User Attention-guided Multimodal Dialog Systems

Chen Cui
Shandong University
chentsuei@gmail.com

Wenjie Wang
Shandong University
wenjiewang96@gmail.com

Xuemeng Song
Shandong University
sxmustc@gmail.com

Minlie Huang
Tsinghua University
aihuang@tsinghua.edu.cn

Xin-Shun Xu
Shandong University
xuxinshun@sdu.edu.cn

Liqiang Nie
Shandong University
nieliqiang@gmail.com

ABSTRACT

As an intelligent way to interact with computers, the dialog system has been catching more and more attention. However, most research efforts only focus on text-based dialog systems, completely ignoring the rich semantics conveyed by the visual cues. Indeed, the desire for multimodal task-oriented dialog systems is growing with the rapid expansion of many domains, such as the online retailing and travel. Besides, few work considers the hierarchical product taxonomy and the users' attention to products explicitly. The fact is that users tend to express their attention to the semantic attributes of products such as color and style as the dialog goes on. Towards this end, in this work, we present a hierarchical User attention-guided Multimodal Dialog system, named UMD for short. UMD leverages a bidirectional Recurrent Neural Network to model the ongoing dialog between users and chatbots at a high level; As to the low level, the multimodal encoder and decoder are capable of encoding multimodal utterances and generating multimodal responses, respectively. The multimodal encoder learns the visual presentation of images with the help of a taxonomy-attribute combined tree, and then the visual features interact with textual features through an attention mechanism; whereas the multimodal decoder selects the required visual images and generates textual responses according to the dialog history. To evaluate our proposed model, we conduct extensive experiments on a public multimodal dialog dataset in the retailing domain. Experimental results demonstrate that our model outperforms the existing state-of-the-art methods by integrating the multimodal utterances and encoding the visual features based on the users' attribute-level attention.

CCS CONCEPTS

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

KEYWORDS

Multimodal Dialog Systems; Multimodal Response Generation; Multimodal Utterance Encoder; Taxonomy-attribute Combined Tree

ACM Reference Format:
Chen Cui, Wenjie Wang, Xuemeng Song, Minlie Huang, Xin-Shun Xu, and Liqiang Nie. 2019. User Attention-guided Multimodal Dialog Systems. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19), July 21–25, 2019, Paris, France.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3331184.3331226

1 INTRODUCTION

In the past few years, dialog systems have penetrated into many aspects of our lives, and have been gaining increasing research interests [4]. Roughly speaking, dialog systems fall into two categories: open-domain dialog systems [5, 15, 29, 34, 36, 41] and task-oriented dialog systems [22, 25, 32, 35, 40, 43]. The former is able to chat with users on a wide range of topics without domain restrictions; whereas the latter helps users to accomplish specific tasks in certain vertical domains, such as the catering and travel. Notably, both research and industrial communities have reached the consensus that a robust and efficient task-oriented dialog system is capable of improving the user experience and thereby boost sales [4, 32]. In the light of this, we focus on improving task-oriented dialog systems in this work, especially the multimodal dialog system in the online retailing domain.

The traditional task-oriented dialog systems usually follow a typical pipeline [13, 22, 40], comprising four components: 1) natural language understanding (NLU); 2) dialog state tracker (DST); 3) policy network; and 4) natural language generation (NLG). To be more specific, the first component NLU is implemented to encode users' utterances towards understanding users' intention via categorization. Following that, DST keeps tracking users' goals and constraints as the conversation continues. And most importantly, it determines the values of predefined slots in each turn. Thereafter, the policy network component is responsible to decide what actions to take at the next step. Ultimately, NLG gives the final responses based on the efforts of the former steps, which technologically can be implemented by the predefined sentence templates [11] or generation-based methods [7]. In addition to the four-stage pipeline, some end-to-end task-oriented dialog systems emerge [25]. Thereinto, the reinforcement learning has proven to be effective in this task [7, 35].

Despite the success of task-oriented dialog systems in various tasks, they still suffer from the following limitations. 1) The

