Multimodal Dialog System: Generating Responses via Adaptive Decoders

Liqiang Nie
Shandong University
nieliqiang@gmail.com

Wenjie Wang
Shandong University
wenjiewang96@gmail.com

Richang Hong
Hefei University of Technology
hongrc.hfut@gmail.com

Meng Wang
Hefei University of Technology
eric.mengwang@gmail.com

Qi Tian
Noahs Ark Lab, Huawei
tian.qi1@huawei.com

ABSTRACT

On the shoulders of textual dialog systems, the multimodal ones, recently have engaged increasing attention, especially in the retail domain. Despite the commercial value of multimodal dialog systems, they still suffer from the following challenges: 1) automatically generate the right responses in appropriate medium forms; 2) jointly consider the visual cues and the side information while selecting product images; and 3) guide the response generation with multi-faceted and heterogeneous knowledge. To address the aforementioned issues, we present a Multimodal diAloG system with adaptIve deCoders, MAGIC for short. In particular, MAGIC first judges the response type and the corresponding medium form via understanding the intention of the given multimodal context. Hereafter, it employs adaptive decoders to generate the desired responses: a simple recurrent neural network (RNN) is applied to generating general responses, then a knowledge-aware RNN decoder is designed to encode the multiform domain knowledge to enrich the response, and the multimodal response decoder incorporates an image recommendation model which jointly considers the textual attributes and the visual images via a neural model optimized by the max-margin loss. We comparatively justify MAGIC over a benchmark dataset. Experiment results demonstrate that MAGIC outperforms the existing methods and achieves the state-of-the-art performance.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**; **Natural language generation**;

KEYWORDS

Multimodal Dialog Systems; Multiform Knowledge-aware Decoder; Adaptive Decoders

* Corresponding author: Liqiang Nie (nieliqiang@gmail.com).

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350923

ACM Reference Format:
Liqiang Nie, Wenjie Wang, Richang Hong, Meng Wang, and Qi Tian. 2019. Multimodal Dialog System: Generating Responses via Adaptive Decoders. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350923

1 INTRODUCTION

With the development of big data and deep learning techniques, we have witnessed the rise of dialog systems in recent years. Generally speaking, dialog systems can be categorized into two groups: open-domain and task-oriented dialog systems. The former is able to chat with users on a wide range of topics without domain restrictions, like chit-chat; whereas the latter satisfies users' specific requests in certain vertical domains, such as accommodation booking. They both have shown promising commercial value in many fields, spanning from the companion chat (*e.g.*, XiaoIce[1]) to customer services (*e.g.*, Cortana[2] and Siri[3]). Despite their progress, most of the existing efforts purely focus on the textual conversation between users and chatbots, overlooking the important visual cues. As the old saying goes, "a picture is worth a thousand words", namely a picture can often vividly express the intentions. As displayed in Figure 1, the user describes his preferred sandals through a product picture, which remarkably facilitates the user to express requirements and enables the chatbot to understand the appearance of products clearly. Inspired by this, seamlessly integrating the visual images into the traditional textual dialog systems, the so-called multimodal dialog systems, deserves our attention.

In this paper, we work towards a task-oriented multimodal dialog system, which undoubtedly relies on the support of large-scale and domain-aware datasets. As a leading study, Saha *et al.* [30] released a multimodal dialog dataset (MMD) in the retail domain. Along with this MMD benchmark dataset, the authors presented two basic tasks, namely the textual response generation and the best image response selection, which are implemented by a multimodal hierarchical encoder-decoder (MHRED) model. On the basis of MHRED model, Liao *et al.* [21] incorporated the style tips into the neural model by a Memory Network [32] and adopted deep reinforcement learning to maximize the expected future reward. As a result, the knowledge-aware multimodal dialog system (KMD)

[1] https://www.msxiaobing.com.
[2] https://www.microsoft.com/en-us/cortana.
[3] https://www.apple.com/siri.

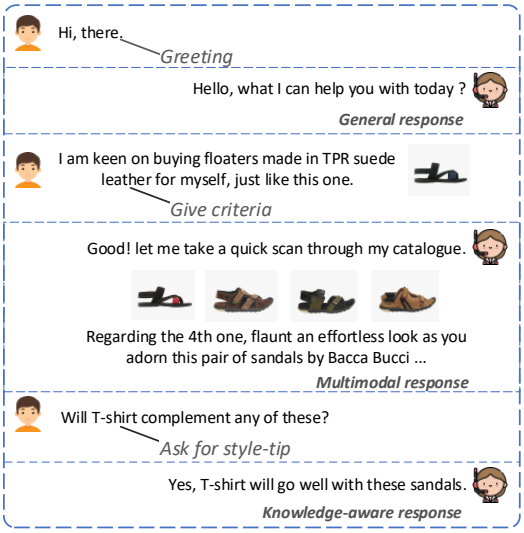

Figure 1: A multimodal dialog system between a shopper and a chatbot. The shopper expresses his requirements step by step as the dialog goes on. And the chatbot generates different responses according to the context.

achieves better performance as compared to MHRED over two basic tasks, respectively.

Although existing task-oriented multimodal dialog systems have shown promising performance, they still suffer from the following issues: 1) As illustrated in Figure 1, the responses from the chatbot express various information in different medium forms, ranging from greetings and visual demonstration to informative explanation, which are demonstrated in either texts or text-image combinations. Pioneer efforts treat text generation and image selection in the multimodal dialog systems as two separate tasks, and generate the responses by selectively assembling the texts and images manually. 2) The image selection task is essentially a product recommendation problem. Recommenders rank the products and return the top ones according to the user's preference conveyed in the context. Existing methods, such as MHRED and KMD, only consider visual images during the selection, totally ignoring the rich side information associated with the products, such as the price, material, size, and style. And 3) as shown in Figure 1, the conversations between the shopper (user) and the chatbot usually involve multiform knowledge in heterogeneous facets, including style tips, product attributes, and product popularity in celebrities. Nevertheless, the KMD method only incorporates the style tips into the neural model and the MHRED one does not explore any kinds of knowledge at all.

