Task-Oriented Conversation Generation Using Heterogeneous Memory Networks

Zehao Lin [1] **Xinjing Huang** [1] **Feng Ji** [2] **Haiqing Chen** [2] **Yin Zhang**[1]*
[1] College of Computer Science and Technology, Zhejiang University
[2] DAMO Academy, Alibaba Group
{georgelin,huangxinjing,zhangyin98}@zju.edu.cn
{zhongxiu.jf,haiqing.chenhq}@alibaba-inc.com

Abstract

How to incorporate external knowledge into a neural dialogue model is critically important for dialogue systems to behave like real humans. To handle this problem, memory networks are usually a great choice and a promising way. However, existing memory networks do not perform well when leveraging heterogeneous information from different sources. In this paper, we propose a novel and versatile external memory networks called Heterogeneous Memory Networks (HMNs), to simultaneously utilize user utterances, dialogue history and background knowledge tuples. In our method, historical sequential dialogues are encoded and stored into the context-aware memory enhanced by gating mechanism while grounding knowledge tuples are encoded and stored into the context-free memory. During decoding, the decoder augmented with HMNs recurrently selects each word in one response utterance from these two memories and a general vocabulary. Experimental results on multiple real-world datasets show that HMNs significantly outperform the state-of-the-art data-driven task-oriented dialogue models in most domains.

1 Introduction

Compared with chitchat, task-oriented dialogue systems aim at solving tasks in specific domains with grounding knowledge. Though far from handling conversation like a real human, existing task-oriented dialogue systems have shown cheerful prospect in a specific domain, e.g. Siri and Cortana are personal assistants, helping people a lot in daily life and business work.

In general, knowledge-grounded task-oriented dialogue system can be divided into three important components: understanding user utterances, fetching right knowledge from external storage and replying right answer. As shown in Figure 1, agent is required to do a point-of-interest navigation. According to dialogue history, agent will fetch related knowledge base information, in our case represented as tuples (e.g. [$hotel\_keen$, poi_type, $rest\_stop$], which indicates the point-of-i nterests type of $hotel\ keen$ is $rest\ stop$), as an external knowledge to answer correctly and complete task.

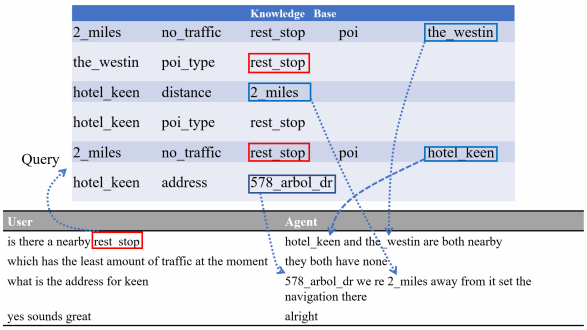


Figure 1: A multi-turn dialogue example. The upper table shows several n-tuples sampled from knowledge base. Lower table shows multi-turn dialogues. Agent needs to retrieve appropriate knowledge tuples to generate the proper response.

Traditional pipeline dialogue systems(Yan et al., 2017; Rojas-Barahona et al., 2017) and some end-to-end dialogue systems rely on the predefined the slot filling labels. Despite the consumption of human efforts, these kinds of systems are difficult to adapt to new domains.

Many works, e.g. (Vinyals and Le, 2015; Shang et al., 2015), show that training a fully data-driven end-to-end model is a promising way to build domain-agnostic dialogue system. Their models mostly try to use the attention mechanism, including memory networks techniques, to fetch the most similar knowledge (Sukhbaatar et al., 2015), then incorporate grounding knowledge into a seq2seq neural model to generate a suitable re-

*Corresponding Author

Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, pages 4558–4567, Hong Kong, China, November 3–7, 2019.

sponse (Madotto et al., 2018).

However, existing memory networks equally treat information from multiple sources, e.g. sequential dialogue history and structure knowledge bases. Therefore two weaknesses arise in such methods: (1) It is difficult to model different types of structured information in only one memory network. (2) It is also difficult to model the effectiveness of knowledge from different sources in such a single memory network. To address these issues, we expand the architecture of memory networks used in a seq2seq neural model.

Our contributions are mainly three-fold:

- We propose a novel seq2seq neural conversation model augmented with Heterogeneous Memory Networks. We first model sequential dialogue history and grounding external knowledge with two different kinds of memory networks and then feed the output of context-aware memory to the context-free memory to search the representations of similar knowledge.