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

SIGIR '19, July 21–25, 2019, Paris, France

ACM ISBN 978-1-4503-6172-9/19/07…$15.00
https://doi.org/10.1145/3331184.3331226

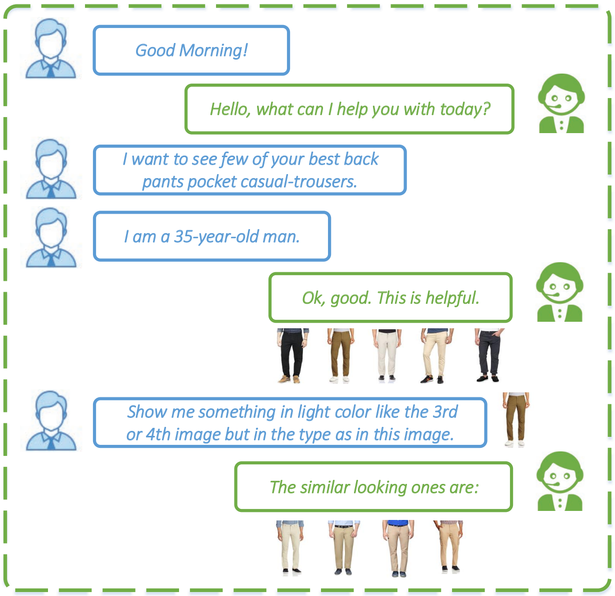

Good Morning!
Hello, what can I help you with today?
I want to see few of your best back pants pocket casual-trousers.
I am a 35-year-old man.
Ok, good. This is helpful.
Show me something in light color like the 3rd or 4th image but in the type as in this image.
The similar looking ones are:

Figure 1: Illustration of a multimodal dialog between a shopper and a chatbot. The shopper expresses his requirements and preference for products step by step as the dialog goes on. And the chatbot generates the multimodal responses based upon the context.

old proverb says, "a picture is worth a thousand words", yet most existing dialog systems only focus on the textual utterances, ignoring the fact that people tend to communicate with multimodal information. 2) To obtain the desired product, users may particularly pay more attention to certain aspects or attributes of products when interacting with the chatbots. As shown in Figure 1, the user expresses the preferred attributes with visual images, such as color and style. However, very limited efforts have been dedicated to the attribute-level attention of users. Accordingly, it is highly desired to devise smarter multimodal dialog systems considering the users' attribute-level attention.

However, it is non-trivial to well-address the aforementioned problems due to the following challenges. 1) In existing e-commerce websites, the products are actually organized as a hierarchical tree structure, whereby similar products share more common properties. Modeling the informative taxonomy when encoding the visual images to learn the distinguishable and interpretable representation is a challenge we are facing. 2) In the dialog settings, users usually describe their attention through text. Thus how to identify the pivotal words in users' descriptions and effectively extract more informative textual features is a tough issue. And 3) indeed, the products share some common attributes, such as color, style and material, semantically describing the key characteristics of products. Users' attention to products is frequently expressed with these attributes, as exemplified in Figure 1. Therefore, how to integrate images and text to explore the users' attribute-level attention is worth studying.

In this paper, we propose a novel User attention-guided Multimodal Dialog system to address the issues mentioned above, named UMD for short. As shown in Figure 2, from the high-level perspective, a bidirectional Recurrent Neural Network (RNN) is applied to model the interaction between the user and the chatbot; As to the low-level perspective, the multimodal encoder and decoder are supposed to encode multimodal utterances and generate multimodal responses, respectively. In the multimodal encoder, in order to acquire more distinguishable and interpretable visual representation, we apply a hierarchy-aware tree encoder to learn the taxonomy-guided attribute-level visual representation. For textual utterances, the textual RNN, augmented by a Convolutional Neural Network (CNN)-based attention mechanism, takes the textual messages as input and outputs attentive textual features. As shown in Figure 3, visual features are extracted by a CNN model, and then fed into a taxonomy-attribute combined tree. Encoded by the hybrid tree, visual features are then interacted and weighted by textual features in the attribute level. Ultimately, the visual and textual features are fed into a Multimodal Factorized Bilinear Pooling (MFB) [42] module to generate the utterance vector. Pertaining to the multimodal decoder, it inputs the context vector from the high level RNN, and then outputs a textual response and the selected images. Overall, a RNN-based response decoder generates the textual response and the model ranks the images by maximizing the margin between the cosine similarities for the positive and negative samples. To justify the effectiveness of our proposed model, we compare it with several state-of-the-art baselines over a multimodal dialog (MMD) dataset. The experimental results show the superiority of our model.

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

- We propose a novel hierarchical encoder to learn the taxonomy-guided attribute-level representation of product images in multimodal dialog systems.
- As far as we know, this is the first work to attentively integrate the images and text to explore users' attribute-level attention to products in dialog systems.
- We comprehensively justify our model by comparing it with several state-of-the-art baselines. In addition, we release our code and data to promote the research in this field[1].

This paper is structured as follows. The related work is introduced in Section 2. In Section 3, we explain the proposed model in detail, followed by the experiments and the analysis of the model performance in Section 4. Finally, Section 5 concludes the work and figures out the future research directions.

