10-2018

Knowledge-aware multimodal dialogue systems

Lizi LIAO
Singapore Management University, lzliao@smu.edu.sg

Yunshan MA

Xiangnan HE

Richang HONG

Tat-Seng CHUA

Knowledge-aware Multimodal Dialogue Systems

Lizi Liao[1], Yunshan Ma[1], Xiangnan He[1], Richang Hong[2], Tat-Seng Chua[1]
[1]National University of Singapore, [2]Hefei University of Technology
{liaolizi.llz, mysbupt, xiangnanhe, hongrc.hfut}@gmail.com, chuats@comp.nus.edu.sg

ABSTRACT

By offering a natural way for information seeking, multimodal dialogue systems are attracting increasing attention in several domains such as retail, travel etc. However, most existing dialogue systems are limited to textual modality, which cannot be easily extended to capture the rich semantics in visual modality such as product images. For example, in fashion domain, the visual appearance of clothes and matching styles play a crucial role in understanding the user's intention. Without considering these, the dialogue agent may fail to generate desirable responses for users.

In this paper, we present a Knowledge-aware Multimodal Dialogue (KMD) model to address the limitation of text-based dialogue systems. It gives special consideration to the semantics and domain knowledge revealed in visual content, and is featured with three key components. First, we build a taxonomy-based learning module to capture the fine-grained semantics in images (*e.g.*, the category and attributes of a product). Second, we propose an end-to-end neural conversational model to generate responses based on the conversation history, visual semantics, and domain knowledge. Lastly, to avoid inconsistent dialogues, we adopt a deep reinforcement learning method which accounts for future rewards to optimize the neural conversational model. We perform extensive evaluation on a multi-turn task-oriented dialogue dataset in fashion domain. Experiment results show that our method significantly outperforms state-of-the-art methods, demonstrating the efficacy of modeling visual modality and domain knowledge for dialogue systems.

KEYWORDS

Multimodal Dialogue, Domain Knowledge, Fashion

ACM Reference Format:

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, Tat-Seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3240508.3240605

1 INTRODUCTION

The design of intelligent assistants that can interact directly with human ranks high on the agenda of current AI research. Recently, we have seen some truly remarkable conversational agents on the market (*e.g.* Apple Siri, Microsoft Cortana, and Google Allo). However, most of these agents only focus on textual (or voice) modality, performing simple tasks and answering factual questions [10]. As evidenced by the increasing demand for multimodal conversational agents in domains like e-commerce retail, travel, and entertainment *etc.*, there exists an urgent need for a more natural and informative way to satisfy user's information need [31]. For example, when a user searches for a dress in a particular style or chooses tourist attractions to visit, an effective multimodal dialogue agent would serve them in a more intuitive and interactive manner as shown in Figure 1. Moreover, the expressive visual modality enables vibrant UI design and alleviates difficulties faced by text response generation such as the ability to describe certain visual attributes of fashion products.

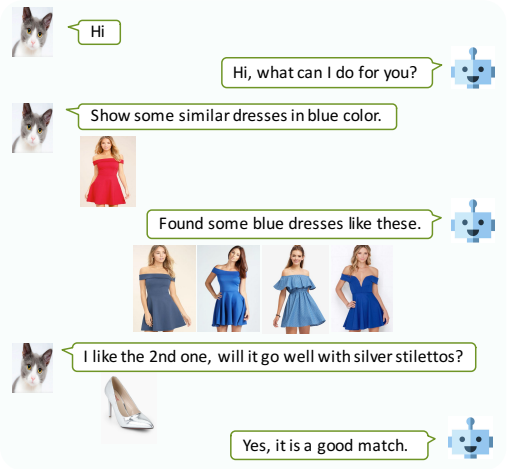

Figure 1: An example of knowledge-aware multimodal dialogue for fashion retail. The agent manages to understand the semantics of product image and modify attributes during back-end retrieval, offer matching suggestions for the user, and generate responses with different modalities.

Although multimodal conversational agents show various advantages in helping users, it is non-trivial to make it really "smart" in generating substantive answers. First, as the fashion domain example illustrated in Figure 1, in order to properly respond to the user's request about similar dresses in blue color, the agent needs to correctly understand the visual semantics of the product image in the first place. Second, when forming queries for product retrieval in the back-end, it should be able to make accurate attribute modifications (*e.g.*, changing from red color to blue). Lastly, to answer user's question about whether the blue skater dress matches with the silver stilettos, the agent should have the capability of leveraging fashion style tips. Similar multimodal scenario can be formulated for other domains such as travel and entertainment. Therefore, we believe that an essential requirement for building intelligent multimodal agents is to capture the semantics in visual

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10…$15.00
https://doi.org/10.1145/3240508.3240605

modality and the underlying domain knowledge. This is especially relevant to specific domains like fashion and travel, where many domain knowledge are multimodal in nature, varying from visual semantics about what does a Gothic architecture look like, to style tips about whether black gown matches with pearl necklace, or what kind of outfit is better suited for which social occasions.

Indeed, there have been several efforts in incorporating domain knowledge into dialogue systems and demonstrated promising results. For example, [15] proposed a rule-based method by filling the response templates with entries from an extracted knowledge base. [12] augmented conversation history with relevant unstructured facts such as Foursquare tips mined from online reviews. [52] built a music domain specific knowledge base to facilitate substantive conversations. However, the above-mentioned efforts on knowledge-aware dialogue systems are all limited to textual modality. The use of visually-aware knowledge, such as the semantics in product or travel images and matching style tips, has not been considered in current dialogue systems. How to leverage such multimodal knowledge to generate better responses in dialogue systems is a challenging yet untapped problem.

In this paper, we propose a knowledge-aware multimodal dialogue model (KMD) as shown in Figure 2 and apply it to the fashion domain. Firstly, to enable the agent to understand fine-grained semantics in product images, we develop a taxonomy-based visual semantic learning module to represent the product in a continuous vector space. Secondly, we embed fashion style tips into the same space and store in a memory network [41]. When generating responses, the agent employs an attention mechanism to adaptively attend to the domain knowledge and decide which knowledge entry is useful. The key idea is that the agent conditions answers based not only on conversation history, but also on the extracted knowledge that are relevant to the current context. For modeling the conversation procedure, we employ the state-of-the-art hierarchical recurrent encoder-decoder (HRED) model [32] as the backbone network. Thirdly, to avoid inconsistent dialogues and error accumulation problems, we apply deep reinforcement learning to model future rewards that characterize good conversations. During training, the agent learns a policy, the parameters of HRED, by optimizing the long-term rewards from dialogues using policy gradient methods.