- Our context-aware memory networks is able to learn the context-aware latent representations and stores them into memory slots, by employing a gating mechanism when encoding dialogue history and user utterance.

- Experimental results demonstrate that our neural approach significantly outperforms the examined neural methods automatic metrics, and context-aware memory networks can learn and store more meaningful representations than the examined memory approaches.

2 Related Works

The end-to-end model uses deep neural net instead of several parts in pipeline models to generate responses. (Rojas-Barahona et al., 2017) propose a data-driven goal-oriented neural dialogue system by adding database operator and policy networks modules to introduce database information and track state which need extra labeling step that breaks differentiability. (Bordes and Weston, 2016) propose a testbed to break down the strengths and shortcomings of end-to-end dialog systems in goal-oriented applications. Those methods treated dialogue system as the problem of learning a mapping policy from dialogue histories to agents' responses.

The booming internet dialogue data lay the foundation of building data-driven models. (Ritter et al., 2011) first applied phrase-based Statistical Machine Translation(Setiawan et al., 2005). It treats the conversation system as a translation problem, a user utterance needs to be translated into an agent response.

(Sutskever et al., 2014) propose Sequence to sequence model (SEQ2SEQ) architecture and apply it to neural machine translation task. SEQ2SEQ has become a general basis of natural language generation tasks, e.g. question answering(Tan et al., 2018) and question generation (Zhou et al., 2017). By applying the RNN based encoder-decoder framework to generate responses, models(Shang et al., 2015; Cho et al., 2014b; Luong et al., 2015b) are able to utilize neural networks to learn the representation of dialogue histories and generate appropriate responses.

To deal with multi-turn information, (Sordoni et al., 2015) propose a model that represents the whole dialogue history (including the current message) with continuous representations or embeddings of words and phrases to address the challenge of the context-sensitive response generation.

By adding a knowledge base module, recent works (Ghazvininejad et al., 2018; Young et al., 2018) have shown the possibility of training an end-to-end task-oriented dialogue system on the sequence to sequence architecture. Ghazvininejad et al. (Ghazvininejad et al., 2018) generalize the SEQ2SEQ approach by conditioning responses on both conversation history and external knowledge, aiming at producing more contextual responses without slot filling.

CopyNet (Gu et al., 2016) and Pointer Networks (Vinyals et al., 2015) improve model's accuracy and ability of handling of out-of-vocabulary words using neural attention. Pointer-Generator networks (See et al., 2017) apply copy mechanism to the neural generation model. Their work shows copy mechanism can improve quality in text generation. (Dhingra et al., 2017) and (Li et al., 2017) apply reinforcement learning to make it differentiable.

Recent works on external memory (Graves et al., 2014; Henaff et al., 2016) provide an efficient method of introducing and reasoning different types of external information. (Sukhbaatar

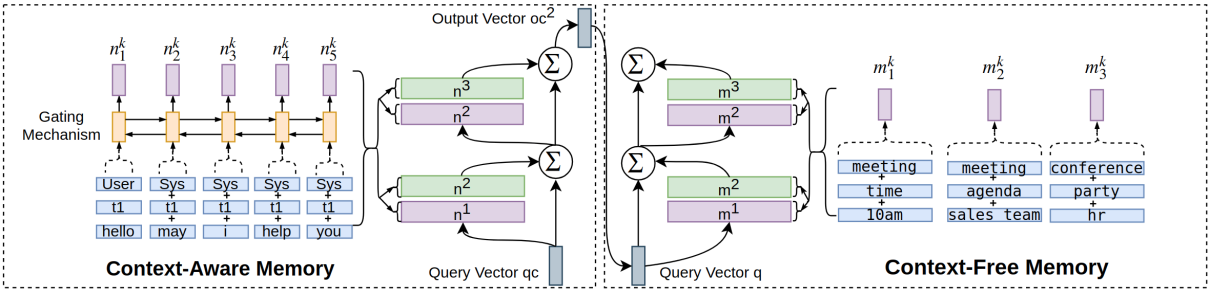


Figure 2: An example of Heterogeneous Memory Networks with two-hop attention. Context-aware memory which encodes dialog history to a context vector oc^2 while context-free memory loads knowledge base information. The output of context-aware memory will be employed as the query vector to the context-free memory.

et al., 2015) propose end-to-end memory networks with multiple attention hops model over a possibly large external memory. (Madotto et al., 2018) propose Mem2Seq that combines the end-to-end memory networks with the idea of pointer networks. (Chen et al., 2018) add the hierarchical structure and the variational memory network to capture both the high-level abstract variations and long-term memories during the dialogue tracking. To take care of information from different sources, (Fu and Feng, 2018) propose an attention mechanism to encourage the decoder to actively interact with the memory by taking its heterogeneity into account.

