Memory Grounded Conversational Reasoning

Seungwhan Moon, Pararth Shah, Anuj Kumar, Rajen Subba
Facebook Assistant
{shanemoon, pararths, anujk, rasubba@}fb.com

Abstract

We demonstrate a conversational system which engages the user through a multi-modal, multi-turn dialog over the user's memories. The system can perform QA over memories by responding to user queries to recall specific attributes and associated media (*e.g.* photos) of past episodic memories. The system can also make proactive suggestions to surface related events or facts from past memories to make conversations more engaging and natural. To implement such a system, we collect a new corpus of memory grounded conversations, which comprises human-to-human role-playing dialogs given synthetic memory graphs with simulated attributes. Our proof-of-concept system operates on these synthetic memory graphs, however it can be trained and applied to real-world user memory data (*e.g.* photo albums, etc.) We present the architecture of the proposed conversational system, and example queries that the system supports.

1 Introduction

In the last few decades, people have been storing an increasing amount of their life's memories in the form of digital multimedia, *e.g.* photos, videos and textual posts. Retrieving one's memories from these memory banks and reminiscing about events from one's personal and professional life is a prevalent desire among many users. Traditionally, the interfaces to access these memories are either (i) keyword based search systems which demand specific keyword combinations to identify and retrieve the correct memories, or (ii) catalog based browsing systems that allow scrolling through memories across a single dimension, most commonly, time of creation. However, we posit that a more natural way of interacting with one's memories is through a flexible interface that can support fuzzy queries by referencing memories through various attributes, such as events, people, locations or activities associated with them, and that can enable the user to explore other relevant memories connected through one of many dimensions, *e.g.* same group of people, same location, etc., thereby not being restricted to browsing only temporally adjacent memories.

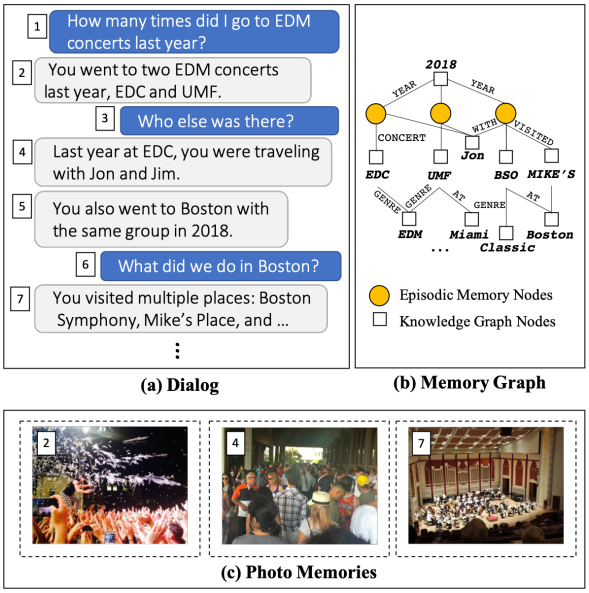

Figure 1: **Memory Grounded Conversational Reasoning** between a user and the assistant with a parallel (a) dialog and (b) memory graph pair. Dialog transitions can be captured as walks over a memory graph. (c) Some of the memories can be inferred from photos or other media, which are surfaced to the user when they are available and relevant.

We present a conversational system that provides a natural interface for retrieving and browsing through one's memories. The system supports conversational QA capability for open-ended querying of memories, and also supports the ability to proactively surface related memories that the user would naturally be interested in consuming. An important element of such an open-ended dialog system is its ability to ground conversations with past memories of users, making the interactions more personal and engaging.