To sum up, the main contributions of this work are as follows:

- Starting from HRED, we propose to further understand fine-grained visual semantics and leverage domain knowledge to enable the agent to generate more substantive responses.
- We integrate the strength of deep reinforcement learning in optimizing for better rewards with the power of hierarchical seq2seq models in modeling sequential utterances.
- We conduct extensive experiments to evaluate the proposed framework in various evaluation metrics and show superior performance over state-of-the-art methods.

2 RELATED WORK

2.1 Multimodal Dialogue Systems

2.1.1 Dialogue systems. Human-computer conversation has attracted increasing attention owning to its promising potentials and alluring commercial values [7]. According to the applications, it can be roughly categorized into two groups: task-oriented systems and non-task-oriented systems. Also known as chat bots, the non-task-oriented systems converse with human typically on open domains to provide reasonable responses and entertainment. In this paper, we focus on the task-oriented systems which aim to assist users to complete certain tasks such as finding products.

The widely applied approaches for task-oriented systems treat the procedure of generating dialogue responses as a pipeline [17, 28, 46]. It first encodes human utterances as an internal dialogue state, then takes some actions according to the policy with respect to the state, and finally transforms the action to form a natural language response. However, such methods suffer from the credit assignment problem since errors from upstream modules can propagate to the rest of the pipeline. Also, the interdependence among processes within these pipeline methods makes online adaptation of certain components challenging. More importantly, the heavy reliance on annotated training data hinders its usage to domains such as fashion or travel which usually involve large volume of data and great information diversity. Thus, it is limited primarily to simple tasks such as querying bus routine, booking movie or flight ticket, and finding restaurants.

Recently, with the development of big data and deep learning techniques, there emerge attempts to build end-to-end task-oriented dialogue systems [42, 51], which can expand the state space representation in the traditional pipeline systems and generate novel utterances with more flexible syntactical structures. Among the first works of end-to-end dialogue systems, [33] extended their HRED model to the dialogue domain. [5, 40] treated a dialog system as a mapping problem between the dialogue history and the system response. They learned this mapping via novel variants of the encoder-decoder model. [51] differed from them by learning a strategic plan using reinforcement learning and jointly optimizing state tracking. [10, 19] trained the end-to-end system as a task completion neural dialogue system with user simulations. However, all these efforts are restricted to textual modality.

2.1.2 Multimodal Dialogue. With great advances in understanding the informative visual modality, multimodal conversational agents are gaining importance. Most recently, the authors of [31] contributed a Multimodal Dialogues (MMD) benchmark dataset in fashion domain. It consists of over 150K conversation sessions and contains domain knowledge curation. The authors extended the HRED model by simply concatenating visual features with text representations. Although 350K fine-grained style tips as well as many fashion attributes and synsets were provided as shown in Table 1, the authors did not leverage such knowledge in conversation modeling. Our work will take a step forward to build knowledge-aware multimodal dialogue systems.

Another body of work relevant to ours would be the Vision-to-Language problems such as image captioning and visual question answering (VQA). Much recent progress in such problems has been achieved through a combination of Convolutional Neural Networks (CNN) and recurrent neural networks [44]. In image captioning, current state-of-the-art methods follow the general framework where a CNN is used as an image 'encoder' to produce image representation, which is then fed into the 'decoder' LSTM to generate a

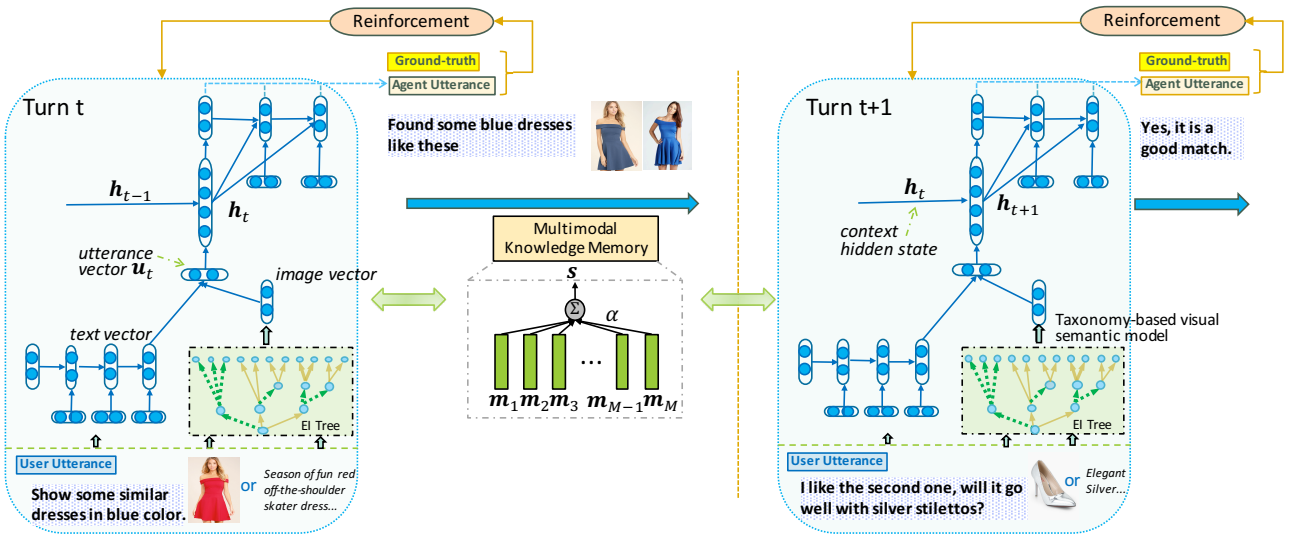

Figure 2: The knowledge-aware multimodal dialogue framework (KMD). Two turns of conversations are illustrated. The agent leverages taxonomy-based visual semantic model to understand user utterances in different forms. It generates various forms of responses enriched with extracted domain knowledge. Deep reinforcement learning measures the goodness of a response through a reinforcement signal and optimizes the long-term rewards that characterize a good conversation.

caption (without attention mechanism [39] or with attention mechanism [23, 45]). In VQA, similar frameworks have been adopted and promising results have been achieved. However, as pointed out in [13, 50], there exist strong language priors which lead to good superficial performance of these models without truly understanding the visual content. Therefore, in our work, a taxonomy-based visual semantic model is built to explicitly represent fashion concepts.