3 Proposed Framework

To generate responses using dialogue history and grounding knowledge, we introduce a novel encoder-decoder neural conversation model augmented with Heterogeneous Memory Networks (HMNs). The encoder module adopts a context-aware memory network to better understand the dialogue history and query. The decoder is enhanced with HMNs, which is able to incorporate external knowledge and dialog history when generating words.

3.1 Encoder

The encoder encodes the dialogue history into a fixed context vector. Here we adopt the context-aware memory as our encoder module. As shown in the left part of Figure 2, each word will be extended to the following parts: 1) token itself. 2) A turn tag. 3) An identity tag. For example, in the first turn a user says "hello" and the response from system is "may I help you", it will be concatenated as [(hello, t1, user),(may, t1, sys),(I, t1, sys),(help, t1, sys), (you, t1, sys)], *sys* means the word comes

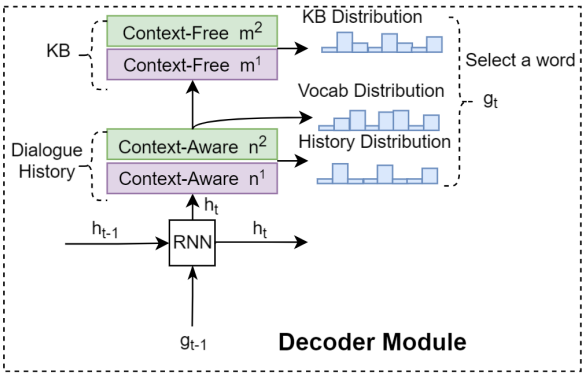


Figure 3: Decoder module with two attention hops

from the dialog system, so does *user*. *t1* indicates the word is from the first turn. Each word can be transformed into vector by embedding lookup, and we sum up vectors in each tuple to be the input sequence of the context-aware memory. Using a fixed vector to query the memory, a context vector c can be obtained.

3.1.1 Context-Aware Memory

To efficiently model the context information of sequential data, we present the context-aware memory. Memory slots n^k are structured by concatenating all input vectors as $n^k = cat[n_1^k, n_2^k, ..., n_l^k]$, k stands for the k-th hop in memory, and l means the length of the input. n_l^k is the sum of each word tag embedding using each hop own randomly initialized embedding matrix C^k. And we adopt the adjacent weight sharing scheme, which means C^k is not only the input embedding matrix in the k-th hop, but also the output embedding matrix in the (k-1)-th hop. We add a gating mechanism between memory cells. The gating mechanism applied is adopted from Bidirectional GRU (Cho et al., 2014a) in our case. Thus the context-dependent representation of in-

puts, denoted as $n^k = [\overrightarrow{n^k}, \overleftarrow{n^k}]$, where $\overrightarrow{n^k}$ and $\overleftarrow{n^k}$ are the forward and backward representation of inputs, respectively. The forward process can be illustrated as equations:

$$r_t = \sigma(W_1 v_t^k + W_2 \overrightarrow{n^k}_{(t-1)} + b_1) \quad (1)$$

$$z_t = \sigma(W_3 v_t^k + W_4 \overrightarrow{n^k}_{(t-1)} + b_2) \quad (2)$$

$$e_t = \tanh(W_5 v_t^k + r_t(W_6 \overrightarrow{n^k}_{(t-1)} + b_3)) \quad (3)$$

$$\overrightarrow{n_t^k} = (1 - z_t)e_t + z_t \overrightarrow{n^k}_{(t-1)} \quad (4)$$

where W_1, W_2, W_3, W_4, W_5, W_6 and b_1, b_2, b_3 are trainable weights and biases.

Given the query vector qc^k, attention weights over memory cells n^k can be calculated by the equation:

$$p^k = Softmax(n^k \cdot qc^k) \quad (5)$$

the readout vector is the sum of output memory matrix n^{k+1} with corresponding attention weights p^k.

$$u^k = \sum_i (p_i^k \cdot n_i^{k+1}) \quad (6)$$

By summing query and readout vector together, we can get the output from the k-th hop.

$$oc^k = qc^k + u^k \quad (7)$$

Note oc^k is also the query vector of the (k+1)-th hop.