Indeed, it is tough to alleviate the aforementioned issues due to the following challenges: 1) Given the multimodal context, we have to determine what type of responses should be generated in advance and then present them in appropriate medium forms. 2) Jointly considering the product images and their side information to select the most relevant products is very necessary, however, it has been untapped adequately to date. And 3) the multi-faceted knowledge are heterogeneous. For example, style tips are usually organized as a graph, whose edges describe the relations among different products; while the product popularity in celebrities always appears as a popularity distribution histogram, and the product attributes are actually organized into a key-value table. How to encode multiform knowledge within a unified decoder is an unsolved problem.

To address the aforementioned challenges, in this work, we present a Multimodal diAloG system with adaptIve deCoders, MAGIC for short, as illustrated in Figure 2. To be more specific, our proposed MAGIC model first embeds the historical utterances via a multimodal context encoder. It then understands users' diverse intentions conveyed in the multimodal context by classifying them into 15 categories, such as greeting, giving criteria, and purchasing. According to our statistics over the MMD dataset, responses to these 15 kinds of intentions are in three variants without exception: general responses in texts, knowledge-enriched responses in texts, and the multimodal responses in the form of texts and images. In the light of this, MAGIC automatically judges the response type and its corresponding medium form by looking up our pre-defined tables with triplet entries (Intention Category, Response Types, Medium Forms). Hereafter, MAGIC employs the adaptive decoders to generate the desired response types, whereby the input of the decoders is the embedding of the historical utterances. In particular, 1) a simple recurrent neural network (RNN) is applied to generating general responses; while 2) a knowledge-aware RNN decoder embeds the multiform domain knowledge into a knowledge vector in a high-dimensional space via the Memory Network [32] and the Key-Value Memory Network [26], and then the knowledge vector is incorporated into a unified RNN decoder to produce more knowledge-enriched responses. And 3) the recommender model learns the product representations by jointly considering the textual attributes and the visual images via a neural model optimized by the max-margin loss. Ultimately, the recommender ranks the product candidates based on the similarity between the product representation and the embedding of the historical utterances. It is worth noting that since the multimodal responses are very complex and sophisticated, mixing general responses, domain knowledge, and visual illustration, they hence integrate the outputs of the simple RNN decoder, the knowledge-aware RNN decoder, and the recommender, simultaneously. Extensive experiments on the MMD dataset demonstrate the superior performance of MAGIC over the baselines. And we release our code and data[4] to facilitate the research in this field.

To sum up, the contributions of our work are threefold:

- In the family of multimodal dialog systems, we are the first to judge the response type and its medium form. It is achieved by an intention understanding component, which enables us to automatically generate the context-adaptive responses in appropriate medium forms.
- We incorporate a novel product recommender into the multimodal dialog systems, which jointly characterizes the visual and side information of products. It is also applicable to recommend multimodal items in other fields.
- We design a multiform knowledge-aware response decoder, encoding various forms of domain knowledge within a unified decoder.

[4]https://acmmultimedia.wixsite.com/magic.

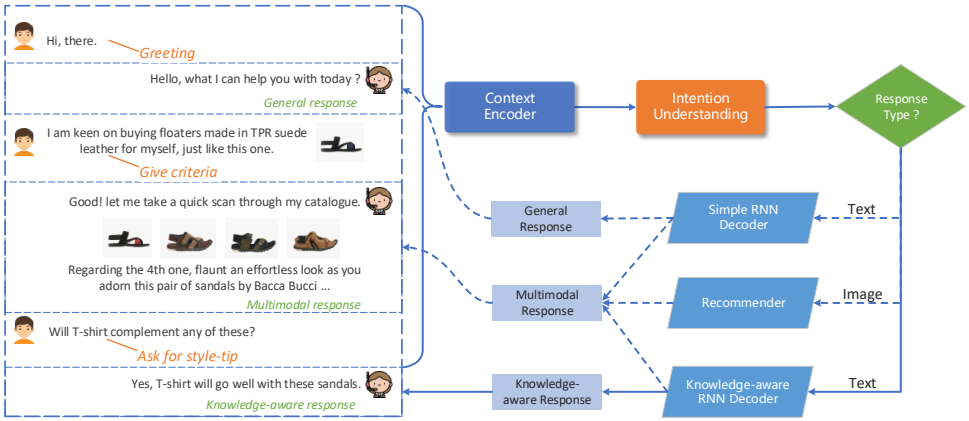

Figure 2: Schematic illustration of our proposed MAGIC model.

2 RELATED WORK

This work is closely related to the textual dialog systems, multimodal dialog systems, and conversational recommender systems.

2.1 Textual Dialog Systems

In recent years, great efforts have been dedicated to developing textual dialog systems, including open-domain and task-oriented dialog systems. The former is usually implemented by template-, retrieval- or generation-based methods. To be more specific, the template-based methods leverage predefined rules and templates to reply to users [36]. They are, however, prone to frequent errors as the results are not always the desired ones. As to the retrieval-based models [38, 41–43], they select proper responses for the current conversation from a repository via response selection algorithms, benefiting from informative and fluent responses. By contrast, the generation-based ones [17, 35] produce more proper responses that have never appeared in the corpus by automatically predicting the generation probability of each word in the response based on the historical context. For example, Serban *et al.* [31] extended the recurrent hierarchical encoder-decoder (HRED) neural network to generate responses. Attention mechanism [2] is also incorporated into the generation-based methods to improve the performance [25, 44]. In addition, leveraging the external knowledge with deep neural models to retrieve or generate responses has been a promising research direction [40, 43]. For example, Yang *et al.* [43] proposed a deep neural matching network, which leverages the external knowledge for response ranking in information-seeking conversation systems. Another example is illustrated in [12], whereby a knowledge-ground generation-based model employs the external non-conversational data to produce more informative responses.

Different from open-domain dialog systems that chat with users without domain restrictions, task-oriented [4, 37] ones focus on assisting users to accomplish specific tasks in vertical domains. Traditional task-oriented dialog systems follow a typical pipeline. They first utilize a natural language understanding component to classify the users' intentions. And then the dialog state tracker tracks the users' requirements and fills the predefined slots. It is followed by the policy network that decides what action to make at the next step. Ultimately, the natural language generation component gives the response through the predefined templates or some generation-based models. Though this pipeline has performed well in certain tasks, it still suffers from the complex framework and the error propagation [16, 21, 27]. To alleviate such problems, several end-to-end task-oriented dialog systems [3, 19, 29, 37], integrating the advantages of supervised learning and deep reinforcement learning methods, are proposed. Besides, they [14, 29, 37] also incorporate domain knowledge into dialog systems by issuing a symbolic query in a structured knowledge base (KB) [19] or estimating the "soft" posterior distribution over the KB [9]. However, all these task-oriented methods only focus on the textual dialog and totally ignore the visual cues.