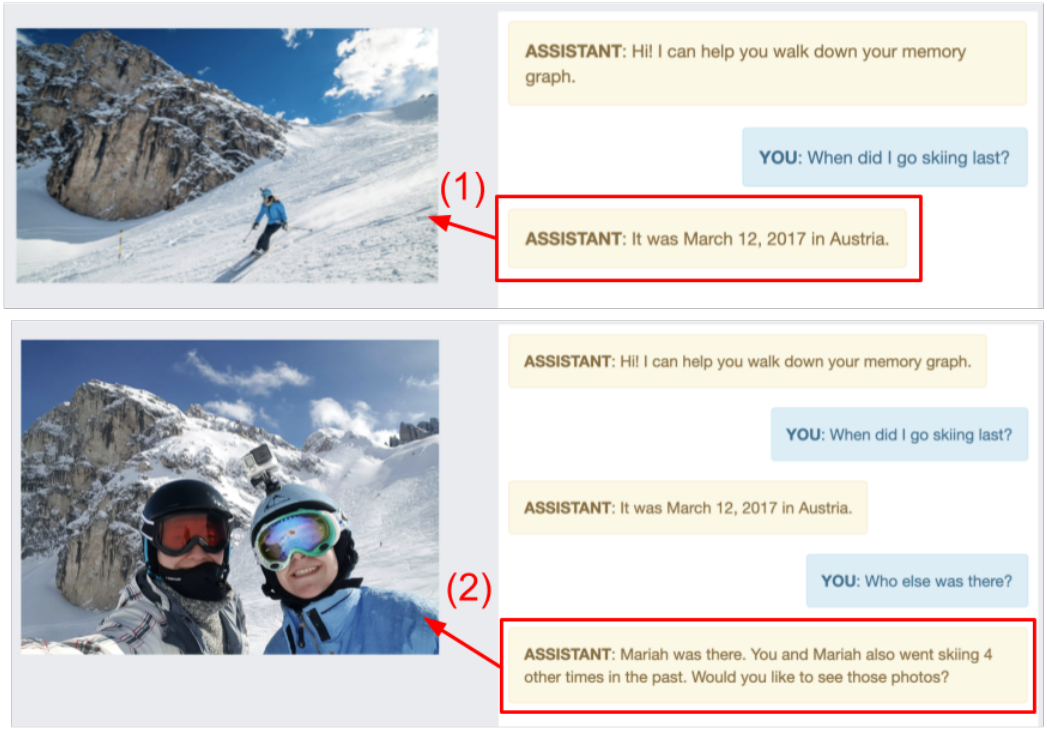

Figure 2: **Memory Walker Chatbot UI** for memory grounded conversations between a user and the assistant.

Figure 1 shows an example interaction supported by the demonstrated system, spanning multiple episodic memories that are represented as a graph composed of memory nodes and related entity nodes connected via relational edges. The example above shows three key novel features of the demonstrated system: 1) the ability of querying a personal database to answer various complex user queries (memory recall QA), 2) surfacing photos most relevant to the dialog, and 3) identifying other memories to surface that are relevant to conversational contexts, resulting in increased engagement and coherent interactions.

Our system consists of several components: First, we use the Memory Graph Networks (MGN) model, which learns natural graph paths among episodic memory nodes, conditioned over dialog contexts. MGN can introduce new memory nodes relevant to conversational contexts when memory nodes are activated as a result of graph walks. Second, the QA module takes as input user query utterances and infers correct answers given candidate memory graph nodes activated with the MGN model. Specifically, we utilize multiple sub-module nets such as `CHOOSE`, `COUNT`, etc., to support discrete reasoning questions that cannot be handled directly via graph networks. Finally, the photo recommender module then uses the MGN graph node embeddings and the generated attention scores to retrieve the most relevant photos for each dialog response.

Section 2 provides a detailed description of the system's user interface, method and data collection setup. Section 3 provides an analysis of demonstrations performed via the system, and Section 4 lists related work.

2 Memory Grounded Conversations

2.1 User Interface

The goal of the demonstrated system is to establish a natural user interface (UI) for interacting with memories. Figure 2 shows the UI of the demonstrated system, composed of two main sections: the media section (left) and the chat section (right; highlighted yellow: assistant, blue: user). This is a simple yet powerful interface for interacting with memories, for the following reasons:

Flexible. The UI enables natural language QA queries through text or voice input. This UI can be deployed in a desktop, web or mobile application, or on a connected home device like smart TVs. For each user memory recall query, the system provides a corresponding answer from the memory graph. In Figure 2 part 1, the user retrieves a memory related to an activity (skiing) through a textual question.