3.2 Decoder

The decoder contains HMNs and an RNN controller, as shown in Figure 3. The controller controls the process of querying HMNs.

In each time step, HMNs will generate: 1) a readout vector oc^1, which is the output of the first hop in history memory, and 2) attention weights of the last hop in two memories, called history word distribution P_{his} and knowledge tuple distribution P_{kb}. The readout vector is concatenated with h_t to predict the vocabulary distribution P_{vocab}. Formally,

$$P_{vocab} = Softmax(W_7[h_t, oc^1]) \quad (8)$$

where W_7 is trainable weight matrix.
We adopt a simple strategy (Section 3.5) to select a word from three distributions P_{vocab}, P_{his} and P_{kb}.

3.2.1 Heterogeneous Memory Networks

HMNs stacks two types of memory: 1) context-aware memory and 2) context-free memory. Dialog history is loaded into context-aware memory, and knowledge base triples are loaded into context-free memory. Firstly, HMNs accepts query vector as inputs, then walk through context-aware memories. The final output u^k in the last hop will be employed to query context-free memory. Context-aware memory has been detailed in Section 3.1.1.

Context-Free memory itself is end-to-end memory networks (Sukhbaatar et al., 2015). Compared with our context-aware memory, it has no gating mechanism. The input to the memory is the summed vectors in each knowledge triple, as depicted in the right part of Figure 2. Each hop owns randomly initialized embedding matrix $C^{k'}$ We denote memory slots as m^k. It accepts a query vector and then follows the same process 5 to 7, the output $u^{k'}$ can be obtained.

3.2.2 Controller

We adopt GRU as our controller. It accepts the output c from encoder as initial hidden state h_0. In each time step, it takes the previous generated word g_{t-1}'s embedding $E(g_{t-1})$ and last time hidden state h_{t-1} as inputs. Formally:

$$h_t = GRU(E(g_{t-1}), h_{t-1}) \quad (9)$$

then h_t is used to query the HMNs.

3.3 Copy Mechanism

We adopt copy mechanism to copy words from memories. Attention weights in the last hop of the two memories,P_{kb} and P_{his} will be the probability of the target word from those memories. If the target word does not appear in inputs, the position index will be the last position in memories, which is a sentinel added in preprocessing stage.

3.4 Joint Learning

To learn the distribution of three vocabularies P_{vocab}, P_{kb} and P_{his} in each time step, the loss in the t-th time step is the negative log-likelihood of the predict probability of the target word for that time step. Formally:

$$Loss = -\frac{1}{T} \sum_{t=0}^{t=T} \sum_i (\log p_{ti}) \quad (10)$$

Note that p_{it} means the t-th word's probability in $i \in \{P_{vocab}, P_{kb}, P_{his}\}$.

Table 1: The statistics of the bAbI-3, 4, 5, DSTC2 and Key-Value Retrieval datasets

Datasets	Key-Value Retrieval dataset	DSTC 2	bAbI-3	bAbI-4	bAbI-5
Avg. History words	25.5	63.4	49.9	20.5	62.6
Avg. KB pairs	64.7	42.7	23.4	7.0	23.6
Avg. Response Length	8.7	10.2	7.2	5.7	6.5
Vocabulary Size	1554	1066	739	1004	1135
Dialogue Turns	2.8	9.9	10.9	4.5	19.3

3.5 Word Selection Strategy

In our case, if words with the highest probability in P_{his} and P_{kb} vocabularies are not on sentinel positions, we directly compare the probability of each word and select the higher one. If one of the vocabularies points to the sentinel position, the model will select the word with the highest probability in the other vocabulary. At last, if both vocabularies get to sentinel positions, the word from P_{vocab} will be selected.

4 Experimental Setup

4.1 Datasets

As the proposed approach is quite general, the model can be applied to any task-oriented dialogue datasets with conversation and knowledge base data. To evaluate and compare the results with the state-of-the-art methods in multiple dimensions, we choose three popular task conversation datasets including DSTC 2, Key-Value Retrieval dataset and the (6) dialog bAbI tasks. Table 1 shows the statistics of datasets.