2.2 Multimodal Dialog Systems

Although the textual utterances convey valuable information, they are limited in describing the visual properties in many cases [6, 21–23]. With the development of many industrial domains, such as e-commerce retail and travel, the demand for multimodal dialog systems is increasing rapidly. In fact, the release of MMD dataset [30] has promoted the development of multimodal dialog systems. Saha *et al.* [30] also proposed two basic tasks along with the dataset: the textual response generation and the best image response selection. Meanwhile, they developed MHRED model regarding these two tasks. Later, Liao *et al.* [21] presented the KMD model, extracting the visual representation using Exclusive&Independent tree [20], incorporating style tips provided in [30] into the MHRED, and leveraging deep reinforcement learning to boost the performance. However, MHRED and KMD consider the product recommendation as an image selection task, which only leverages visual features and overlooks the side information of products. Moreover, they neglect other forms of domain knowledge required in the retail domain, such as product attributes and popularity.

In addition to multimodal dialog systems, visual question answering (VQA) [1, 13] and visual dialog [7, 8] are also somehow related to our work. VQA is an emerging problem in computer vision and natural language processing that has engaged a large amount of interest from various communities of the deep learning, computer vision, and natural language processing. In VQA, an algorithm needs to answer text-based questions about images. Since the release of the first VQA dataset [1] in 2014, additional datasets [24] have been released and many algorithms [13] have been proposed. Different from VQA that only consists of a single-round natural language interaction, visual dialog [7] requires the agent to conduct a meaningful dialog with humans in natural language regarding the visual content. Specifically, given an image,

a dialog history, and a question about the image, the agent has to understand the question about the image, infer the context from history, and ultimately answer the question in natural language. By comparison, multimodal dialog systems encode the multimodal dialog history with many images, and integrate the utterances in multiple medium forms to interact with humans at every turn.

2.3 Conversational Recommender System

Research on conversational recommender systems is comparatively sparse. Considering the success of recommender systems in helping users find the preferred items, integrating the strength of recommendation methods into the dialog systems to serve users has much commercial value. In the former efforts [5, 18, 33], the chatbots usually continue to propose predefined questions in the conversation until they collect enough information to make a recommendation. Many slots are defined manually in most existing methods [5, 33] to characterize the products, such as prize and location. Thereafter, the methods extract the values of these slots from the users' feedback. For example, Sun *et al.* [33] selected five item attributes as slots for the food in the Yelp challenge recommendation dataset[5] and generated dialog scripts to train the proposed Conversational Recommender System. Whereas, all these conversational recommender systems merely focus on textual dialog and product information, totally ignoring the significance of the massive visual cues in the recommendation.

3 OUR PROPOSED MAGIC MODEL

This section details our proposed MAGIC model, which can embed the multimodal context, understand users' intention, and adaptively generate the responses for users in appropriate medium forms. As shown in Figure 2, given the multimodal context $\{u_0, u_1, ..., u_n\}$, where each utterance u_i consists of several textual sentences, or integrates textual sentences and visual images, MAGIC first leverages a context encoder to embed the multimodal context into a context vector $\mathbf{c}$, and then understands the user's intention via classification. To generate multiform responses for various user intentions, we leverage three parallel components (*i.e.*, Simple RNN Decoder, Knowledge-aware RNN Decoder, and Recommender) to produce three types of responses: general responses, knowledge-aware responses, and multimodal responses. It is worth noting that general responses refer to the highly-frequent responses in the conversations, which smooth the conversation without any practical information, for example, *"What can I help you with today?"* As to the knowledge-aware responses, they are the responses incorporated with multiform domain knowledge to satisfy users' specific demands, such as responses to the question *"Will T-shirt match any of these sandals?"* In addition, multimodal responses comprise a general response in courtesy, the visual images of recommended products, and a knowledge-aware response to introduce the product attributes.

3.1 Context Encoder

To encode the multimodal context $\{u_0, u_1, ..., u_n\}$, we design a deep hierarchical neural model as illustrated in Figure 3. To be more specific, at the low level, a RNN is used to encode the

[5]https://www.yelp.com/academic_dataset.

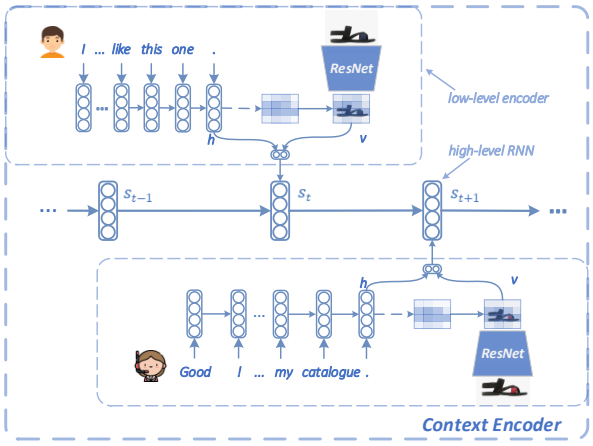

Figure 3: Context Encoder by a hierarchical neural model.

textual utterance word by word and a ResNet [15] augmented with soft visual attention [39] is employed to extract the visual feature of products. As to the high level, the textual features or the concatenation of textual and visual features is fed into the sentence-level RNN in every utterance. Notably, the utterance could be either textual or multimodal, therefore, the high-level RNN sometimes only takes textual features as inputs.