While VQA [2] involved a single question and response, the work of visual dialog [8, 27] handled a sequence of QA pairs with a single image forming a dialogue. [9] introduced a two-player guessing game to collect a large-scale dataset and proposed end-to-end optimization of goal-driven and visually grounded dialogue systems in [36]. However, as pointed out in [27, 31], the problem setting for these works actually belongs to image-grounded QA rather than multimodal dialogues. For example, most of these works focus on reasoning from a single image and the responses are always textual. In natural conversations among humans, there could be multiple images providing context and the context images could change across turns during the course of the conversation. Also, the system should be able to retrieve and organize images as responses, which encourages the computer to play its strength in processing speed and storage.

2.2 Incorporation of Knowledge

As illustrated in Figure 1, domain knowledge is essential for dialogue systems to generate reasonable and substantive responses. There have been several efforts incorporating domain knowledge into dialogue systems and shown promising results [12, 15, 52]. The knowledge can either be from external knowledge bases such as WordNet[25], DBPedia[3] or NELL[26], or from unstructured facts mined from online resources.

Based on the way of incorporating knowledge, existing studies can be categorized into different types. One type of work leverages symbolic features derived from knowledge bases (KBs) [29]. This is not ideal as the symbolic features have poor generalization ability. Moreover, such systems query outside knowledge base (KB) by issuing a symbolic query to retrieve entries [10, 19] and the retrieval operation is non-differentiable. Another type of work learns distributed representations of structured knowledge from large KBs [6, 14, 47] and has been widely applied to generative QA and dialogue systems. For example, inspired by the key-value memory networks [24], the authors in [11] augmented existing recurrent network architectures with a differentiable attention-based key-value retrieval mechanism. Moreover, [10] replaced symbolic queries with an induced "soft" posterior distribution over the knowledge base which indicates the user's interests. [1] generated knowledge-related words by copying from the description of the predicted fact while [12] took unstructured text as external knowledge to enhance the traditional chit-chat dialogue systems. However, these studies are largely restricted to textual knowledge. In our work, we not only incorporate taxonomy-based structured knowledge into visual semantic understanding, but also enrich the dialogue context with extracted multimodal knowledge.

3 METHOD

Figure 2 illustrates our knowledge-aware multimodal dialogue (KMD) method. There are three major components. (1) In each turn, given a multimodal utterance, the agent attempts to understand the semantics inherent in product images via a taxonomy-based learning model which captures the category and attributes of product. (2) Besides modeling utterances using the HRED network, the agent employs an attention mechanism over the extracted domain knowledge and decides which knowledge is relevant to the current context. The agent thus generates responses based on the conversation history and relevant knowledge stored in a memory network. (3) Based on the extended HRED backbone network, we apply deep reinforcement learning that accounts for future awards to optimize the neural conversational model using policy gradient methods.

Table 1: Domain Specific Knowledge Base Statistics.

Knowledge Base Statistics		Examples
#Fashion Synsets	716	shirt, trouser, tuxedo, loafer, stilletoes, sunglasses, handbag, hat
#Coarse-grained StyleTips	8871	shirt & trouser, tshirt & sneakers, tuxedo & cufflinks, suit & formal shoes, kurta & jeans
#Fine-grained StyleTips	350K	white shirt & black trousers, light tshirt & dark jacket, black gown & pearl necklace

Formally, at turn t, given the user utterances $u_1, u_2, \cdots, u_t$ and previous agent responses $\hat{u}_1, \hat{u}_2, \cdots, \hat{u}_{t-1}$, the agent needs to generate a response $\hat{u}_t$. Each utterance (u or $\hat{u}$) can be of different modalities. For example, there might be products mentioned by either user or agent during the conversation, and these products are usually described as images. An utterance may contain multiple product images. To introduce the domain knowledge, we present the basic statistics in Table 1 and provide a detailed example of the query in Example 1.

Example 1. **User:** I like the 2nd one, will it go well with silver stilettos?

The relevant style tips would be $\mathcal{G} = \{g_1, g_2, g_3, \cdots\}$ where
$g_1 = \{blue\ skater\ dress, match\ with, silver\ stilettos\}$
$g_2 = \{blue, match\ with, silver\}$
$g_3 = \{skater\ dress, match\ with, stilettos\}$

Note that products such as *blue skater dress* in style tips and the *silver stilettos* might not co-occur in the whole training corpus. Therefore, it would be rather hard for the agent to generate proper response without leveraging these external knowledge.

3.1 Taxonomy-based Visual Semantic Learning

In multimodal dialogue systems, the correct understanding of semantics of product images is essential for generating appropriate responses. To achieve this, we may directly apply the well-trained convolutional neural networks to extract image features. However, such generic approach is unable to capture the rich domain knowledge in specific domain [49]. In the fashion domain, the agent should have a clue about human perception of product organization and product similarity. For instance, a T-shirt belongs to the up_cloth category but not the bottom_cloth; and thus it does not have details such as rise or fry. Also, details such as sleeve length and color can be decided independent of each other.

Therefore, we build an Exclusive&Independent tree (EI tree) structure which organizes the semantic concepts from general to specific, where exclusive and independent relationships are integrated among siblings. For example, sibling concepts involving product categories usually share exclusive relationship, while sibling concepts involving attributes are often characterized by independent relationships. We crawled product hierarchies from 40 e-commerce sites such as *amazon.com* and applied the Bayesian Decision approach provided in [38] to unify the hierarchies. We then extracted the exclusive as well as independent relationships and built the EI tree manually by a fashion expert. Figure 3 shows part of the resulting fashion EI tree with top level concepts such as $up, bottom, color, pattern$ *etc.*

To learn the visual semantics of product images, we leverage the taxonomy knowledge enriched EI tree to train an EITree model as depicted in Figure 4 using the product images and text descriptions provided in meta-data. The EITree model is trained by mapping