- *Key-Value Retrieval dataset* (Eric and Manning, 2017). This dataset releases a corpus of 3,031 multi-turn dialogues. The dialogues consist of three different domains: calendar scheduling, weather information retrieval, and point-of-interest navigation.

- *The (6) dialog bAbI tasks* (Bordes and Weston, 2016). The (6) dialog bAbI tasks are a set of five subtasks within the goal-oriented context of restaurant reservations. Conversations are grounded with an underlying knowledge base of restaurants and their properties (location, type of cuisine, etc.). As task 1 and 2 have been achieved very well, we only test our model on task 3 to 5 and their OOV(out-of-vocabulary), where entities (e.g. restaurant names) in test sets may not have been able to see during training.

- *The Dialog State Tracking Challenge 2* (DSTC 2). DSTC 2 is a research challenge focused on improving the state-of-the-art in tracking the state of spoken dialogue systems. DSTC 2's training dialogues were gathered using Amazon Mechanical Turk related to restaurant search.

For all datasets, we employ the original conversation and knowledge base information only and drop the other labels e.g. slot filling labels. We take several metrics over all datasets to evaluate the performance on multiple dimensions. And to evaluate the context-aware memory networks, we also test the HMNs with only context-free memory on the dialog bAbI tasks.

4.2 Evaluation Method

To compare with the original datasets baselines, we apply evaluation methods on each datasets the same as datasets' original papers described in 4.1.

- Bilingual Evaluation Understudy (BLEU) (Papineni et al., 2002). BLEU has been widely employed in evaluating sequence generation including machine translation, text summarization, and dialogue systems. BLEU calculates the n-gram precision which is the fraction of n-grams in the candidate text which is present in any of the reference texts.

- F1 Score (F-measure): F1 evaluates the model's ability in terms of precision and recall, which is more comprehensive than just using precision or recall measure. We adopt F1 to evaluate if a model can extract information from a knowledge base precisely.

- Per-response accuracy and Per-dialog accuracy. Per-response and Per-dialog accuracy count the percentage of responses that are correct. Any incorrect words will make a response or a dialogue negative. Accuracy shows if the model is able to learn the distribution of reproducing factual details.

Table 2: Results on Key-Value Retrieval dataset. F1 score, including Entity F1, is micro-average over the entire set, and three subtasks. Human results are reported by Eric et al.(Eric and Manning, 2017)

Model name	BLEU	Ent. F1	Scheduling Ent. F1	Weather Ent. F1	Navigation Ent. F1
Human*	13.5	60.7	64.3	61.6	55.2
SEQ2SEQ	11.07	30.5	30.7	**46.4**	13.4
SEQ2SEQ+Attn.	11.19	35.6	40.5	44.0	23.0
Mem2Seq	12.06	31.1	51.8	34.3	12.3
HMNs	**14.46**	**43.1**	**61.3**	40.3	**32.3**

Table 3: Results of Per-response accuracy and Per-dialog accuracy (in brackets) on bAbI dialogues. Per-dialog accuracy presents the accuracy of complete dialogues.

Task	SEQ2SEQ	SEQ2SEQ+Attn.	Mem2Seq	HMNs-CFO	HMNs
T3	74.8(0)	74.8(0)	83.9(15.6)	93.7(55.9)	**93.6(56.1)**
T4	56.5(0)	56.5(0)	97.0(90.5)	96.8(89.3)	**100(100)**
T5	98.9(82.9)	**98.6(83)**	96.2(46.4)	97.1(58.2)	98.0(69.0)
T3-OOV	74.9(0)	74.0(0)	83.6(18.1)	92.3(45.2)	**92.5(48.2)**
T4-OOV	56.5(0)	57.0(0)	97.0(89.4)	96.1(90.3)	**100(100)**
T5-OOV	67.2(0)	67.6(0)	71.4(0)	78.3(0)	**84.1(2.6)**

Table 4: The results on the DSTC 2

Model name	F1	BLEU
SEQ2SEQ	69.7	55.0
SEQ2SEQ+Attn.	67.1	**56.6**
SEQ2SEQ+Copy	71.6	55.4
Mem2Seq	75.3	55.3
Our model	**77.7**	56.4

4.3 Baselines and Training Setup

The hyper-parameter settings are adopted as the best practice settings for each training set following the Madotto's (Madotto et al., 2018) and Manning's (Manning and Eric, 2017) best experimental results on baselines SEQ2SEQ and Mem2Seq. Detailed models and their settings are as follows:

- Sequence to sequence. For SEQ2SEQ, we adopt one layer LSTMs as encoder and decoder. For Key-Value Retrieval dataset, hidden size is placed at 512 and the dropout rate is 0.3. On dataset bAbI, the hidden size and dropout rate are 128 and 0.1 for task 3, 256 and 0.1 for task 4 and 5. Learning rates are set to 0.001 for bAbI and 0.0001 for DSTC 2 and Key-Value Retrieval dataset.