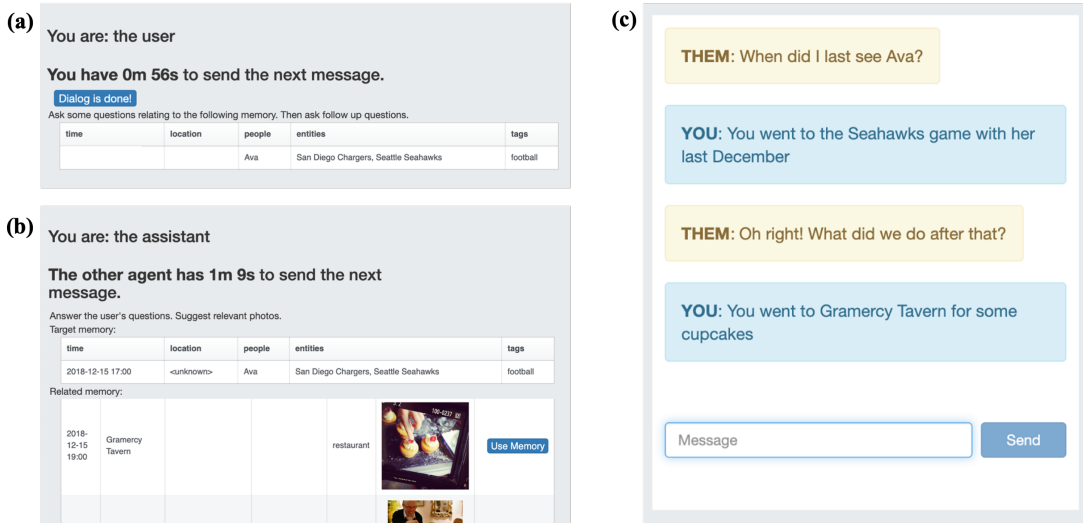

Figure 3: ***Memory Dialog* Dataset Collection Interface**, with an example. (**a**) User-playing agent is provided with partial memory information to query about. (**b**) Assistant-playing agent has the ground-truth information about the target memory as well as all related memories. (**c**) The two agents generate dialogs, grounded on the synthetic memories that they are presented with.

Visual. The UI can display visual content connected to the memories, and allow for further queries into the content of the image or video. Authors' personal photos are used in the demonstration, attached with synthetically generated memory graphs.

Contextual. The system keeps track of the conversational context within a user session, allowing the user to refer to entities present in the dialog or media. In Figure 2 part 2, the user refers to the event mentioned in the previous system response and asks a further question regarding who attended the event.

Proactive. The system can insert conversational recommendations for exploring related memories based on the system's model of which memories are naturally interesting for users to consume in a particular context. In Figure 2 part 2, the system suggests to the user other memory instances that share the same activity and set of people. The system can make the suggestions more personalized by learning the sequences in which users like to explore memories, from the user's past sessions.

2.2 Dataset: *Memory Dialog*

Synthetic Memory Graph: For the proposed system, we first bootstrap the large-scale memory collection through a synthetic memory graph generator, which creates multiple artificial episodic memory graph nodes with connections to real entities appearing on common-fact KGs (*e.g.* locations, events, public entities). We first build a synthetic social graph of users, where each user is assigned to a random interest profile as a probability distribution over the 'activity' space. We then iteratively generate a memory node and its associated attributes and entities by sampling activities, participants, locations, entities, time, *etc.* each from a manually defined ontology. Through the realistic memory graph that is synthetically generated, we avoid the need for extracting memory graphs from other structured sources (*e.g.* photo albums) which are often private or limited in size. Note also that by representing the memory database in a graph format, it allows for flexible operations required for complex QA and conversational reasoning.

Wizard-of-Oz Setup: We collect the *Memory Dialog* dataset in a Wizard-of-Oz setting (Shah et al., 2018) by connecting two crowd-workers to engage in a role-playing chat session either as a user or an assistant, with the joint goal of creating natural and engaging dialogs (Figure 3). The user-playing agent is given a memory node from the synthetic memory graph with some of the attributes hidden, and asked to initiate a conversation about those missing attributes to simulate a memory recall query. The assistant-playing agent is provided with a set of memories and photos relevant to user's questions, and is instructed to use one or more of these memories to answer the question and frame a free-form conversational response. In addition, the assistant agent is encouraged to proactively bring up any other memo-