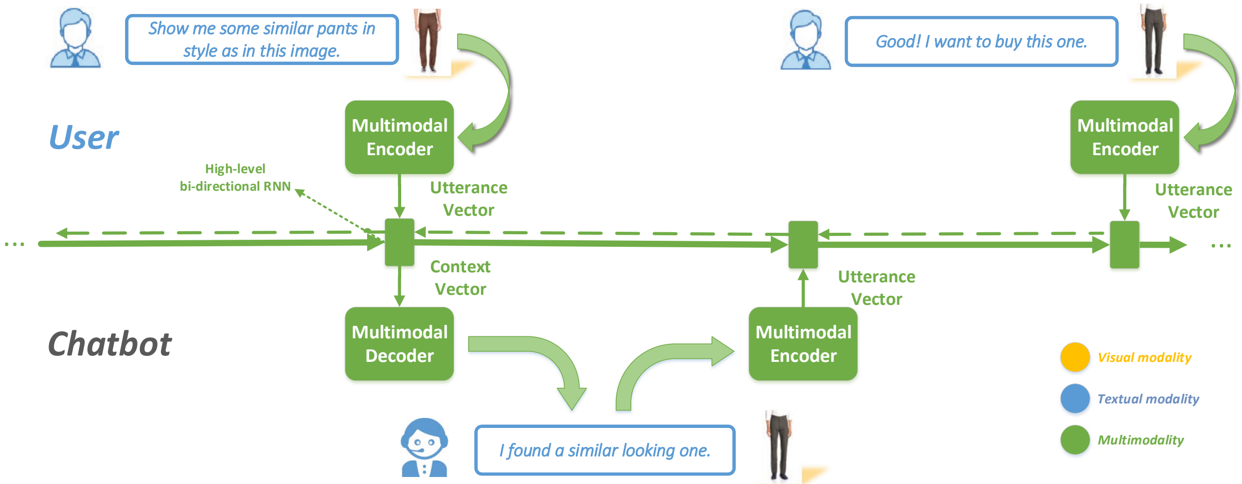
2 RELATED WORK

Our work is closely related to a variety of dialog systems, which could be roughly divided into two categories: the text-based dialog systems and the multimodal ones.

2.1 Text-based Dialog Systems

Extensive research efforts have been dedicated to the study of the text-based open-domain and task-oriented dialog systems over the past few years. According to their applications, open-domain dialog systems [5, 36] aim to chat with people in diverse topics, while task-oriented ones [32] focus on assisting users to

[1] https://github.com/ChenTsuei/UMD



Show me some similar pants in style as in this image.

Good! I want to buy this one.

I found a similar looking one.

Figure 2: Schematic illustration of our proposed model. At the high level, a bidirectional RNN is used to model the ongoing dialog; As to the low level, the multimodal encoder and decoder are applied to encode multimodal utterances and generate multimodal responses, respectively.

accomplish some specific tasks. The former is technologically implemented by retrieval- or generation-based methods. Retrieval-based methods [38, 39, 41] leverage the dialog history to rank the response candidates, and then return the top one to users. By contrast, the generation-based ones model the mapping between the dialog history and its response using an encoder-decoder framework [33]. Recently, the attention mechanism [2] has been incorporated into these generation-based methods to improve the performance [21]. As for the multi-turn text-based dialog systems, HRED [27] encodes the multi-turn context and generates its response hierarchically while VHRED [28] adds latent stochastic variables to HRED for diverse responses. Besides, deep reinforcement learning is also used to strengthen the generation-based dialog systems [16].

Different from formulating the response generation as a mapping problem, task-oriented dialog systems follow a typical pipeline [13, 22]. They usually encode user utterances firstly and then determine the current state. They next decide the following policy, take the corresponding action and give the final response orderly according to the current state. However, this pipeline brings several serious problems [14]. 1) Errors from upstream components would propagate and accumulate along the pipeline. 2) There is heavy interdependence among the components. And 3) the training and testing of these task-oriented dialog systems require large-scale annotated data in specific domains. To alleviate these issues, many end-to-end dialog systems [3, 17, 35] integrating the strength of reinforcement learning and supervised learning have been introduced recently. Knowledge [9] is also integrated into dialog systems to generate more informative responses. Many of them [25, 35] issue a symbolic query to retrieve the required entries or relations from a knowledge base (KB), which is replaced by "soft" posterior distribution over KB in subsequent methods [9]. Lei et al. [14] developed a belief spans-based framework to avoid the complex architecture. Although the existing dialog systems have made much progress, these efforts neglect the importance of visual information in the human-machine dialog.

2.2 Multimodal Dialog Systems

Due to the rich visual semantics conveyed by product images[10], the demand for multimodal dialog systems is increasing. However, the study in this area has been limited due to the lack of large-scale multimodal dialog datasets. To this end, Saha et al. [26] constructed a MMD benchmark dataset. Besides, the authors developed two benchmark baselines for two tasks: the textual response generation and the best image response selection. Later, Liao et al. [18] presented a knowledge-aware multimodal dialog (KMD) model to generate more substantive responses, where deep reinforcement learning is integrated with the hierarchical neural models to improve the performance. Different from KMD, in this work, we pay more attention to the user requirements explicitly in the attribute level and encode the dialog history dynamically based on users' attention.

