knowledge*			*Dialogue History + Textual + Visual Know.*		
Model	**BLEU**	**Entity F1**	**PPL**	**BLEU**	**Entity F1**	**PPL**	**BLEU**	**Entity F1**	**PPL**
Mem2Seq	50.14	63.25	59.71	52.35	68.04	56.58	-	-	-
GLMP	62.02	72.13	48.13	64.21	73.90	45.11	-	-	-
DF-Net	63.63	72.55	47.85	64.44	74.13	44.79	-	-	-
GraphDialog	64.33	71.85	48.67	65.58	73.56	45.20	-	-	-
UniMF (K=1)	67.17	76.35	43.48	67.38	76.70	41.88	68.93	78.77	38.57
UniMF (K=3)	**67.17***	**76.35***	**43.48***	**68.29***	**77.49***	**41.10***	**69.98***	**79.83***	**37.26***
UniMF (K=6)	67.17	76.35	43.48	66.72	76.13	42.57	68.65	78.94	40.15

Table 2: Main results on the *MMDialKB* test set. The numbers with * indicates that the improvement of our model over all baselines is statistically significant with $p < 0.05$ under t-test.

Model	**BLEU**	**Entity F1**	**PPL**
UniMF (Full model)	**69.98**	**79.83**	**37.26**
- w/o Textual Know. Retriever	68.64	78.21	39.53
- w/o Visual Know. Retriever	67.81	77.23	41.06
- w/o Multi-Granularity Fusion	68.25	78.54	39.78
- w/o RoBERTa Encodings	66.67	76.43	41.59
- w/o UniLM Finetuning	66.49	76.03	42.14

Table 3: Ablation study on the *MMDialKB* test set.

Model	**Fluency**	**Correctness**	**Humanlike**
GLMP	4.08	3.81	3.22
DF-Net	4.16	3.96	3.30
GraphDialog	4.13	4.15	3.29
UniMF (Full model)	**4.33**	**4.37**	**3.68**
Human	4.85	4.63	4.59

Table 4: Human evaluation results on randomly selected responses from *MMDialKB* test set.

model in the encoder with standard GRU (See Appendix C for details), the performance has sharply decreased by 3%∼4% across various metrics. This shows that the general knowledge captured by the pre-training is beneficial to the model performance. This is also verified by replacing the UniLM with GRU as the decoder (Details in Appendix C). When removing UniLM, the model performance has significantly dropped by about 4% on average across all metrics.

5.5 Case Study

We provide one case study to conduct in-depth analysis of the model dynamics within our designed knowledge retrievers. Specifically, we utilize the last hop attentions in both knowledge retrievers for investigation. We find that our model has assigned higher weights for the textual entities `frankie_and_bennys` and `Italian` in the TextualKnowledgeRetriever during the first turn of the dialogue in Figure 1. For the follow-up turns, the retriever looks more on the visual entities which are the `dining hall` and `parking lot` images in the VisualKnowledgeRetriever. This shows that the knowledge retrievers successfully learn to retrieve the appropriate textual and visual knowledge from the external KBs according to the dialogue context. See visualization results in Figure 9 in the appendix.

5.6 Human Evaluations

Following prior work [Qin *et al.*, 2020], we conduct human evaluations on the generated responses from three aspects: *Fluency*, *Correctness*, and *Humanlikeness*. *Fluency* is utilized to measure the fluency of the outputs (e.g., contain repetitions or not). *Correctness* is used to measure the correctness in terms of dialogue flow, grammar and provided entities. *Humanlikeness* is utilized to evaluate the probability of the generated responses spoken by persons. We compare our work with previous state-of-the-art models and the original responses as well. We randomly select 300 different dialogue samples from the test set and ask human experts to judge the quality of the responses and score them according to the three metrics ranging from 1 to 5. The results are shown in Table 4. We can see that our model has outperformed all the baselines across all the three metrics, which is consistent with the observations in automatic evaluations. Details about the scoring criterions are listed in Appendix D.

6 Conclusion

In this paper, we propose *MMDialKB*, a new large-scale dialogue dataset focusing on multimodal KBs. Different from existing dialogue datasets with KBs such as *SMD* and *MultiWOZ 2.1*, *MMDialKB* features a unique multimodal KB and high-quality dialogue data collected upon it. To handle multimodal KB inputs, we propose a novel framework which integrates pre-trained language models with external multimodal KB reasoning. To further improve the performance, we introduce a novel multi-granularity fusion mechanism to capture multi-grained semantics in the dialogue history. Both automatic and human evaluations demonstrates the effectiveness of the proposed framework. We hope these efforts could facilitate new advances towards task-oriented dialogue modeling with multimodal KBs.

References

[Bordes *et al.*, 2016] Antoine Bordes, Y-Lan Boureau, and Jason Weston. Learning end-to-end goal-oriented dialog. *arXiv preprint arXiv:1605.07683*, 2016.

[Das *et al.*, 2017] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[De Vries *et al.*, 2017] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4466–4475, 2017.

[Dong *et al.*, 2019] Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[Eric and Manning, 2017] Mihail Eric and Christopher Manning. 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*, pages 468–473, 2017.

[Eric *et al.*, 2017] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, 2017.

[Huang *et al.*, 2020] Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. Semi-supervised dialogue policy learning via stochastic reward estimation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 660–670, 2020.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lei *et al.*, 2018] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 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*, pages 1437–1447, 2018.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[Madotto *et al.*, 2018] Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. Mem2Seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1468–1478, 2018.

[Moon *et al.*, 2020] Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*, 2020.

[Mostafazadeh *et al.*, 2017] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. Image-grounded conversations: Multimodal context for natural question and response generation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing*, pages 462–472, 2017.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

[Qin *et al.*, 2020] Libo Qin, Xiao Xu, Wanxiang Che, Yue Zhang, and Ting Liu. 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*, pages 6344–6354, 2020.

[Sukhbaatar *et al.*, 2015] Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. End-to-end memory networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, 2015.

[Wolf *et al.*, 2019] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, pages arXiv–1910, 2019.

[Wu *et al.*, 2016] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016.

[Wu *et al.*, 2019] Chien-Sheng Wu, Richard Socher, and Caiming Xiong. Global-to-local memory pointer networks for task-oriented dialogue. In *ICLR*, 2019.

[Yang *et al.*, 2020] Shiquan Yang, Rui Zhang, and Sarah Erfani. 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*, pages 1878–1888, 2020.

[Zhao *et al.*, 2017] Tiancheng Zhao, Allen Lu, Kyusong Lee, and Maxine Eskenazi. Generative encoder-decoder models for task-oriented spoken dialog systems with chatting capability. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 27–36, 2017.