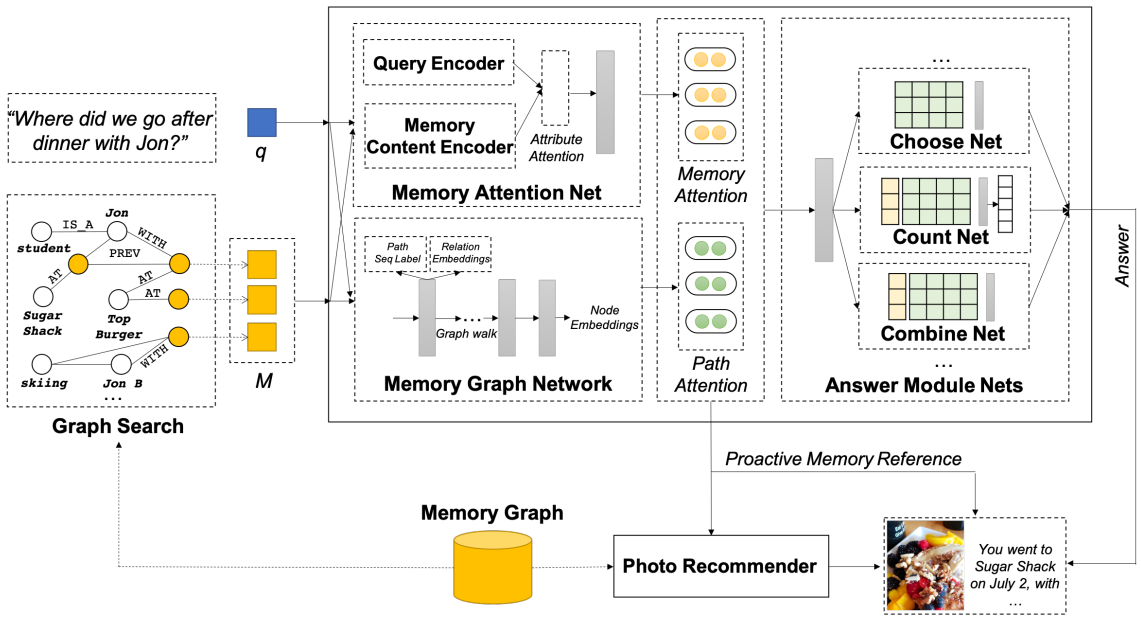

Figure 4: **Overall architecture** of the demonstrated system. Candidate memory nodes $\mathbf{m} = \{\mathbf{m}^{(k)}\}$ are provided as input memory slots for each query $\mathbf{q}$. The Memory Graph Network then traverses the memory graph to expand the initial memory slots and activate other relevant entity and memory nodes. The output paths of MGN are then used to trigger proactive memory reference, if relevant. The Answer Module executes the predicted neural programs to decode answers given intermediate network outputs. The photo recommender is then called to retrieve relevant photos (*e.g.* photos of the reference memory that includes the answer to a query).

ries relevant to conversational contexts that would make the interaction more engaging and interesting (*e.g.* memories with the same group of people, at the same location as the reference memory, etc.). The two agents continue this process to explore the given memory graph, until one of the agents decides to end the conversation.

2.3 Method

Figure 4 illustrates the overall architecture and the model components of the demonstrated system.

Input Module: For a given query $\mathbf{q}$, its relevant memory nodes $\mathbf{m} = \{\mathbf{m}^{(k)}\}_{k=1}^{K}$ for slot size K are provided as initial memory slots via graph searches. The Query Encoder then encodes the input query with a language model. Specifically, we represent each textual query with a state-of-the-art attention-based Bi-LSTM language model (Conneau et al., 2017). The Memory Encoder then encodes each memory slot based on both its structural features (graph embeddings) and contextual multi-modal features from its neighboring nodes (*e.g.* attribute values). We construct memory graph embeddings to encode structural contexts of each memory node via the graph embeddings projection approaches (Bordes et al., 2013), in which semantically similar nodes are distributed closer in the embeddings space.