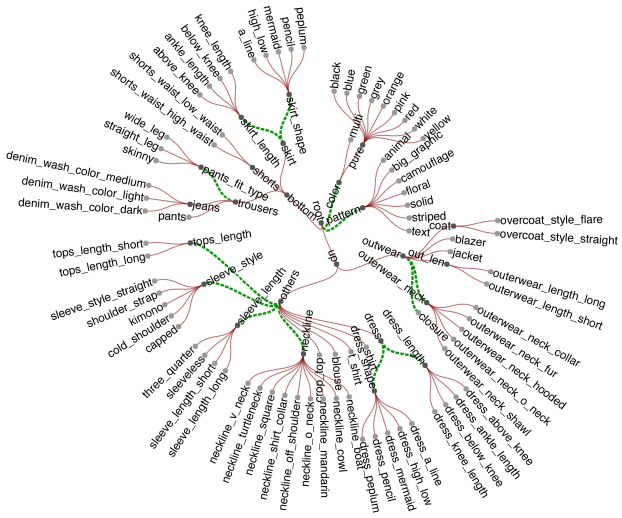

Figure 3: Part of an EI tree taxonomy for fashion concepts. The green dash lines denote independent relations among siblings while red solid lines denote exclusive relations.

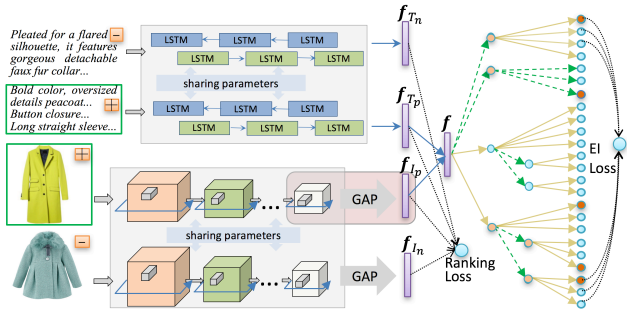

Figure 4: Taxonomy-based visual semantic model (EITree).

the implicit deep features to explicit fashion concepts via our constructed EI tree. Each concept is traced from the root to itself along the tree and a probability is generated based on the tracing path, which mimics the general to specific recognition procedure. Intuitively, a softmax constraint is put among the exclusive siblings, forcing the model to choose only one of them; the independent siblings are decided independently. Formally, suppose $c_0 \rightarrow c_n$ is the semantic path to concept c_n, $\mathbf{f}$ is the integration of visual and textual features, and $\mathbf{W}_{EI} \in \mathbb{R}^{2048 \times |C|}$ is the EI weight matrix (c_0 denotes the root), the probability of concept c_n is:

$$p(c_n|\ c_0 \rightarrow c_n, \mathbf{f}, \mathbf{W}_{EI}) = p(c_1|\ c_0, \mathbf{f}, \mathbf{W}_{EI}) \cdot p(c_2|\ c_1, \mathbf{f}, \mathbf{W}_{EI}) \cdots p(c_n|\ c_{n-1}, \mathbf{f}, \mathbf{W}_{EI}),$$

which can be viewed as a sequence of steps along the path. Note that there are two kinds of steps in Figure 4: the green dashed line denotes the independent step $l_{c_{n-1}c_n} \in \mathcal{E}_I$ while the brown solid line denotes the exclusive step $l_{c_{n-1}c_n} \in \mathcal{E}_E$. We keep exclusive siblings of each node as ES_{c_n}. Thus, the probability of each step is:

$$p(c_n|\ c_{n-1}, \mathbf{f}, \mathbf{W}_{EI}) = \begin{cases} \frac{exp(\mathbf{f}^T \cdot \mathbf{W}_{EI} \cdot \mathbf{c}_n)}{\sum_{k \in ES_{c_n}} exp(\mathbf{f}^T \cdot \mathbf{W}_{EI} \cdot \mathbf{c}_k)} & l_{c_{n-1}c_n} \in \mathcal{E}_E \\ \sigma(\mathbf{f}^T \cdot \mathbf{W}_{EI} \cdot \mathbf{c}_n) & l_{c_{n-1}c_n} \in \mathcal{E}_I \end{cases}$$

where $\mathbf{c}_n$ denotes the one hot vector for node c_n, and $\sigma(\cdot)$ denotes the sigmoid function.

To fulfill the whole training procedure, we compare the outputs for each leaf concept against the ground truth labels, which resumes the cross-entropy loss. In order to match the textual and visual modalities, a bi-directional ranking loss is used as a regularizer. We apply the Adam optimizer to train this taxonomy-based visual semantic embedding model and then use it for extracting semantic representations. For more details about the EITree model, please refer to [20]. We use the **f** (as denoted in Figure 4) concatenated with last layer outputs as the representations for products.

3.2 Incorporation of Domain Knowledge

To model sequential utterances, we resort to the powerful hierarchical sequence-to-sequence models such as HRED and extend it to our multimodal scenario. We then embed style tips into vector space using the features extracted from the EITree model and incorporate such knowledge into the HRED structure via memory network.

3.2.1 Basic HRED and Extensions. The general procedure for the original text-based HRED [33] is shown in Figure 5. At the word level, the encoder RNN maps each utterance to an utterance vector representation $\mathbf{u}_t$, which is the hidden state obtained after the last token of the utterance has been processed. At the utterance level, the context RNN keeps track of past utterances by iteratively processing each utterance vector and generates the hidden state $\mathbf{h} \in \mathbb{R}^d$. Each hidden state $\mathbf{h}_t$ of the context RNN represents a summary of dialogue up to and including the user utterance in turn t, which is used to predict the response in turn t. The response prediction is performed by means of a decoder RNN, which takes the hidden state of the context RNN and produces a probability distribution over the tokens in the next response.

In our multimodal dialogue scenarios, the utterance vector $\mathbf{u}_t$ is a concatenation of the encoded text representation and the visual product representation (when available). In producing text responses, we couple a standard RNN decoder (GRU cells) with an attention model which learns to attend to different time-steps of the second level encoder. It has been used successfully for various natural language generation tasks including text conversation systems [34]. In generating image responses, we treat it as a ranking task to rank a given set of images depending on their relevance to the context. We train the model using a max margin loss. Specifically, we compute the cosine similarity between the learned image representation and the encoded multimodal context representation. The model is then trained to maximize the margin between the cosine similarity for the correct and incorrect images.