Specifically, the textual utterance is encoded by a low-level RNN, and the final hidden state $\mathbf{h_t}$, embedding the information in the whole utterance, is treated as the representation of the input textual utterance. As for the visual feature extraction, considering that the users' visual attention on image regions differs, we leverage the soft visual attention to attentively extract visual features, which has shown its effectiveness in several cross-modal tasks, such as image captioning [39]. In particular, the feature map of ResNet is divided into L regions; meanwhile, the final hidden state $\mathbf{h_t}$ of the low-level RNN describing the users' current requirements for products, such as the color or size preference, is used to decide what the model should pay attention to. Formally, the attention weight α_i^t for each region at the time step t is calculated by,

$$e_i^t = f_{att}(\mathbf{h_t}, \mathbf{v_i^t}), \quad \alpha_i^t = \frac{exp(e_i^t)}{\sum_{j=1}^{L} exp(e_j^t)}, \tag{1}$$

where $\mathbf{h_t}$ is the final hidden state of the low-level RNN at the time step t, $\mathbf{v_i^t}$ refers to the visual feature of region i at the step t, the function f_{att} is implemented by a Multi-Layer Perceptron (MLP) in this work and $\boldsymbol{\alpha}$ denotes the attention weighs over the visual features of L regions. Thereafter, the visual feature $\mathbf{v_t}$ at the step t can be obtained as follows,

$$\mathbf{v_t} = \sum_{i=1}^{L} \alpha_i^t \mathbf{v_i^t}. \tag{2}$$

Note that for multimodal utterances, the visual feature $\mathbf{v_t}$ and the textual one $\mathbf{h_t}$ are concatenated first, and then fed into the high-level RNN. And if an utterance consists of several images, the images will be unrolled into a sequence of visual utterances and fed into the high-level RNN with textual features one by one. As for utterances of mere texts, only textual features are calculated at the high level. Therefore, from a high-level perspective, the RNN iteratively processes the utterances, characterizes the user-related

Table 1: The categories of users' intentions and their corresponding responses in different medium forms.

Id	Intention Category	Response Type	Medium Form	Id	Intention Category	Response Type	Medium Form
1	greeting	general response	Text	9	filter results	multimodal response	Text+Image
2	give self-info	general response	Text	10	ask for style tips	knowledge-aware response	Text
3	give criteria	multimodal response	Text+Image	11	ask for attributes	knowledge-aware response	Text
4	like specific items, show more	multimodal response	Text+Image	12	ask for popularity in celebrities	knowledge-aware response	Text
5	dislike specific items, show more	multimodal response	Text+Image	13	switch back to former items	general response	Text
6	show orientations of products	multimodal response	Text+Image	14	buy	general response	Text
7	show similar items	multimodal response	Text+Image	15	exit	general response	Text
8	sort results	multimodal response	Text+Image				

information in the dialog step by step, and ultimately outputs the final hidden state as the context vector $\mathbf{c}$. Thereafter, the context vector $\mathbf{c}$ will be fed into the intention understanding, the recommender and two RNN decoders as the multimodal context representation.

3.2 Intention Understanding

Given the context vector, this component aims to understand the users' intention and thereafter to decide the corresponding decoder for response generation. In particular, the intention of users in the MMD dataset can be classified into 15 categories as summarized in Table 1. Here, we leverage the MLP network to predict the probability distribution over the 15 intentions based on the context vector $\mathbf{c}$ generated by the context encoder. Besides, a cross-entropy loss is applied to optimizing the network, and ultimately the model achieves superior accuracy up to 98.9%.

As aforementioned analyses over the MMD dataset, we noticed that the responses to these 15 intentions can be exclusively divided into three types, namely general response, knowledge-aware response, and multimodal response. Thereinto, the multimodal response is expressed in both texts and images; whereas the others are only in texts. Inspired by this, we design a lookup table, containing many triplets in the format of $(intention\ category, response\ type, medium\ form)$. Once given the intention category of the multimodal context, our model MAGIC can select the right decoder to generate the corresponding responses in appropriate medium forms.

3.3 Simple RNN Decoder

The objective of the simple RNN decoder is to generate general responses based on the context vector $\mathbf{c}$, which are common in the MMD dataset and do not need any domain knowledge. As utilizing knowledge-aware RNN decoder to produce general responses may bring additional computing burden, incorporate noise and mislead the optimization of the model, we introduce the simple RNN to generate general responses separately. The hidden state $\mathbf{h_0}$ of the simple RNN is initialized by the context vector $\mathbf{c}$, and then updated iteratively by the following equation,

$$\mathbf{h_t} = f(\mathbf{h_{t-1}}, \mathbf{e_{w_{t-1}}}), \tag{3}$$

where $\mathbf{h_t}$ refers to the hidden state at the step t, and $\mathbf{e_{w_{t-1}}}$ denotes the embedding of the token w_{t-1} in the target response. Thereafter, the model linearly projects the hidden state at each step to a one-dimensional vector in the vocabulary size and outputs the probability distribution of every token. Eventually, the cross-entropy error function is applied to maximizing the prediction probability of sequential tokens in the target response.

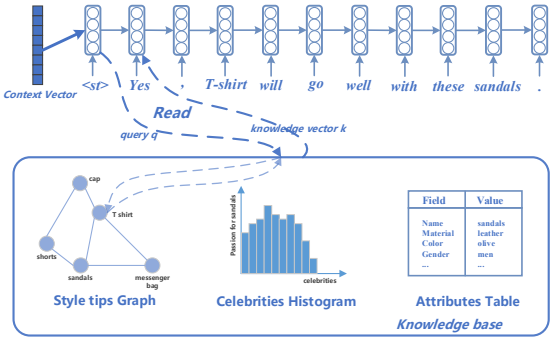

Figure 4: Multiform knowledge-aware RNN Decoder.

3.4 Knowledge-aware RNN Decoder

In the MMD dataset, the shoppers tend to express their requirements and collect enough product information before the final purchase. The product information usually involves three kinds of domain knowledge, namely style tips, product attributes, and the product popularity among celebrities. To be more specific, 1) style tips describe the matching status between different clothes, such as neckties going well with white shirts; 2) product attributes are organized in a key-value table to record the common attributes of products, such as prize, brand, and material; and 3) as to the product popularity among celebrities, it presents the preference distribution of celebrities over all kinds of products. For example, some celebrities favor black trousers instead of blue ones. Based upon the intention understanding results, as summarized in Table 1 (ID 10, 11 and 12), MAGIC can easily determine what kind of domain knowledge to incorporate. Specifically, if the user's intention is to query style tips, product attributes or product popularity in Table 1, MAGIC will embed the corresponding domain knowledge into a knowledge vector and incorporate it into the RNN decoder.