- SEQ2SEQ + Attention. We adopt the attention mechanism (Luong et al., 2015a) commonly used in neural machine translation. On dataset bAbI, hidden size and the dropout rate are 256 and 0.1 for task 3 and 4, 128 and 0.1 for task 5. For Key-Value Retrieval dataset, hidden size and dropout rate are 512 and 0.3. On the DSTC 2 task, hidden size is set to 353 and word embedding size is 300 (same with original work).

- Mem2Seq. Except 128 in task 3, hidden size in other tasks is 256. The dropout rate is set to 0.2 in task 3, 4 and Key-Value Retrieval dataset, 0.1 in task 5 and DSTC 2 dataset. We adopt three hops in DSTC 2 and Key-Value Retrieval dataset.

- HMNs with context-free only (HMNs-CFO). To test the performance of context-aware memory, we apply other context-free to encode dialogue history instead of context-aware memory in HMNs. All the other structure and parameter settings are the same as HMNs in this model.

All models are tested with various hyper-parameter settings to get their best performance, e.g. hidden size selected from [64, 128, 256, 512]. Note that settings from datasets are also tested like SEQ2SEQ + Attention's hidden size is 353 on Key-Value Retrieval dataset.

During the training, all experiments employ the teacher-forcing scheme, feeding the gold target of last time or highest probability word into decoder with probability 50%. We also randomly mask input with UNK according to the dropout rate.

Table 5: A generated example from Key-Value Retrieval dataset with correct knowledge entities in **bold**. Given the knowledge base and user's request, we list the generated responses of three models and the gold answers. This example is randomly selected from all generated sentences and we only show tuples been used by models.

Dataset	
Knowledge Base	... valero poi_type gas_station valero distance 2_miles valero address 200_alester_ave 1_miles moderate_traffic parking_garage poi palo_alto_garage_r ...
User	address to the **gas_station**
Gold	**valero** is located at **200_alester_ave**
Generated Sentence	
SEQ2SEQ+Attn.	the closest **gas_station** is located at **200_alester_ave** 7_miles away would you like directions there
Mem2Seq	there is a **valero** 1_miles away
HMNs	there is a **gas_station** located **2_miles** away at **200_alester_ave**

4.4 HMNs Training Settings

We test the hidden size in [64, 128, 256] and set dropout rate in [0.1, 0.2]. Learning rate is initiated with 0.001 and training batch is set to 64. The metrics results are coming from the best result settings for each dataset. We select hidden size and dropout rate at (256, 0.1) on bAbI task 3 and task 5, (256,0.2) on task 4. On the DSTC 2 task, we set hidden size and dropout rate at (128, 0.1). For Key-Value Retrieval dataset, the setting is hidden size 256 and dropout rate 0.1. Except for bAbI tasks' 1 layer, all HMNs and Mem2Seq tasks employ 3 layer memories.

5 Results and Discussion

5.1 Results and Analysis

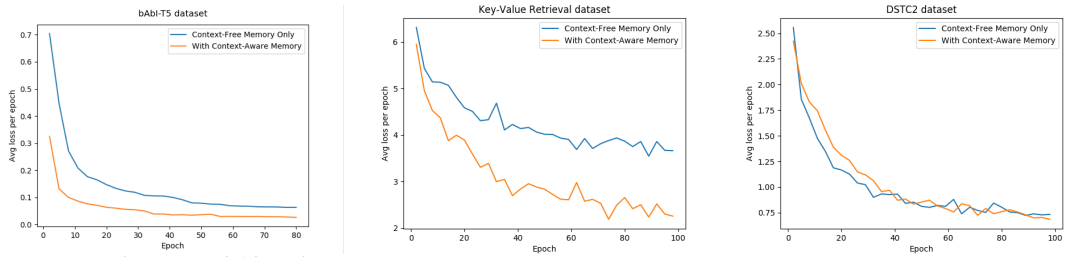
The best results of the baselines and HMNs are gathered into tables and figures. Table 2 show the result of models on Key-Value Retrieval dataset. Except for F1 scores on Calendar Scheduling, HMNs get significantly better results on all benchmarks comparing the state-of-the-art models. HMMs' BLEU score is even higher than human results which are reported in (Eric and Manning, 2017). Results show our model's outstanding performance in generating a fluent and accurate response in most tasks.