3.2.2 Incorporation of Knowledge via Memory Network. In modeling the conversation, similar to [31], we extended a HRED model that contains a dialog-RNN sitting on top of a recurrent block. At each time step t, we regard $\mathbf{h}_t$ as the summary of the input so far. However, it might not be sufficient. For instance, if the user asks for advice about matching tips of gladiator sandals, the matching candidates such as the denim skirts might not co-occur with it in the conversation context or even the whole training corpus. Therefore, it would be rather hard for the agent to generate answer containing denim skirts. To address such problem, a proper way is

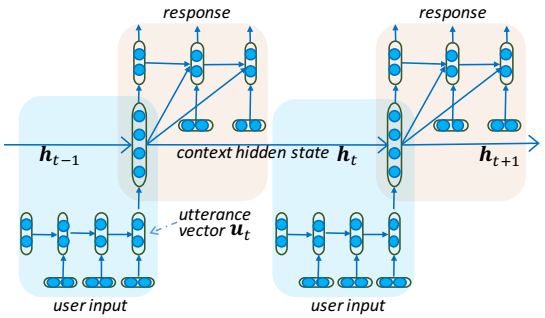

Figure 5: An illustration of text-based HRED backbone.

to incorporate multimodal knowledge via memory networks [41]. As widely used in QA to make inferences based on facts, it uses an associative memory for modeling the knowledge, then retrieves and weights the appropriate knowledge based on the input query. In our case, we need to capture the multimodal knowledge of the fashion items mentioned in a conversation. Thus, we apply the trained EITree model to extract multimodal feature representations for style tips provided in the dataset [31], such as black trousers go well with white shirts. We obtain average representations for black trousers and white shirts respectively. We then concatenate such representations to obtain a knowledge entry $\mathbf{g}_i \in \mathbb{R}^v$ similar to [1]. Such entries are stored in the memory network and stay fixed.

We then incorporate these knowledge into encoder state. Note that $\mathbf{g}_i$ refers to the vector representation of knowledge i, we have:

$$\mathbf{m}_i = \mathbf{A}\mathbf{g}_i \tag{1}$$

$$\mathbf{o}_i = \mathbf{B}\mathbf{g}_i \tag{2}$$

$$\alpha_i = \frac{exp(\mathbf{h}_t^T \mathbf{m}_i)}{\sum_{k=0}^{M} exp(\mathbf{h}_t^T \mathbf{m}_k)} \tag{3}$$

$$\mathbf{s} = \sum_{i=1}^{M} \alpha_i \mathbf{o}_i \tag{4}$$

where M is the total number of knowledge entries. $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{d \times v}$ are the parameters of the memory network. The former one embeds the knowledge $\mathbf{g}_i$ into memory representation while the later one transforms it to a corresponding output vector. Then the hidden state of the decoder RNN is initialized with $\mathbf{h}'_t$ which is a synthesization of input sentence and the external facts as below:

$$\mathbf{h}'_t = \mathbf{h}_t + \mathbf{s}. \tag{5}$$

Instead of summing up dialogue encoding and integrated knowledge as in Equation 5, we also experimented with other operations such as concatenation, but summation seemed to yield the best results. Similar observations are reported in [12].

3.3 End-to-End Reinforcement Learning

With the HRED backbone and multimodal knowledge incorporated, the agent manages to take in multimodal utterances and generate responses turn by turn. However, one of the drawbacks of training the agent in a supervised learning setup (as in HRED) is that it may result in inconsistent dialogues and that errors can accumulate over time. For example, the agent might tend to answer "Sorry could not find anything similar" due to the high frequency of occurrence of such responses in the training data, and its compatibility with a

diverse range of conversational contexts. However, this response is not a good one since it closes down the conversation. Therefore, we propose to fine-tune the response generation by adapting the popular REINFORCE [43] algorithm with proper bias correction using the learned "baseline". By applying reinforcement learning method, we can measure the goodness of a response through a reinforcement signal. Adjustments can be made to increase the chance of selecting the responses receiving positive reinforcement and to reduce the chance of responses with negative reinforcement.

Formally, a dialogue can be represented as an alternating sequence of utterances $\{u_1, \hat{u}_1, u_2, \hat{u}_2, \cdots, u_t, \hat{u}_t\}$ generated by the user and the agent. We view the agent generated utterances as actions that are taken according to a policy defined by the HRED backbone network. After taking an action, the agent receives a reward and back-propagates to the HRED model. The parameters of the network are optimized to maximize the expected future reward. Following numerous works on applying the encoder-decoder architectures with RL methods [4, 22, 30], we use the BLEU score as a reward signal R_{BLEU} to fine-tune the text response network which is trained with a cross-entropy loss. Since the average number of words in agents' text responses is not large (only 14 as shown in Table 2), we consider up to 4-grams for BLEU. For image responses, we directly apply the similarity between target image and the positive/negative images $R_{SIM} = sim(\mathbf{I}, \mathbf{I}_{pos}) - sim(\mathbf{I}, \mathbf{I}_{neg})$ as the reward signal. We fine-tune the image response network which is trained with a max margin loss. For ease of simplicity, we use R to uniformly denote the R_{BLEU} and R_{SIM} for different response networks.

Like in imitation learning, we have a training set of optimal sequences of actions. During training we choose actions according to the current policy and observe rewards by comparing the actions from the current policy against the optimal actions. The goal is to find the parameters of the agent that maximize the expected reward. Thus, we define our loss as the negative expected reward:

$$\mathcal{L}_\theta = -\mathbb{E}[R(\hat{u}_t, C)], \tag{6}$$

where $C = [u_1, \hat{u}_1, u_2, \hat{u}_2, \cdots, u_t]$ denotes the previous context utterances. In our experiment, we choose up to five previous utterances of $\hat{u}_t$. The gradient is estimated using the likelihood ratio trick as below:

$$\nabla \mathcal{L}_\theta = -R(\hat{u}_t, C) \nabla \, log \, p_{RL}(\hat{u}_t|C), \tag{7}$$

where $p_{RL}(\hat{u}_t|C)$ is the probability of generating $\hat{u}_t$ under C by RL. We update the parameters in the HRED networks using stochastic gradient descent. A baseline strategy is employed to decrease the learning variance, similar to [30]. Thus, the updated gradient is

$$\nabla \mathcal{L}_\theta = - \nabla \, log \, p_{RL}(\hat{u}_t|C) \, [R(\hat{u}_t, C) - \bar{R}], \tag{8}$$

where $\bar{R}$ is the reward for a baseline method. The model either encourages a utterance choice $\hat{u}_t$ if $R(\hat{u}_t, C) > \bar{R}$, or discourages it if $R(\hat{u}_t, C) < \bar{R}$. In our implementation, for the text response task, the baseline is estimated by a linear regressor which takes as input the hidden states of HRED. The regressor is an unbiased estimator of future rewards since it only uses past information. For the image response task, similar to [18, 48], we leverage an additional neural model that takes as inputs the generated target and the initial source and outputs a baseline value.