In addition, our work is relevant to several cross-modal problems, such as visual question answering (VQA) [1], visual dialog [6] and image captioning [37]. The difference is that multimodal dialog systems focus more on multi-turn multimodal interaction between users and chatbots.

3 USER ATTENTION-GUIDED MULTIMODAL DIALOG SYSTEM

In this section, we will detail the proposed model. To avoid the heavy reliance on the annotated data of the typical pipeline, we present a unified neural model to accomplish two tasks: the textual response generation and the best image response selection. As shown in Figure 2, the proposed scheme models the multimodal dialog hierarchically: from the low-level perspective, the multimodal encoder and decoder are able to encode multimodal utterances and generate multimodal responses, respectively. Meanwhile, the high-level RNN model characterizes the entire dialog process at the

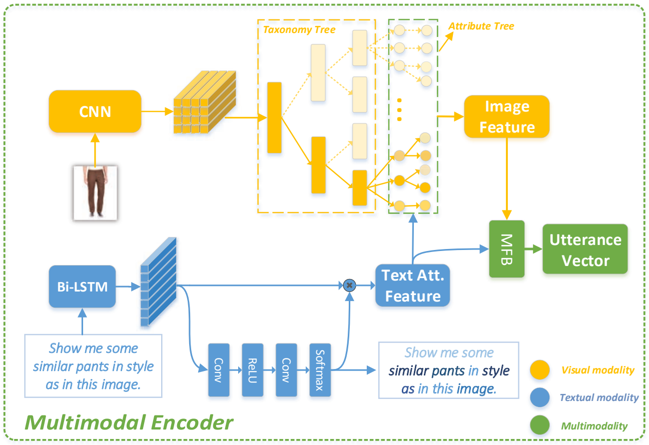


Figure 3: Schematic illustration of the multimodal encoder. A taxonomy-attribute combined tree is applied to learn the visual representation. The attention-augmented RNN encoder is incorporated to output attentive textual features and then the visual features are weighted by textual ones in the attribute level. They are ultimately fed into a multimodal fusion layer (MFB module) to generate the utterance vector.

utterance level. To be more specific, the multimodal encoder takes users' and chatbots' multimodal utterances as input and outputs the utterance vector. The high-level RNN inputs the utterance vector, and outputs the hidden state as the context vector at each step. As for the multimodal decoder, it generates a textual response and selects several images on the basis of current context vector. Notably, some utterances in the dialog may be only presented in the textual modality, where the encoder and decoder do not process visual images. Ultimately, the multimodal response is fed back to users.

3.1 Multimodal Encoder

In this component, we integrate the textual utterances and visual images to learn the multimodal utterance representation. Given a textual utterance $\mathcal{U}$ and a product image i, we leverage a taxonomy-attribute combined tree and attention-augmented RNN to learn the taxonomy-guided attribute-level visual representation and extract attentive textual features, respectively.

3.1.1 Attention-augmented RNN. It is well-known that the words in textual utterances are not equally important. Some words could convey important information regarding users' intention and preferences, while others may be some common or supportive words in our daily conversations, such as "hello", "is" and "me". The latter is extremely frequent in the training data, heavily hindering the propagation of users' requirements. To alleviate this problem, we leverage a CNN-based attention mechanism to attentively weigh the words in textual utterances in order to maximize the useful information about users' requirements.

The textual utterance $\mathcal{U}$ is first fed into a RNN model equipped with bidirectional Long Short Term Memory units (Bi-LSTM), and then weighted by the scores of CNN-based attention mechanism to output the final attentive textual features. Formally, a sequence of hidden states of the RNN encoder are calculated as follows,

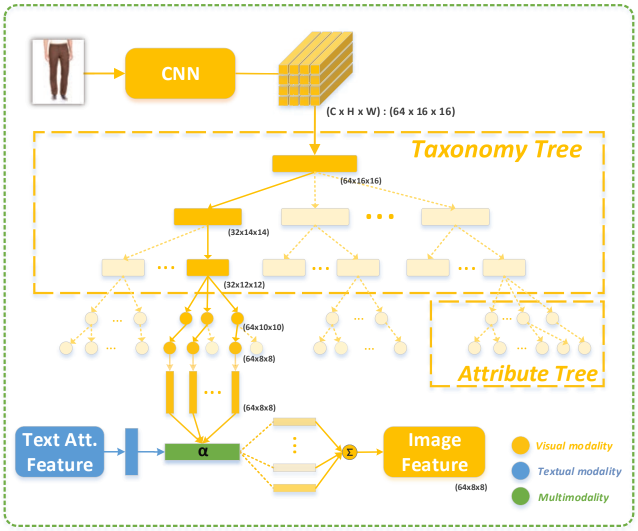


Figure 4: The proposed taxonomy-attribute combined tree. The solid lines connect the nodes that the image will pass through from top to bottom; whereas the dash lines denotes the irrelevant categories. Notably, all products share N common attribute nodes in the attribute tree.