Examples generated by our approach and baselines are given in Table 5. These two examples are randomly selected from all generated sentences. Comparing the generated sentences by humans, although entities and sentences are different with gold answer in example one, our approach is able to produce more fluent and accurate sentences. However, the result on task weather forecasting neither HMNs and Mem2Seq can outperform SEQ2SEQ. We will discuss it in the next section.

Table 4 shows our model gets the best F1 score on dataset DSTC 2, while SEQ2SEQ with attention gets the best BLEU result.

Table 3 shows results of models on bAbI tasks. HMNs and Mem2Seq adopt one hop attention only and note that all results are the best performance of each model in 100 epochs. HMNs achieved the best results on most tasks except T5. HMNs-CFO also outperforms the other models. This demonstrates that both training multiple distributions over heterogeneous information and employment of context-aware memory benefit the end-to-end dialogue system. The improvements in per-dialogue accuracy on out-of-vocabulary tests are even more significant. Figure 4 shows the changes of HMNs and HMNs-CFO's total loss across time. HMNs learns significantly faster.

Though automatic metrics cannot really examine human beings' diversified expression, existing dialogue systems aim at generating sentence by learning the patterns of training data, so we believe BLEU is still a metric of great concern in comparing the similar models' ability in learning the sentence patterns. Though human results show end-to-end machines have still a long way to go (60.7 to 43.1). Compared to other models, HMNs sig-



(a) Loss changes on bAbI T5

(b) Loss by epochs on Key-Value Retrieval dataset

(c) Loss by epochs on the DSTC2 dataset

Figure 4: The total loss by epochs on three different datasets in 100 epochs. One line with context-aware memory and the other with only context-free memory

nificantly improves performance in retrieving correct knowledge entities.

5.2 Discussion

5.2.1 Context-Aware Memory

To show whether context-aware memory benefits conversation learning, on bAbI tasks, we also tested HMNs-CFO memory only. From Table 3, we observe that HMNs-CFO is significantly better than original Mem2Seq as well as SEQ2SEQ + attention in several results and only loses slightly on task 4 (89.3 to 90.5). One reason is that one memory is difficult to learn best distribution over different sources. Respectively encoding sequential dialogue history and grounding knowledge can learn two better distributions than one general but not best distribution. This also indicates that using the query vector generated by history memory to retrieve information in knowledge base memory is reasonable.

As the HMNs model get the best results in all tasks except one, in addition the results of training speed of HMNs and HMNs-CFO (Figure 4), the context-aware memory is clearly to learn representation of the dialogue history much better and faster and also demonstrates that the importance of incorporating context information for dialogue systems. HMNs outperform the HMNs-CFO not only on BLEUs but also entity F1 on most tasks, showing building a good representation of dialogue history benefits knowledge reasoning, and help to improve the context-free memory by issuing a good query vector.

From above all, we can conclude that both stacked memory networks architecture and using context-aware memory to load sequential information can improve the performance of retrieving knowledge and generating sentences.

5.2.2 Shortcomings

From the results in Table 2, we note that HMNs and Mem2Seq failed on weather forecasting task. We analysed the average knowledge pairs of weather forecasting tasks and find it near three times the knowledge pairs of the other two tasks. Then we carried out another experiment that first narrows the KB candidates by performing a matching preprocessing operation, and the Weather Ent. F1 result of our method will climb to more than 48 which is the best. This may indicates that this kind of memory networks may have difficulties in handling large scale knowledge base. So perform a matching operation to narrow the candidate knowledge space is critical in a real-world large scale knowledge base.

And in this paper, we only show sequential data and knowledge triples data. For more types of information to integrate, model needs to add other memory networks, e.g. graph neural networks augmented memory networks (Zhou et al., 2018) for graph structured data.

6 Conclusion

In this paper, we propose a model that is able to incorporate heterogeneous information in an end-to-end dialogue system. The model applies Heterogeneous Memory Networks (HMNs) to model sequential history and structured database. Results on several datasets show model can significantly improve the performance of generating the response. Our proposed context-aware memory networks show outstanding performance in learning the distribution over dialogue history and retrieving knowledge. We present the possibility of efficiently using various structured data in end-to-end task-oriented dialogue without any extra labeling and module training.