Formally, the knowledge-aware RNN is initialized by the context vector $\mathbf{c}$ and updated as follows,

$$\mathbf{s_t} = f(\mathbf{s_{t-1}}, [\mathbf{e_{w_{t-1}}}, \mathbf{a}, \mathbf{k}]), \tag{4}$$

where $\mathbf{s_t}$ refers to the hidden state of the knowledge-aware RNN decoder at the step t, and $[\mathbf{e_{w_{t-1}}}, \mathbf{a}, \mathbf{k}]$ is the concatenation of the embedding of the token w_{t-1} in the response, the attentive context vector $\mathbf{a}$, and the knowledge vector $\mathbf{k}$. Specifically, due to the close correlation between the last contextual sentence and the target response, we introduce the context vector $\mathbf{a}$ to attentively combine the hidden states of the last contextual sentence via the attention mechanism. In particular, the hidden state $\mathbf{s_{t-1}}$ is utilized to calculate the attention weights by inner product, and then the weighted hidden states of the last contextual sentence

are linearly added together to get the attentive context vector $\mathbf{a}$. Besides, the knowledge vector $\mathbf{k}$ is acquired from the multiform knowledge base by leveraging the hidden state at the previous step as the query $\mathbf{q}$. Notably, the knowledge vector $\mathbf{k}$ at the first step is acquired by using the context vector $\mathbf{c}$ as the query $\mathbf{q}$ and the first word fed into the RNN decoder is a special token <st>. Given the query $\mathbf{q}$, the following subsections will demonstrate how to embed three kinds of knowledge into the same high-dimensional space, respectively.

3.4.1 Incorporation of Style Tips. The style tips in the retail domain naturally appear as an undirected graph and the edge between two kinds of products implies that one goes well with the other. Therefore, we can describe the graph with pairwise entries, such as $(T\text{-}shirts,\ sandals)$, and then we incorporate the pairwise entries into the RNN decoder by a Memory Network [32]. In particular, we first embed each item of the pair into a vector and thereafter concatenate them to obtain a knowledge entry $\mathbf{e}$. Finally, all these knowledge entries are stored in the single-layer Memory Network. Given the query $\mathbf{q}$, the knowledge vector $\mathbf{k}$ is computed by,

$$\begin{cases} \mathbf{m_i} = \mathbf{A}\mathbf{e_i}, \\ \mathbf{o_i} = \mathbf{B}\mathbf{e_i}, \\ p_i = \dfrac{exp(\mathbf{q}^\mathrm{T}\mathbf{m_i})}{\sum_{j=1}^{N} exp(\mathbf{q}^\mathrm{T}\mathbf{m_j})}, \\ \mathbf{k} = \sum_{i=1}^{N} p_i \mathbf{o_i}, \end{cases} \tag{5}$$

where $\mathbf{e_i}$ denotes the knowledge entry and N is the number of knowledge entries. In addition, $\mathbf{A}$ and $\mathbf{B}$ are the embedding matrices in the Memory Network, which convert the input $\mathbf{e_i}$ into the memory vector $\mathbf{m_i}$ and the output vector $\mathbf{o_i}$, respectively.

3.4.2 Incorporation of Product Attributes. As for the product attributes, we apply the Key-Value Memory Network to acquire the knowledge vector $\mathbf{k}$ since the attributes are always presented as key-value pairs $\{(k_1, v_1), (k_2, v_2), ..., (k_M, v_M)\}$, such as $(material, leather)$. Formally, the knowledge vector $\mathbf{k}$ can be calculated on the basis of query $\mathbf{q}$ by the following equations,

$$p_i = \frac{exp(\mathbf{q}^\mathrm{T}\mathbf{k_i})}{\sum_{j=1}^{M} exp(\mathbf{q}^\mathrm{T}\mathbf{k_j})}, \quad \mathbf{k} = \sum_{i=1}^{N} p_i \mathbf{v_i}, \tag{6}$$

where $\mathbf{k_i}$ and $\mathbf{v_i}$ are the embeddings of the key and value of the i-th attribute, respectively. Meanwhile, M is the number of attributes.

3.4.3 Incorporation of Product Popularity. Supposing that there are N_c celebrities and N_p kinds of products included in the given data corpus, the product popularity among the celebrities can be expressed as a matrix $\mathbf{P} \in \mathbb{R}^{N_c \times N_p}$, where each row denotes a passion distribution of one celebrity over the N_p kinds of products. We treat the passion distribution of one celebrity as a knowledge entry $\mathbf{e}$ and store N_c knowledge entries in the Memory Network. Thereafter, the acquisition of the knowledge vector $\mathbf{k}$ is similar to the incorporation of style tips.

3.5 Recommender

Given the context vector $\mathbf{c}$, N_{pos} positive products, and N_{neg} negative ones, the recommender will rank the product candidates based on the similarity between the context vector $\mathbf{c}$ and the product representation $\mathbf{y}$. Different from the existing methods that rank product images by simply considering the visual features, we jointly incorporate both the visual features and the side information into the recommender. As shown in Figure 5, the product attributes and images are both fed into the product encoder to learn the product representation $\mathbf{y}$. In particular, for each product image, we first arrange its keys in alphabetical order, and then represent each pair by concatenating its key and value embeddings into one vector. Following that, we feed the ordered pairs into a RNN model step by step. The hidden state of the last step is regarded as the representation of textual attributes; whereas the visual representation is extracted by the pre-trained ResNet. Eventually, the textual and visual representations are concatenated and then linearly projected into the same high-dimensional space with the context vector $\mathbf{c}$. A max-margin loss is adopted to optimize the model for better recommendation, formulated as,

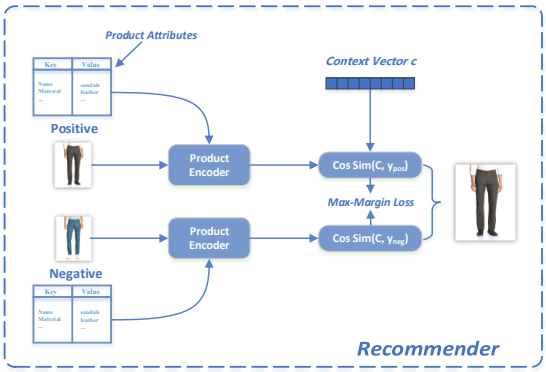

Figure 5: Illustration of the proposed recommender.