Memory Graph Networks: To utilize encoded candidate memory nodes for memory recall QA and proactive memory reference, we first utilize the Memory Graph Networks (MGN) (Moon et al., 2019b). MGN stores memory graph nodes as initial memory slots, where additional contexts and answer candidates can be succinctly expanded and reached via graph traversals. For each $(\mathbf{q}, \mathbf{m}^{(k)})$ pair, MGN predicts optimal memory slot expansion steps: $\mathbf{p}^{(\mathbf{k})} = \{[\mathbf{p}_{e,t}^{(k)}; \mathbf{p}_{n,t}^{(k)}]\}_{t=1}^{T}$ for edge paths $\mathbf{p}_e$ and corresponding node paths $\mathbf{p}_n$. The LSTM-based sequence model is trained to learn the optimal path with ground-truth node and relation paths. The attended memory nodes are then used to answer user memory recall queries (QA module), and to predict relevant memory nodes to surface as a response to the previous conversational contexts (Recommender Module).

QA Modules: An estimated answer $\hat{\mathbf{a}} = \text{QA}(\mathbf{m}, \mathbf{q})$ is predicted given a query and MGN graph path output from initial memory slots.

We then define the memory attention to attenuate or amplify all activated memory nodes based on their compatibility with query, formulated as follows:

$$\beta = \text{MLP}(\mathbf{q}, \{\mathbf{m}^{(k)}\}, \{\mathbf{p}^{(k)}\}) \quad (1)$$

$$\alpha = \text{Softmax}(\mathbf{W}_{\beta}^{\top} \beta) \in \mathbb{R}^{K} \quad (2)$$

Question and Answer	Model Preidction	
	Top-k Answers	Other Relevant Memory Nodes
Q: *Where did Emma and I go after we watched the new Star Wars movie?* // **A**: *Neptune Oyster*	Neptune Oyster AMC Theatre	{Star Wars IV, John, Mar 2014, ...} {Emma, hiking, July 2017, ...}
Q: *When did I last go skiing with Emily?* **A**: *Feb 2017*	Feb 2017 Jan 2016	{skiing, Emily, Dec 2016, ...} {Emily, soccer, ...}
Q: *Show me the photos of when I went to the Dodgers game with Mia this year.* // **A**: [*photo 1*]	[*photo 1*] [*photo 2*]	{baseball, Mia, Dodgers, May 2016, ...} {Mia, movie, April 2018, ...}

Table 1: **Example output**: Model predictions the top-k answers and the attended memory nodes are partially shown for each question and ground-truth answer pair.

Next, the model outputs a module program $\{\mathbf{u}^{(k)}\}$ for several sub-module networks (*e.g.* `CHOOSE`, `COUNT`, ...) via the multi-layer perception module selector network, which outputs the module label probability $\{\mathbf{u}^{(k)}\}$ for each memory node:

$$\{\mathbf{u}^{(k)}\} = \text{Softmax}(\text{MLP}(\mathbf{q}, \{\mathbf{m}^{(k)}\})) \quad (3)$$

Each module network produces an answer vector, the aggregated result of which determines the top-k answers to be returned.

Photo Memory Recommender: The memory attention values (α) and the path outputs are then used to proactively recommend relevant photos associated with each activated memory node $\mathbf{m}^{(k)}$:

$$\{\mathbf{i}^{(k)}\} = \text{Softmax}(\text{MLP}(\alpha^{(k)}, \mathbf{p}^{(k)}))\forall k \quad (4)$$

where the output score is used to rank the candidate photos. Photos with the top score (above threshold) are then finally surfaced along with their memory nodes. We leave this threshold as a tunable hyper-parameter that can determine the proactive behavior of the system in introducing other relevant memories given previous conversational contexts.

3 Demonstration

Table 1 shows some of the example output from the demonstrated system given the input query and memory graph nodes. It can be seen that the model is able to predict answers by combining answer contexts from multiple components (walk path, node attention, neural modules, etc.) In general, the model successfully explores the respective single-hop or multi-hop relations within the memory graph. The activated nodes via graph traversals are then used as input for each neural module, the aggregated results of which are the final top-k answer predictions. The model also attends on other relevant memory nodes which often have some of the key attributes shared with the target reference memory (*e.g.* same activity, people, location, etc.). At test time, we can proactively present these relevant memories to users along with their associated media contents for more engaging memory-grounded conversations.