4 EXPERIMENTS

In this section, the experiments are carried out to answer the following research questions:

- **RQ1** Can our KMD model help the agents to make good use of multimodal knowledge and generate better responses?
- **RQ2** What are the effects of incorporating different knowledge components in the proposed KMD method?
- **RQ3** Will deep reinforcement learning help to improve the performance of our multimodal dialogue systems?

4.1 Experimental Setups

4.1.1 Datasets. Arguably the greatest bottleneck for statistical approaches to dialogue system development is the collection of appropriate training dataset, and this is especially true for task-oriented dialogue systems [40]. Fortunately, [31] contributed a dataset consisting of over 150K conversation sessions between shoppers and sales agents. The dialogues seamlessly incorporate multimodal data in utterances and also demonstrate domain-specific knowledge during the series of interactions. Detailed information about these dialogues are listed in Table 2. We carried out experiments on this dataset as provided in [31]. Part of an example dialogue session is shown in Figure 6. We train our method to generate responses as the agent in Figure 6.

Table 2: Multimodal Dialogue Dataset Statistics.

Dataset Statistics	Train	Valid	Test
#Dialogues (chat sessions)	105,439	22,595	22,595
Proportion in terms of dialogues	70%	15%	15%
Avg. #Utterances per dialogue	40	40	40
#Utterances with shopper's questions	2M	446K	445K
#Utterances with image response	904K	194K	193K
#Utterances with text response	1.54M	331K	330K
Avg. #Positive images in image response	4	4	4
Avg. #Negative images in image response	4	4	4
Avg. #Words in shopper's Question	12	12	12
Avg. #Words in text Response	14	14	14

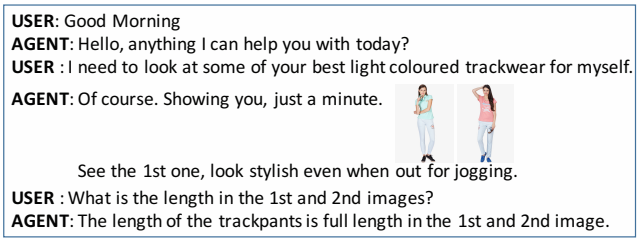

Figure 6: Example of a partial dialogue session.

4.1.2 Comparing Methods. To evaluate the effectiveness of the proposed knowledge-aware multimodal dialogue framework, we compare it with the following three representative solutions. a) **HRED (text-only)** [32] adopts the hierarchical recurrent encoder-decoder network to minimize a single objective function in an end-to-end fashion with minimum reliance on hand-crafted features. b) **MemNN** [5] constructs an end-to-end dialogue system based on Memory Networks that can store historical dialogues and short-term context to reason about the required response. To capture visual information, it concatenates visual features with text

Table 3: Performance of the different models on text response generation and image response generation (RQ1).

Method	Text Response		Image Response (k = 5)		
	BLEU	Diversity (unigram)	R@1	R@2	R@3
HRED (text-only)	0.3174	0.00369	0.4323	0.6217	0.7486
MemNN	0.5013	0.00435	0.7800	0.8372	0.9091
MHRED	0.5195	0.00426	0.7980	0.8859	0.9345
KMD	**0.6731**	**0.00534**	**0.9198**	**0.9552**	**0.9755**

features. c) **MHRED** (short for multimodal HRED) [31] extends the basic HRED with both textual and visual modalities. We adopt the version of MHRED with attention mechanism. Note that it can be seen as our **KMD** model without the three major components as described in Section 3. For all these baselines, we apply the 4096 dimensional representation provided in [31] which is obtained from the FC6 layer of a VGGNet-16, while all of our methods use the features extracted via the taxonomy-based visual semantic embedding model (tagged as **+TK**). In order to analyze the effect of incorporating the major components, we also compare the performance of three variants of our model as follows: d) **MHRED+TK** which only incorporates taxonomy knowledge, e) **MHRED+TK+EK** which additionally handles extracted external knowledge such as style tips, and f) **MHRED+TK+RL** which optimizes rewards that characterize good conversations.

4.1.3 Training Setups. For the first stage of training, we built on prior work of predicting a generated target utterance given the dialogue history using the knowledge enriched multimodal HRED model in a supervised fashion. We trained the knowledge enriched multimodal HRED model on the training dataset. Each response turn in the dataset was treated as a target and the concatenation of five previous utterances were treated as context. We used the Adam optimizer with the learning rate initialized to 0.001 and decayed under default settings. The batch size was set to 64.

For the second stage of training, following the popular strategy in RL training as in [18, 30, 37, 43], we initialized the policy model using the knowledge enriched multimodal HRED model trained during the first stage. This ensures that we start off with a much better policy than random because the model can now focus on a relatively good part of the search space.

4.1.4 Evaluation Protocols. For the text response generation, we use the BLEU scores following [31, 35]. It is based on the idea of modified n-gram precision, where the higher score denotes better performance. We also report the lexical diversity scores by calculating the number of distinct unigrams in generated responses. The value is scaled by the total number of generated tokens to avoid favoring long sentences [18]. For image response generation, we use Recall@top-k as the evaluation metrics where k is varied from 1 to 3, and the model prediction is considered to be correct only if the true response is among the top-k entries in the ranked list.

4.2 Evaluating the Text Response

Table 3 shows the performance comparisons between different models on response generation. We first focus our comparison on the text responses. We report the corresponding scores on BLEU and unigram diversity. Four sample responses of the MHRED and the KMD method are provided in Figure 7. Due to space limitation, we omit the former utterances. The key observations are as follows.

First of all, compared to the pure text-based HRED method, the other methods working on multimodal information perform significantly better. It suggests that adding images indeed improves the response capability of fashion agents and validates the motivation behind the building of multimodal conversation systems. Intuitively, fashion domain involves multimodal data by nature. As the sample responses illustrated in Figure 7, there are rich semantics expressed via images, and there are many visual traits of fashion items that are not easily translated into words. For example, it might be hard to describe the watches by pure text in the last example. Therefore, building multimodal dialogue systems that can handle both text and image is a viable way to better assist the customers.