$$\begin{cases} \mathcal{U} = \{w_1, w_2, ..., w_T\}, \\ \mathbf{h_t} = f(\mathbf{h_{t-1}}, \mathbf{e_{w_t}}), \end{cases} \tag{1}$$

where w_t denotes the t-th token in the textual utterance $\mathcal{U}$, T is the length of the textual utterance, $\mathbf{e_{w_t}}$ refers to the embedding vector of w_t, $\mathbf{h_t}$ is the hidden state of the RNN at time t, and f is the non-linear function in LSTM units. Thereafter, the hidden states are put into a CNN model to estimate their weights,

$$\begin{cases} \mathbf{s} = CNN(\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_T}), \\ \alpha_i = \dfrac{exp(s_i)}{\sum_{j=1}^{T} exp(s_j)}, \end{cases} \tag{2}$$

where the textual CNN structure is shown in Figure 3. Thereinto, the symbol $\mathbf{s}$ refers to the output of the CNN, and the weights $\alpha_i s$ are acquired by the softmax of $\mathbf{s}$. Ultimately, the textual feature $\mathbf{t}$ is calculated by

$$\mathbf{t} = \sum_{i=1}^{T} \alpha_i \mathbf{h_i}. \tag{3}$$

3.1.2 Taxonomy-attribute combined tree. In many e-commerce websites, extensive products are divided into various categories, and organized into a hierarchical tree structure. Intuitively, the same kind of products share a lot of common visual features[20]. Taking the pants in Figure 1 as an example, they are similar in many visual properties, such as shape, proportion and appearance, which facilitate users to navigate or recognize the desired products (men's pants) easily. Another observation is that these pants are distinguishable in the attribute level and the customers always select them by these detailed attributes, such as color, style and material. Therefore, in order to extract more representative and distinguishable visual features, we introduce a taxonomy-based hierarchical encoder. Besides, we define N common attributes for products, and then construct a key-value attribute tree to explore users' attention to products in the attribute level. The keys correspond to the N common attributes, such as color; while the values are the specific values of the attributes. For example, the attribute "color" has several values such as blue, black and yellow.

As shown in Figure 4, given a product image i, it is firstly encoded by a CNN module,

$$\mathbf{v^0} = CNN(i), \tag{4}$$

where the CNN module is implemented by several pre-trained layers based upon the Deep Residual Network [12] and the specific parameters are listed in Figure 4. It is followed by a taxonomy-based tree, consisting of L layers and M leaves. Each leaf node denotes a kind of products and the categories are organized as a hierarchical tree. Notably, there is only one path where an image walks through from top to bottom because a product image only falls into one leaf category. Next, the features are fed into N parallel attribute nodes, and then their corresponding value nodes. Formally, supposing the given image i belongs to the path $\mathcal{P} = \{p_1, p_2, ..., p_L\}$ and has attribute value encoders $\mathcal{A} = \{a_1^v, a_2^v, ..., a_N^v\}$, we can update the visual features $\mathbf{v}$ with the guide of taxonomy information in the attribute level,

$$\begin{cases} \mathbf{v^1} = p_1(\mathbf{v^0}), \\ ...... \\ \mathbf{v^{L-1}} = p_{L-1}(\mathbf{v^{L-2}}), \\ \mathbf{v^L} = p_L(\mathbf{v^{L-1}}), \\ \mathbf{v_1} = a_1^v(a_1^k(\mathbf{v^L})), \\ ...... \\ \mathbf{v_N} = a_N^v(a_N^k(\mathbf{v^L})), \end{cases} \tag{5}$$

where p_L denotes the CNN encoder of the image i in the L-th layer of the taxonomy tree, $\mathbf{v^L}$ refers to the output of p_L, a_N^k means the encoder of the N-th attribute key node, a_N^v corresponds to the encoder of the N-th attribute value node for image i, and $\mathcal{V} = \{\mathbf{v_1}, \mathbf{v_2}, ...\mathbf{v_N}\}$ is the output of the taxonomy-attribute combined tree encoder for the image i.