$$\ell = max(0, 1 - Sim(\mathbf{c}, \mathbf{y_{pos}}) + Sim(\mathbf{c}, \mathbf{y_{neg}})), \tag{7}$$

where $\mathbf{y_{pos}}$ and $\mathbf{y_{neg}}$ refer to the representations of positive and negative products, respectively, and the function $Sim(\mathbf{x}, \mathbf{y})$ denotes the cosine similarity between $\mathbf{x}$ and $\mathbf{y}$. When training the model, we minimize the max-margin loss ℓ to optimize the parameters; whereas at the test period, the recommender ranks the product candidates based on the similarity between the context vector $\mathbf{c}$ and the product representation $\mathbf{y}$.

4 EXPERIMENTS

4.1 Dataset

To build task-oriented multimodal dialog systems, we utilized the MMD dataset in the retail domain to train and evaluate our proposed model MAGIC. The MMD dataset consists of over 150k conversations between users and chatbots, and each conversation comprises approximately 40 utterances. Among them, every user's utterance in the conversation is labeled with one of the 15 intention categories. And over 1 million fashion products with a variety of domain knowledge were crawled from several well-known online retailing websites, such as Amazon[6] and Jabong[7]. Different from MHRED and KMD that utilize the fixed visual features of products obtained from the FC6 layer of the VGGNet-16 in [30], we

[6]https://www.amazon.com/.
[7]https://www.jabong.com/.

crawled the original pictures of these products from the websites to facilitate the extraction of visual features. Similar to MHRED and KMD, we treated every utterance of chatbots in the conversations as a target response and its former utterances as the context. Apart from that, we classified them into three types of responses to train the corresponding decoders separately. More details of the MMD dataset can be found in [30].

4.2 Experimental Settings

4.2.1 Hyper parameters. Following the former studies [21, 30], we utilized two-turn utterances prior to the target response as the context and the vocabulary size was set as 26,422. In the context encoder and response decoders, the RNN models are both implemented by the Gate Recurrent Units with 512 cells. In addition, the length of the knowledge vector $\mathbf{k}$ is 512 and the margin in the max-margin loss of the recommender is 1. Besides, the numbers of positive and negative products in recommendation are 1 and 4, respectively. We used Adam [10] to optimize the whole neural model and the learning rate was initialized as 0.0001.

4.2.2 Evaluation Metrics. To compare with the existing methods, we evaluated the performance of MAGIC over two basic tasks separately. For the task of textual response generation, we integrated the simple RNN decoder and the knowledge-aware RNN decoder to produce all textual responses. And we utilized Bleu-N [28], and Nist [11] to measure the similarity between the predicted and target responses. As the length of 20.07% target responses in the MMD dataset is less than 4, such as *"Yes!"* and *"That's right!"*, we calculated Bleu-N by varying N from 1 to 4. In particular, higher Bleu scores indicate that more n-gram overlaps exist between the predicted and target responses. And based on Bleu, Nist considers the weights of n-grams dynamically. The rarer a n-gram is, the more weight it will be given. As to the best image selection, we judged it by the *Recall@-m* metric similar to [30] and [21], where *m* is varied from 1 to 3. And the selection is correct only if the positive product is ranked in the top-m ones.

4.2.3 Baselines. To justify the performance of MAGIC, we compared MAGIC with several representative methods: Seq2seq [34], HRED [31], MHRED [30], Attention-based MHRED (AMHRED) [30], and KMD [21]. In particular, 1) Seq2seq is a classic encoder-decoder framework and achieves superior performance in many natural language processing tasks. 2) HRED is the most representative method in text-based multi-turn dialog systems. 3) MHRED is the first work on multimodal task-oriented dialog systems in the retail domain. 4) AMHRED is proposed along with MHRED, which incorporates the attention mechanism into MHRED at the sentence level. And 5) KMD is the state-of-the-art method in multimodal task-oriented dialog systems.

4.3 Objective Evaluation

4.3.1 Evaluating the best image selection. Table 2 displays the performance comparison with respect to Recall@m on the best image selection. From Table 2, we have the following findings: 1) MAGIC outperforms all the baselines in this task. Specifically, the recall scores of MAGIC approach 100%. According to our analyses, it is probably because: a) MAGIC extracts visual features from the original product images by ResNet while the baselines leverage the fixed visual features provided by [30]. And b) the rich side information of products is incorporated into our recommendation, which provides abundant product attributes for MAGIC to distinguish the positive product from the negative ones. And 2) the multimodal methods surpass the text-only methods. Indeed, text-only methods totally ignore the visual features in the multimodal context and only calculate the similarity between the textual features of the context and the visual features of products. Therefore, the poor performance of text-only methods verifies that the similarity of visual features significantly matters in the selection of products.

Table 2: Performance comparison between our proposed MAGIC model and baselines on the best image selection.

Methods		Recall@1	Recall@2	Recall@3
Text-only	Seq2seq	0.5926	0.7395	0.8401
	HRED	0.4600	0.6400	0.7500
Multimodal	MHRED	0.7200	0.8600	0.9200
	AMHRED	0.7980	0.8859	0.9345
	KMD	0.9198	0.9552	0.9755
	MAGIC	**0.9813**	**0.9927**	**0.9965**

Table 3: Performance comparison between the baselines and MAGIC on textual response generation.