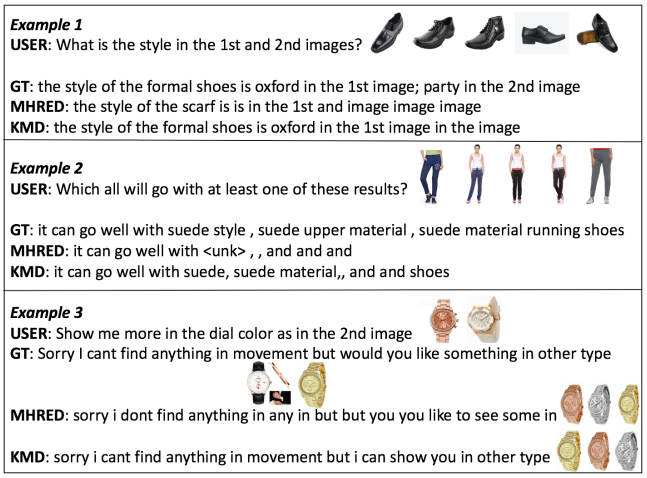

Figure 7: Sample responses generated from the MHRED and the KMD method (GT stands for ground truth response).

Secondly, the proposed method KMD achieves the best performance among all methods. The performance improvements of KMD over the other methods are significant. For example, in terms of BLEU score, KMD improves the performance of text response generation by 25.5% and 22.8% as compared to the multimodal information enriched MemNN and MHRED, respectively. Note that the backbone network of KMD is actually the MHRED. Thus, the superior performance of the proposed method demonstrates the usefulness of incorporating knowledge. The higher unigram diversity score also indicates that the proposed method generates more diverse outputs when compared against the other methods. More detailed analysis will be provided in the ablation study later.

4.3 Evaluating the Image Response

We also compare the results of different models on the image response generation task in Table 3, where k refers to the size of target image set to be ranked by the model (one is correct and the rest are incorrect).

Clearly, we observe that the proposed KMD method outperforms all the other baselines. Specifically, the text-only HRED performs the worst which is as expected. This is because only textual information is captured in the context hidden state which is then leveraged to calculate similarities with candidate images for ranking. Many useful visual information along the conversation is ignored, thus resulting in poor performance. When such visual information is incorporated, we observe a performance jump as in the MemNN and MHRED methods. Further, when different knowledge components are integrated into the KMD framework, the best performances are achieved, which again lends support to the knowledge-aware design.

One thing to note here is that the candidate set we use is relatively small. Thus the Recall@top-k scores are rather high. In real life scenarios where there might be a large number of candidate images instead of five, we may see a sharp decline in the performance. To alleviate the effect, we will need to process and organize the product repository beforehand to generate better candidate list.

Table 4: Performance of ablation study of the KMD framework on text response (RQ2 & RQ3).

Method	BLEU	Diversity (unigram)
MHRED	0.5195	0.00426
MHRED+TK	0.5729	0.00428
MHRED+TK+EK	0.5988	0.00534
MHRED+TK+RL	0.6368	0.00420
KMD	**0.6731**	**0.00534**

4.4 Ablation Study on Major Components

There are mainly three major components in the proposed knowledge-aware multimodal dialogue framework. In order to better understand the contribution of each component, we carried out ablation study for detailed analysis.

For text response generation, Table 4 shows that all the three components have positive contributions to the performance. For instance, the BLEU score is increased by 10.3%, 15.3%, 22.6% for MHRED+TK, MHRED+TK+EK and MHRED+TK+RL respectively as compared to MHRED. Specifically, the taxonomy-based visual semantic model captures the structured knowledge such as the different relationships between product categories and attributes. It thus learns more informative representations for fashion products. For instance, our method manages to understand that the first shoe in Figure 7 is a formal shoe. Similarly, the memory network component stores extracted multimodal knowledge such as style tips. Therefore, the agent can generate responses not only based on the conversation context but also the external knowledge. Such external knowledge helps to boost the performance and also increase the unigram diversity of generated responses which is demonstrated by the third example in Figure 7. It is clearly evidenced that the further incorporation of the RL component produces the highest increase in the BLEU score. Such result is as expected since it fine-tunes the MHRED backbone network and directly optimize the BLEU score as rewards.

For image response generation, the three components also help to boost the performance as shown in Table 5. Instead of directly using the image features obtained from the FC6 layer of a VGGNet-16, MHRED+TK applies the features extracted via the taxonomy-based visual semantic model – the EITree model, in which more informative representations of fashion products are learned. Therefore, we observe a performance increase over the MHRED backbone method. In MHRED+TK+EK, similar to the text response generation, the external multimodal knowledge helps to introduce relevant information which is not available in the conversation context. Thus, it enables the agent to answer some questions that pure seq2seq models fail to answer, such as recommending matching items. For the RL component, it fine-tunes the MHRED backbone network and optimizes the similarity based rewards. A baseline method is leveraged to decrease learning variance. With such proper training, it also helps to generate better answers.

Table 5: Performance of ablation study of the KMD framework on image response (RQ2 & RQ3).

Method	R@1	R@2	R@3
MHRED	0.7980	0.8859	0.9345
MHRED+TK	0.8281	0.9141	0.9532
MHRED+TK+EK	0.8478	0.8889	0.8947
MHRED+TK+RL	0.8823	0.9387	0.9473
KMD	**0.9198**	**0.9552**	**0.9755**

5 SUMMARY

In this paper, we proposed a general knowledge-aware multimodal dialogue model named KMD. It was constructed around a taxonomy-based visual semantic learning model and introduced an attention mechanism to adaptively attend to the multimodal knowledge extracted and stored in memory network. To avoid the inconsistent dialogues and error accumulation problems, a deep reinforcement learning method was adapted to optimize multimodal dialogues in an end-to-end fashion. The proposed KMD model can be applied to domains like retails, travel, and entertainment etc. We evaluated it on the fashion dialogue application. Experimental results demonstrated the effectiveness of the proposed framework in integrating domain knowledge into the systems, leading to better performance as compared to the state-of-the-art approaches.