Next, the attentive textual feature $\mathbf{t}$ and visual feature $\mathcal{V}$ are used to calculate the users' attention scores in the attribute level,

$$\begin{cases} e_i = MFB(\mathbf{v_i}, \mathbf{t}) \quad i = 1...N, \\ \beta_i = \dfrac{exp(e_i)}{\sum_{j=1}^{N} exp(e_j)}, \end{cases} \tag{6}$$

where MFB integrates the multimodal features and outputs the corresponding scores, and $\beta_i s$ denote the attention scores of visual features. Ultimately, the taxonomy-guided attribute-level visual representation $\mathbf{v}$ is updated by

$$\mathbf{v} = \sum_{i=1}^{N} \beta_i \mathbf{v_i}. \tag{7}$$

3.1.3 Multimodal fusion layer. Rather than simply concatenating the textual feature $\mathbf{t}$ and visual feature $\mathbf{v}$, we leverage a MFB layer to get the multimodal utterance vector, which has demonstrated the effectiveness and efficiency of combining multimodal features in the VQA task [42]. In this way, the multimodal utterance vector $\mathbf{u}$ is formulated as,

$$\mathbf{u} = SumPooling(\mathbf{U^T t} \circ \mathbf{V^T v}, k), \tag{8}$$

where $\mathbf{U^T}$ and $\mathbf{V^T}$ are the transform matrices projecting $\mathbf{t}$ and $\mathbf{v}$ to the common high dimensional space, respectively, $\circ$ denotes the element-wise product, and the function $SumPooling(\mathbf{x}, k)$ means using a one-dimensional non-overlapped window with the size k to perform sum pooling over vector $\mathbf{x}$.

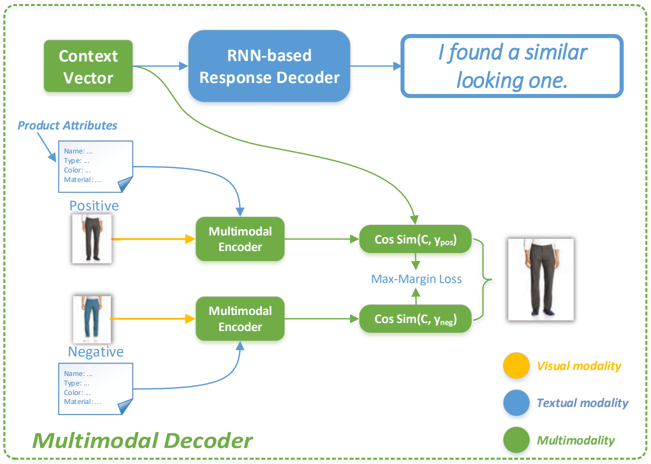


Figure 5: Illustration of the multimodal decoder. A RNN-based decoder is to generate a textual response; And the image selection is to maximize the consine similarity for positive and negative samples.

3.2 Utterance-level RNN

Utterance-level RNN transfers the information between users and chatbots. Utterance vectors are from the multimodal encoder; whereas context vectors are fed into the multimodal decoder.

As shown in Figure 2, the utterance RNN takes the utterance vector $\mathbf{u_i}$ from the multimodal encoder, and then calculates a current hidden state as the context vector $\mathbf{c_i}$ at the step i,

$$\mathbf{c_i} = f(\mathbf{c_{i-1}}, \mathbf{u_i}), \tag{9}$$

where $\mathbf{u_i}$ refers to the multimodal utterance vector at the step i, and f is the non-linear function in LSTM units.

3.3 Multimodal Decoder

3.3.1 RNN-based response decoder. The objective of this component is to generate a textual response on the basis of context vector $\mathbf{c}$. The RNN is initialized by the vector $\mathbf{c}$, and updated by

$$\mathbf{s_t} = f(\mathbf{s_{t-1}}, \mathbf{e_{w_{t-1}}}), \tag{10}$$

where $\mathbf{s_t}$ denotes the hidden state at the step t and $\mathbf{e_{w_{t-1}}}$ refers to the embedding of token $\mathbf{w_{t-1}}$ in the response. The RNN decoder calculates the probability of every token in the response by linearly projecting the hidden state to a one-dimensional vector in the vocabulary size,

$$p(w_i | \mathbf{c}, w_1, ..., w_{i-1}) = \mathbf{o_t} \cdot \sigma_s(\mathbf{W_p s_t} + \mathbf{b_p}), \tag{11}$$

whereby $\mathbf{W_p}$ and $\mathbf{b_p}$ are the parameters of the linear projection layer, σ_s means the softmax function, w_i is the i-th token in the response and $\mathbf{o_i}$ is the one-hot vector of w_i. Formally, the probability of generating the whole response $p(w_1, w_2, ..., w_T|\mathbf{c})$ is given by

$$p(w_1, ..., w_T|\mathbf{c}) = p(w_1|\mathbf{c}) \prod_{i=2}^{T} p(w_i|\mathbf{c}, w_1, ..., w_{i-1}). \tag{12}$$

The loss function of a textual response is formulated as,

$$\ell_{\text{text}} = -\log p(w_1|\mathbf{c}) - \sum_{i=2}^{T} \log p(w_i|\mathbf{c}, w_1, ..., w_{i-1}), \tag{13}$$

where the smaller ℓ_{text} implies the higher probability to generate the entire target response $\{w_1, w_2, ..., w_T\}$.