Methods		Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist
Text-only	Seq2seq	35.39	28.15	23.81	20.65	3.3261
	HRED	35.44	26.09	20.81	17.27	3.1007
Multimodal	MHRED	32.60	25.14	23.21	20.52	3.0901
	AMHRED	33.56	28.74	25.23	21.68	2.4600
	MAGIC	**50.71**	**39.57**	**33.15**	**28.57**	**4.2135**

4.3.2 Evaluating the textual response generation. The performance of the baselines and MAGIC on the textual response generation is summarized in Table 3. By contrast, we have the following observations: 1) MAGIC surpasses the baselines regarding the Bleu and Nist scores, demonstrating that the adaptive decoders generate more informative and meaningful responses by incorporating multiform domain knowledge dynamically. 2) The Bleu-1 score of MAGIC is relatively high. By analyzing the generated responses, we found that MAGIC produces more accurate short responses (*e.g.*, *"Yes"* and *"No"*) for the knowledge-aware queries (*e.g.*, *"Does Cel_28 (celebrity ID) like this T-shirt?"*). This verifies that incorporating the abundant domain knowledge into the decoder is crucial to the generation of knowledge-aware responses. And 3) the performance of Seq2seq and HRED is comparable with MHRED, indicating that the generation of textual responses depends more on the textual features in the multimodal context. In addition, by comparing MHRED and AMHRED with MAGIC, we can conclude that the incorporation of domain knowledge is extremely crucial to the improvement of their performance.

4.4 Subjective Evaluation

Considering that some responses different from the target ones may also make sense in some cases, we conducted subjective comparison between the baselines and MAGIC. We first sampled 200 multimodal contexts randomly from the testing set, and then compared their responses generated by MAGIC with each baseline separately. In this way, we acquired 800 response pairs in total, each containing a response from MAGIC and the other

Table 4: Human evaluation over the responses of MAGIC and other baselines regarding four evaluation factors.

	Fluency				Relevance			
Opponent	Win	Loss	Tie	Kappa	Win	Loss	Tie	Kappa
MAGIC vs. Seq2seq	**16.9%**	13.8%	69.3%	0.46	**37.8%**	7.6%	54.7%	0.43
MAGIC vs. HRED	9.3%	**11.1%**	79.6%	0.51	**29.8%**	8.9%	61.3%	0.37
MAGIC vs. MHRED	**39.6%**	1.3%	59.1%	0.68	**35.6%**	5.3%	59.1%	0.58
MAGIC vs. AMHRED	**81.3%**	2.7%	16.0%	0.60	**74.7%**	2.2%	23.1%	0.56
	Logical Consistency				**Informativeness**			
Opponent	Win	Loss	Tie	Kappa	Win	Loss	Tie	Kappa
MAGIC vs. Seq2seq	**50.7%**	8.4%	40.9%	0.65	**20.0%**	16.9%	63.1%	0.49
MAGIC vs. HRED	**40.9%**	8.0%	51.1%	0.40	12.0%	**25.3%**	62.7%	0.55
MAGIC vs. AMHRED	**53.3%**	5.3%	41.3%	0.61	**48.0%**	2.2%	49.8%	0.65
MAGIC vs. AMHRED	**78.2%**	1.3%	20.4%	0.74	**80.0%**	2.7%	17.3%	0.67

from one of the four baselines. The responses in the pairs were randomly shuffled and three annotators were invited to judge which response is better in the context. If two responses are both meaningful or inappropriate, the comparison of this pair is treated as "tie". The annotators judged the response pairs based on the multimodal context from four aspects: fluency, relevance, informativeness, and logical consistency. Ultimately, we averaged the results of three annotators and reported them in Table 4. The kappa scores indicate a moderate agreement among the annotators. From Table 4, we found that: 1) MAGIC performs well in most of the comparisons, demonstrating that MAGIC is capable of producing more meaningful responses in the textual response generation. 2) HRED generates more informative and fluent responses than MAGIC does but its relevance and logical consistency are poor. By analyzing the specific cases, we found that although the responses of HRED are usually long and fluent, many of them are nonsense. It is partly because that it only considers the dependence among textual features and neglects some key information for lacking of images and knowledge. And 3) MAGIC significantly outperforms MHRED and AMHRED due to its advantage of leveraging knowledge-aware adaptive decoders.

4.5 Discussion

4.5.1 Case Study. Two representative samples are provided in Figure 6. The part context is omitted due to the limited space. From Figure 6, we can observe that the user asks about product attributes in the first case and seeks for product popularity in the second one. Moreover, MAGIC generates more accurate answers than MHRED does due to the incorporation of product attributes and popularity. Indeed, it is unlikely to generate accurate knowledge-aware responses without related domain knowledge. Only utilizing the similarity among the training samples is far from enough. The samples in Figure 6 intuitively explain the importance of incorporating multiform knowledge and the reason why MAGIC produces higher Bleu and Nist scores.

4.5.2 Model Ablation. To examine the effectiveness of adaptive decoders, we conducted the ablation test on MAGIC. We eliminated the incorporation of multiform knowledge into the decoder and leveraged a simple RNN decoder to produce the general and knowledge-aware responses. Table 5 presents the comparison between MAGIC with and without knowledge-aware RNN decoder. From Table 5, we can observe that the performance drops significantly if the multiform knowledge is removed from MAGIC, demonstrating the effectiveness of incorporating multiform domain knowledge into the textual response generation.

Table 5: The ablation test on knowledge-aware decoder.

Methods	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Nist
No Knowledge	35.90	23.79	17.34	13.33	2.5495
MAGIC	**50.71**	**39.57**	**33.15**	**28.57**	**4.2135**

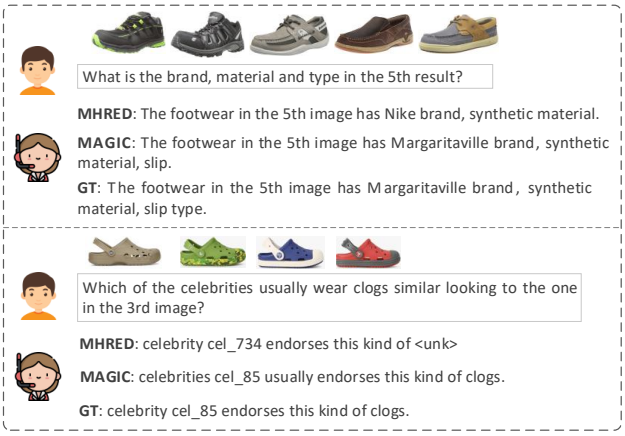

Figure 6: Case Study. "GT" denotes the ground truth response.