We built a demo for this work as in [21]. It is worth noting that current end-to-end models are still far from perfect. In order to build intelligent multimodal agents, the road ahead of us is still long. Nevertheless, we would nudge towards the goal gradually. There are several possible research directions that would be rather helpful. (1) In order to elevate user satisfaction, personalizing the conversation agent might be a good option, because each person has his/her own personality and preferences [16] which largely affect the responses he/she expects to get. (2) Due to the difficulties in collecting dialogue data in certain domains, to harvest knowledge in a specific domain and transfer to other domains would be useful. (3) In real life scenario, multimodal conversations might contain both task-oriented and non-task-oriented dialogues. It might also be natural to handle such dialogues simultaneously.

ACKNOWLEDGMENT

This research is part of NExT++ project, supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@Singapore Funding Initiative. This work is also supported in part by the project from the National Science Foundation of China under grant 61722204 and 61732007.

REFERENCES

[1] Sungjin Ahn, Heeyoul Choi, Tanel Pärnamaa, and Yoshua Bengio. 2016. A neural knowledge language model. *arXiv preprint arXiv:1608.00318* (2016).

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *ICCV*. 2425–2433.

[3] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. DBpedia: A nucleus for a web of open data. In *The semantic web*. 722–735.

[4] Dzmitry Bahdanau, Philemon Brakel, Kelvin Xu, Anirudh Goyal, Ryan Lowe, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. An actor-critic algorithm for sequence prediction. In *ICLR*.

[5] Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning end-to-end goal-oriented dialog. In *ICLR*.

[6] Antoine Bordes, Jason Weston, Ronan Collobert, Yoshua Bengio, et al. 2011. Learning Structured Embeddings of Knowledge Bases. In *AAAI*. 301–306.

[7] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A Survey on Dialogue Systems: Recent Advances and New Frontiers. *SIGKDD Explor. Newsl.* (2017), 25–35.

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *CVPR*. 1080–1089.

[9] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. GuessWhat?! Visual object discovery through multi-modal dialogue. In *CVPR*. 5503–5512.

[10] 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 *ACL*. 484–495.

[11] Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D Manning. 2017. Key-Value Retrieval Networks for Task-Oriented Dialogue. In *SIGDIAL*. 37–49.

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

[13] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *CVPR*. 6904–6913.

[14] Kelvin Guu, John Miller, and Percy Liang. 2015. Traversing Knowledge Graphs in Vector Space. In *EMNLP*. 318–327.

[15] Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *SIGDIAL*. 129–133.

[16] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[17] Sungjin Lee and Maxine Eskenazi. 2013. Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In *SIGDIAL*. 414–422.

[18] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *EMNLP*. 1192–1202.

[19] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Celikyilmaz. 2017. End-to-End Task-Completion Neural Dialogue Systems. In *IJCNLP*. 733–743.

[20] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-seng Chua. 2018. Interpretable Multimodal Fashion Retrieval for Fashion Products. In *MM*.

[21] Lizi Liao, You Zhou, Yunshan Ma, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware Multimodal Fashion Chatbot. In *MM*.

[22] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. [n. d.]. Optimization of image description metrics using policy gradient methods. In *ICCV*. 873–881.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. 375–383.

[24] Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *EMNLP*. 1400–1409.

[25] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* (1995), 39–41.

[26] Tom M Mitchell, William W Cohen, Estevam R Hruschka Jr, Partha Pratim Talukdar, Justin Betteridge, Andrew Carlson, Bhavana Dalvi Mishra, Matthew Gardner, Bryan Kisiel, Jayant Krishnamurthy, et al. 2015. Never Ending Learning.. In *AAAI*. 2302–2310.

[27] Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. Image-Grounded Conversations: Multimodal Context for Natural Question and Response Generation. In *IJCNLP*. 462–472.

[28] Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *ACL*. 1777–1788.

[29] Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. 2016. A review of relational machine learning for knowledge graphs. In *Proceedings of the IEEE*. 11–33.

[30] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *ICLR*.

[31] Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. In *AAAI*. 696–704.

[32] Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical Neural Network Generative Models for Movie Dialogues. *arXiv preprint arXiv:1507.04808* (2015).

[33] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models.. In *AAAI*. 3776–3784.

[34] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*.

[35] 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*. 196–205.

[36] Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*. 2765–2771.

[37] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. 2000. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*. 1057–1063.

[38] Jie Tang, Juanzi Li, Bangyong Liang, Xiaotong Huang, Yi Li, and Kehong Wang. 2006. Using Bayesian Decision for Ontology Mapping. *Journal of Web Semantics* (2006).

[39] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*. 3156–3164.

[40] TH Wen, D Vandyke, N Mrkšíc, M Gašíc, LM Rojas-Barahona, PH Su, S Ultes, and S Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *EACL*. 438–449.

[41] Jason Weston, Sumit Chopra, and Antoine Bordes. 2015. Memory Networks. In *ICLR*.

[42] Jason D Williams and Geoffrey Zweig. 2016. End-to-end LSTM-based dialog control optimized with supervised and reinforcement learning. *arXiv preprint arXiv:1606.01269* (2016).

[43] Ronald J Williams. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*. 5–32.

[44] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton van den Hengel. 2016. What value do explicit high level concepts have in vision to language problems?. In *CVPR*. 203–212.

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*. 2048–2057.

[46] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping.. In *AAAI*. 4618–4626.

[47] Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. 2015. Embedding entities and relations for learning and inference in knowledge bases. In *ICLR*.

[48] Wojciech Zaremba and Ilya Sutskever. 2015. Reinforcement learning neural turing machines-revised. *arXiv preprint arXiv:1505.00521* (2015).

[49] Hanwang Zhang, Zheng-Jun Zha, Yang Yang, Shuicheng Yan, Yue Gao, and Tat-Seng Chua. 2013. Attribute-augmented semantic hierarchy: towards bridging semantic gap and intention gap in image retrieval. In *MM*. 33–42.

[50] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Yin and yang: Balancing and answering binary visual questions. In *CVPR*. 5014–5022.

[51] Tiancheng Zhao and Maxine Eskenazi. 2016. Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning. In *SIGDIAL*. 1.

[52] Wenya Zhu, Kaixiang Mo, Yu Zhang, Zhangbin Zhu, Xuezheng Peng, and Qiang Yang. 2017. Flexible End-to-End Dialogue System for Knowledge Grounded Conversation. *arXiv preprint arXiv:1709.04264* (2017).