3.3.2 Pairwise ranking. Given a set of visual images, this component is to rank them based on the relevance between the image and the context. In addition, considering the connection between the context and the product attributes behind visual images, we especially incorporate the textual attributes into the ranking process.

Formally, the textual attributes are organized as a sequence of words, and then fed into the multimodal encoder with visual images, finally outputting the multimodal product representation. In this task, given some products comprising N_{pos} positive samples and N_{neg} negative ones for a dialog sample, we calculate the cosine similarity between their product representations and the context vector. When training the model, a max-margin loss is applied to maximize the margin between the similarities for the positive and negative samples,

$$\ell_{\text{image}} = max(0, 1 - Sim(\mathbf{c}, \mathbf{y_{pos}}) + Sim(\mathbf{c}, \mathbf{y_{neg}})), \tag{14}$$

where $\mathbf{y_{pos}}$ and $\mathbf{y_{neg}}$ are the representations of positive and negative samples, respectively, and the function $Sim(\mathbf{a}, \mathbf{b})$ calculates the cosine similarity of $\mathbf{a}$ and $\mathbf{b}$. As for the testing period, the model ranks the images based on the cosine similarity.

4 EXPERIMENTS

In this section, we first introduce the experimental dataset and settings, including hyper parameters, evaluation metrics and several baselines. It is followed by the objective and subjective comparison between the baselines and UMD. Ultimately, we present some representative cases and the error analysis.

4.1 Dataset

A large-scale multimodal dialog dataset in vertical domains plays a pivotal role in developing the task-oriented multimodal dialog systems. Fortunately, Saha et al. [26] contributed a MMD benchmark dataset in the retailing domain with over 150k conversations between customers and chatbots, and each conversation describes a complete online shopping process. During the conversations, the user proposes his/her requirements in multimodal utterances and the chatbot introduces different products step by step until they make a deal. Each multimodal conversation involves images and text, and is constructed by the in-house annotators using a *semi-automated manually intense iterative manner* [26] under the supervision of domain experts. Over 1 million fashion products with their available semi/un-structured information are collected from the well-known online retailing websites, such as Amazon[2], Jabong[3], and Abof[4]. Notably, we crawled their visual images from the websites additionally and released them in our experimental data. Based on the MMD benchmark dataset, Saha et al. also proposed several research tasks, including the textual response generation and the best image response selection. For the former task, the textual responses of chatbots are selected as the predicted ones and their preceding multimodal utterances are treated as the dialog context in each conversation. When it comes to the latter, the annotators selected several negative samples for each image in the conversation. The MMD dataset provides five target images for each sample and only one image is correct. Considering that the users tend to express their attention to the products in the attribute level, we incorporated the attributes of products into the selection of images. We chose several common attributes from the product information and the key-value attributes were organized as a sequence of words. More detailed information about the MMD dataset is summarized in Table 1.

Table 1: Detailed statistics of the MMD dataset.

Dataset Statistics	Train	Valid	Test
#Dialogs(chat sessions)	105,439	22,595	22,595
#Proportion in terms of dialogs	70%	15%	15%
Avg. #Utterances per dialog	40	40	40
#Utterances with shopper's question	2M	446K	445K
#Utterances with agent's image response	904K	194K	193K
#Utterances with agent's text response	1.54M	331K	330K
Avg. #Positive images in agent's image response	4	4	4
Avg. #Negative images in agent's image response	4	4	4
Avg. #Words in shopper's question	12	12	12
Avg. #Words in agent's text response	14	14	14
#Vocabulary Size (threshold frequency>=4)	26,422	-	-

4.2 Experimental Settings

4.2.1 Hyper parameters. In the training period, following the parameters settings in MMD [26], we used two-turn preceding utterances before the response as the context. The vocabulary size is 26,422 and the low frequency words out of the vocabulary is mapped to a special token "UNK". In the multimodal encoder, textual utterances are encoded by a bidirectional LSTM model with one layer and 1,024 cells. The kernel sizes of the two-layer textual CNN model are (128 × 1 × 1) and (30 × 1 × 1), respectively. As for the taxonomy-attribute combined tree, we defined 3 layers, 87

[2] https://www.amazon.com/.

[3] https://www.jabong.com/.

[4] https://www.abof.com/.

leaves, and 6 attributes. Other detailed visual CNN kernel sizes are displayed in Figure 4. The window size k in MFB is set as 2. In the multimodal decoder, textual responses are generated by a LSTM model with 1,024 cells. The margin in the max-margin loss is set as 1. We optimized the parameters of the unified model using Adam [8] with the learning rate initialized as 0.0004.