5 CONCLUSION AND FUTURE WORK

In this work, we present a multimodal task-oriented dialog system with adaptive decoders to generate general responses, knowledge-aware responses, and multimodal responses dynamically based on various user intentions. In particular, the proposed model first understands the users' intentions based on the multimodal context, and then leverages three parallel decoders, namely simple RNN decoder, knowledge-aware RNN decoder, and recommender, to generate responses in different medium forms. Extensive experiments exhibit the superiority of our proposed model in two basic tasks over the existing methods, demonstrating the effectiveness of three adaptive decoders.

Since the proposed model has performed well in two basic tasks, there is still a long way before applying it in practice. Firstly, the number of products in the retail domain is huge while the products for retrieval in the MMD dataset are limited. Secondly, the dialogs in the training dataset are restricted in the retail domain, thus extending the application domain of the proposed multimodal dialog system is a tough issue. In the future, we will further explore these issues and improve the practicability of the proposed model.

ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.:61772310, No.:61702300, No.:61702302, No.:61802231, and No.:U1836216; the Project of Thousand Youth Talents 2016; the Shandong Provincial Natural Science and Foundation, No.:ZR2019JQ23, No.:ZR2019QF001; the Future Talents Research Funds of Shandong University, No.:2018WLJH 63.

REFERENCES

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2425–2433.

[2] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

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

[4] Zheqian Chen, Rongqin Yang, Zhou Zhao, Deng Cai, and Xiaofei He. 2018. Dialogue act recognition via crf-attentive structured network. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 225–234.

[5] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the International Conference on knowledge Discovery and Data Mining*. ACM, 815–824.

[6] Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User Attention-guided Multimodal Dialog Systems. In *The 42nd International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 445–454.

[7] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1080–1089.

[8] Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5503–5512.

[9] 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*. ACL, 484–495.

[10] Jimmy Lei Ba. Diederik P. Kingma. 2015. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

[11] George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*. Morgan Kaufmann Publishers Inc., 138–145.

[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 *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI Press, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6325–6334.

[14] Sangdo Han, Jeesoo Bang, Seonghan Ryu, and Gary Geunbae Lee. 2015. Exploiting knowledge base to generate responses for natural language dialog listening agents. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. SIGDIAL, 129–133.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 770–778.

[16] Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 1437–1447.

[17] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technologies*. ACL, 110–119.

[18] Raymond Li, Samira Ebrahimi Kahou, Hannes Schulz, Vincent Michalski, Laurent Charlin, and Chris Pal. 2018. Towards Deep Conversational Recommendations. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 9748–9758.

[19] Xiujun Li, Yun-Nung Chen, Lihong Li, Jianfeng Gao, and Asli Çelikyilmaz. 2017. End-to-End Task-Completion Neural Dialogue Systems. In *Proceedings of the 8th International Joint Conference on Natural Language Processing*. AFNLP, 733–743.

[20] Lizi Liao, Xiangnan He, Bo Zhao, Chong-Wah Ngo, and Tat-Seng Chua. 2018. Interpretable multimodal retrieval for fashion products. In *Proceedings of the ACM Multimedia Conference on Multimedia Conference*. ACM, 1571–1579.

[21] Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware Multimodal Dialogue Systems. In *Proceedings of the ACM Multimedia Conference on Multimedia Conference*. ACM, 801–809.

[22] Meng Liu, Liqiang Nie, Meng Wang, and Baoquan Chen. 2017. Towards Micro-video Understanding by Joint Sequential-Sparse Modeling. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 970–978.

[23] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 843–851.

[24] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. [n. d.]. Hierarchical question-image co-attention for visual question answering. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press.

[25] Hongyuan Mei, Mohit Bansal, and Matthew R. Walter. 2017. Coherent Dialogue with Attention-Based Language Models. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence*. AAAI Press, 3252–3258.

[26] Alexander H. Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-Value Memory Networks for Directly Reading Documents. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. ACL, 1400–1409.

[27] Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. 2016. Learning from multiple social networks. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 8, 2 (2016), 1–118.

[28] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of Annual Meeting of the Association for Computational Linguistics*. ACL, 311–318.

[29] 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. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*. ACL, 438–449.

[30] Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards Building Large Scale Multimodal Domain-Aware Conversation Systems. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. AAAI Press.

[31] 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 *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. AAAI Press, 3776–3784.

[32] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 2440–2448.

[33] Yueming Sun and Yi Zhang. 2018. Conversational Recommender System. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 235–244.

[34] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. In *Proceedings of the Neural Information Processing Systems Conference*. MIT Press, 3104–3112.

[35] Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat more: Deepening and widening the chatting topic via a deep model. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 255–264.

[36] Jason D. Williams, Antoine Raux, Deepak Ramachandran, and Alan W. Blac. 2013. The dialog state tracking challenge. In *Proceedings of the SIGDIAL Conference on Discourse and Dialogue*. SIGDIAL, 404–413.

[37] 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*.

[38] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017. Sequential Matching Network: A New Architecture for Multi-turn Response Selection in Retrieval-Based Chatbots. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. ACL, 496–505.

[39] 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 *Proceedings of International conference on machine learning*. JMLR.org, 2048–2057.

[40] Zhen Xu, Bingquan Liu, Baoxun Wang, Chengjie Sun, and Xiaolong Wang. 2017. Incorporating loose-structured knowledge into conversation modeling via recall-gate LSTM. In *Proceedings of the International Joint Conference on Neural Networks*. INNS, 3506–3513.

[41] Rui Yan, Yiping Song, and Hua Wu. 2016. Learning to Respond with Deep Neural Networks for Retrieval-Based Human-Computer Conversation System. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. ACM, 55–64.

[42] Rui Yan, Dongyan Zhao, and Weinan E. 2017. Joint Learning of Response Ranking and Next Utterance Suggestion in Human-Computer Conversation System. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 685–694.

[43] Liu Yang, Minghui Qiu, Chen Qu, Jiafeng Guo, Yongfeng Zhang, W. Bruce Croft, Jun Huang, and Haiqing Chen. 2018. Response Ranking with Deep Matching Networks and External Knowledge in Information-seeking Conversation Systems. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 245–254.

[44] Kaisheng Yao, Geoffrey Zweig, and Baolin Peng. 2015. Attention with Intention for a Neural Network Conversation Model. *arXiv preprint arXiv:1510.08565*.