4 Related Work

End-to-end dialog systems: There have been a number of studies on end-to-end dialog systems, often focused on task or goal oriented dialog systems such as conversational recommendations (Bordes et al., 2017; Sun and Zhang, 2018), information querying (Williams et al., 2017; de Vries et al., 2018; Reddy et al., 2018), etc. Many of the public datasets are collected via bootstrapped simulations (Bordes et al., 2017), Wizard-of-Oz setup (Zhang et al., 2018; Wei et al., 2018; Moon et al., 2019a), or online corpus (Li et al., 2016). In our work, we propose a unique setup for dialog systems called memory-grounded conversations, where the focus is on grounding human conversations with past user memories for both the goal-oriented task (memory recall QA) and the more open-ended dialogs (proactive memory reference). Our *Memory Dialog* dataset uses the popular Wizard-of-Oz setup between role-playing human annotators, where the reference memories are bootstrapped through memory graph generator.

QA Systems: Structured QA systems have been very popular due to the popularity of the fact-retrieval assistant products, which solve fact-retrieval QA queries with large-scale common fact knowledge graphs (Bordes et al., 2015; Xu et al., 2016; Dubey et al., 2018). Most of the work typically utilize an entity linking system and a QA model for predicting graph operations *e.g.* through template matching approaches, etc. For QA systems with unstructured knowledge sources (*e.g.* machine reading comprehension), the approaches

that utilize Memory Networks with explicit memory slots (Weston et al., 2014; Sukhbaatar et al., 2016) are widely used for their capability of transitive reasoning. In our work, we utilize Memory Graph Networks (MGN) (Moon et al., 2019b) to store graph nodes as memory slots and expand slots via graph traversals, to effectively handle complex memory recall queries and to identify relevant memories to surface next.

Visual QA systems answer queries based on the contexts from provided images (Antol et al., 2015; Wang et al., 2018; Wu et al., 2018). Jiang et al. (2018) propose the visual memex QA task which addresses similar domains given a dataset composed of multiple photo albums. We extend the problem domain to the conversational settings where the focus is the increased engagement with users through natural multi-modal interactions. Our work also extends the QA capability by utilizing semantic and structural contexts from memory and knowledge graphs, instead of relying solely on meta information and multi-modal content available in photo albums.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. *ICLR*.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory network. *arxiv*.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *NIPS*.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *EMNLP*.

Mohnish Dubey, Debayan Banerjee, Debanjan Chaudhuri, and Jens Lehmann. 2018. Earl: Joint entity and relation linking for question answering over knowledge graphs. *ESWC*.

Lu Jiang, Junwei Liang, Liangliang Cao, Yannis Kalantidis, Sachin Farfade, and Alexander Hauptmann. 2018. Memexqa: Visual memex question answering. *arxiv*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. *ACL*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019a. Opendialkg: Explainable conversational reasoning with attention-based walks over knowledge graphs. *ACL*.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019b. Walk the memory: Memory graph networks for explainable memory-grounded question answering. *CoNLL*.

Siva Reddy, Danqi Chen, and Christopher D Manning. 2018. Coqa: A conversational question answering challenge. *arXiv preprint arXiv:1808.07042*.

Pararth Shah, Dilek Hakkani-Tur, Bing Liu, and Gokhan Tur. 2018. Bootstrapping a neural conversational agent with dialogue self-play, crowdsourcing and on-line reinforcement learning. In *NAACL*.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2016. End-to-end memory networks. *NIPS*.

Yueming Sun and Yi Zhang. 2018. Conversational recommender system. *SIGIR*.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *ECCV*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *PAMI*.

Wei Wei, Quoc Le, Andrew Dai, and Jia Li. 2018. Airdialogue: An environment for goal-oriented dialogue research. In *EMNLP*.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Jason D Williams, Kavosh Asadi, and Geoffrey Zweig. 2017. Hybrid code networks: practical and efficient end-to-end dialog control with supervised and reinforcement learning. *ACL*.

Qi Wu, Chunhua Shen, Peng Wang, Anthony Dick, and Anton van den Hengel. 2018. Image captioning and visual question answering based on attributes and external knowledge. *PAMI*.

Kun Xu, Siva Reddy, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2016. Question answering on freebase via relation extraction and textual evidence. *ACL*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *ACL*.