Acknowledgement

We thank the anonymous reviewers for their insightful comments on this paper. This work was supported by the NSFC (No.61402403), DAMO Academy (Alibaba Group), Alibaba-Zhejiang University Joint Institute of Frontier Technologies, Chinese Knowledge Center for Engineering Sciences and Technology, and the Fundamental Research Funds for the Central Universities.

References

Antoine Bordes and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. *CoRR*, abs/1605.07683.

Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1653–1662.

Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014a. On the properties of neural machine translation: Encoder-decoder approaches. In *Proceedings of SSST@EMNLP 2014, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation, Doha, Qatar, 25 October 2014*, pages 103–111.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014b. 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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734.

Bhuwan Dhingra, Lihong Li, Xiujun Li, Jianfeng Gao, Yun-Nung Chen, Faisal Ahmed, and Li Deng. 2017. Towards end-to-end reinforcement learning of dialogue agents for information access. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 484–495, Vancouver, Canada. Association for Computational Linguistics.

Mihail Eric and Christopher D Manning. 2017. Key-value retrieval networks for task-oriented dialogue. *arXiv preprint arXiv:1705.05414*.

Yao Fu and Yansong Feng. 2018. Natural answer generation with heterogeneous memory. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 185–195.

Marjan Ghazvininejad, Chris Brockett, Ming-Wei Chang, Bill Dolan, Jianfeng Gao, Wen-tau Yih, and Michel Galley. 2018. A knowledge-grounded neural conversation model. pages 5110–5117.

Alex Graves, Greg Wayne, and Ivo Danihelka. 2014. Neural turing machines. *arXiv preprint arXiv:1410.5401*.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.

Mikael Henaff, Jason Weston, Arthur Szlam, Antoine Bordes, and Yann LeCun. 2016. Tracking the world state with recurrent entity networks. *CoRR*, abs/1612.03969.

Xuijun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-end task-completion neural dialogue systems. In *Proceedings of The 8th International Joint Conference on Natural Language Processing*.

Thang Luong, Hieu Pham, and Christopher D. Manning. 2015a. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1412–1421.

Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015b. Addressing the rare word problem in neural machine translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 11–19.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 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, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1468–1478.

Christopher D. Manning and Mihail Eric. 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, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*, pages 468–473.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, EMNLP 2011, 27-31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 583–593.

Lina Maria Rojas-Barahona, Milica Gasic, Nikola Mrksic, Pei-Hao Su, Stefan Ultes, Tsung-Hsien Wen, Steve J. Young, and David Vandyke. 2017. A network-based end-to-end trainable task-oriented dialogue system. pages 438–449.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1073–1083.

Hendra Setiawan, Haizhou Li, Min Zhang, and Beng Chin Ooi. 2005. Phrase-based statistical machine translation: A level of detail approach. In *Natural Language Processing - IJCNLP 2005, Second International Joint Conference, Jeju Island, Korea, October 11-13, 2005, Proceedings*, pages 576–587.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, pages 1577–1586.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Denver, Colorado, USA, May 31 - June 5, 2015*, pages 196–205.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 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*, pages 2440–2448.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Chuanqi Tan, Furu Wei, Nan Yang, Bowen Du, Weifeng Lv, and Ming Zhou. 2018. S-net: From answer extraction to answer synthesis for machine reading comprehension. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 2692–2700.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. *CoRR*, abs/1506.05869.

Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building task-oriented dialogue systems for online shopping. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA.*, pages 4618–4626.

Tom Young, Erik Cambria, Iti Chaturvedi, Hao Zhou, Subham Biswas, and Minlie Huang. 2018. Augmenting end-to-end dialogue systems with commonsense knowledge. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Hao Zhou, Tom Young, Minlie Huang, Haizhou Zhao, Jingfang Xu, and Xiaoyan Zhu. 2018. Commonsense knowledge aware conversation generation with graph attention. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden.*, pages 4623–4629.

Qingyu Zhou, Nan Yang, Furu Wei, Chuanqi Tan, Hangbo Bao, and Ming Zhou. 2017. Neural question generation from text: A preliminary study. In *Natural Language Processing and Chinese Computing - 6th CCF International Conference, NLPCC 2017, Dalian, China, November 8-12, 2017, Proceedings*, pages 662–671.