4.2.2 Evaluation Metrics. To evaluate our proposed model, we adopted several objective metrics following the former studies [26]. In the task of textual response generation, BLEU-N [23] is applied to measure the similarity between the predicted response and the reference. To be more specific, BLEU-N is formally defined as,

$$BLEU\text{-}N = \exp(min(1 - \frac{r}{c}, 0) + \sum_{n=1}^{N} w_n \log p_n), \tag{15}$$

where p_n refers to the modified n-gram precision in [23], w_n is equal to $\frac{1}{N}$, and r and c denote the lengths of the reference response and the predicted one, respectively. Intuitively, higher BLEU scores mean more n-gram overlaps between the compared responses, and thereby indicate better performance. Meanwhile, with more grams varied from unigram to 4-gram, BLEU-4 is used more frequently in the task of machine translation and dialog systems [23, 26].

For the best image response selection, we used *Recall@top-m* to measure the performance of the models, where m is varied from 1 to 3. The result is correct only if the positive sample is ranked in the *top-m* samples.

4.2.3 Baselines. To demonstrate the effectiveness of our proposed user attention-guided multimodal dialog system, we compared our model with several representative methods.

- SEQ2SEQ+Attention: As a representative encoder-decoder framework, attention-based SEQ2SEQ [2] has demonstrated its effectiveness in many natural language processing tasks, therefore it is generally used as a baseline in generation-based dialog systems.
- HRED: In text-based multi-turn dialog systems, HRED [27] is a state-of-the-art method by modeling the long context hierarchically. A word-level RNN encodes each word in one sentence step by step; and a sentence-level RNN is applied to encode the sentence representation.
- MHRED: Multimodal hierarchical encoder decoder (MHRED) architecture is proposed by Saha et al. [26] along with the MMD dataset. It is the first work to construct the multimodal dialog system, which incorporates the visual features into the text-based HRED model and achieves the promising performance.
- KMD: KMD [18] incorporates memory network [31] and deep reinforcement learning into the multimodal dialog systems and achieves the state-of-the-art performance. However, the complex structure and the special need for semi-structured product data heavily hinder the reproducibility of KMD. Without the required data and code, we failed to reproduce the performance of the model and thus only compared it with UMD by the metrics reported in [18].

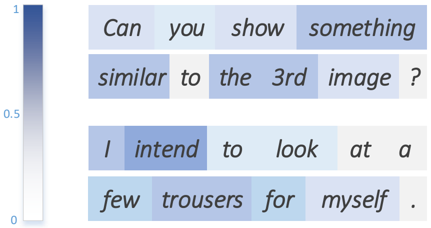


Figure 6: The visualization of the attention scores of two sentences in the context.

4.3 Objective Performance

4.3.1 Evaluating the textual response generation. For the task of the textual response generation, we applied BLEU-N metrics to measure the performance, where N varies from 1 to 4. Table 2 presents the results of the baselines and UMD. From that, we observed the following points:

- UMD surpasses the baselines in BLEU scores, proving that on average, UMD generates more overlaps between the predicted responses and the reference than other methods.
- By analyzing the generated responses, we found that the relatively high BLEU-1 score owes to the more accurate short responses *(e.g., "Yes" and "No")* for the queries related to attributes of products, which demonstrates the incorporation of the taxonomy-attribute combined tree is efficient in representing users' attention to products in the attribute level.
- The text-only methods show comparable performance with the multimodal ones, demonstrating that the generation of textual utterances depends more on the textual features than visual features.

4.3.2 Evaluating the best image response selection. We evaluated the performance of the best image response selection by comparing the recall scores. We can observe the following findings from Table 2:

- UMD performs very well and surpasses all the baselines in this task. Almost all positive images are ranked at the top when testing the performance of UMD. In our opinion, we can analyze the reasons from three aspects: 1) The baselines leverage the 4,096 dimensional visual features extracted by VGGNet-16 [30] as the visual representation of products, which heavily restricts the effect of product images; whereas UMD applies the taxonomy-attribute combined tree to learn more distinguishable visual representation of the original product images. 2) The textual utterances in the context involve many users' requirements about the attributes of products, such as name, type, material and color. The experimental results implied that incorporating the attributes into the selection of images is really helpful. And 3) the positive and negative samples in MMD dataset usually have different categories or distinguishable attributes, partly leading to the superior performance of UMD.
- The performance of the multimodal methods is much better than the text-only ones in this task. It is because that the

Table 2: Objective performance of UMD and the baselines in the tasks of the textual response generation and the best image selection. *In particular, we failed to compare KMD with other methods by these metrics in the same textual testing data due to its special need for the semi-structured data constructed by its authors.

Methods		Text Task				Image Task		
	Metrics	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**Recall@1**	**Recall@2**	**Recall@3**
Text-only Methods	**SEQ2SEQ**	35.39	28.15	23.81	20.65	0.5926	0.7395	0.8401
	HRED	35.44	26.09	20.81	17.27	0.4600	0.6400	0.7500
Multimodal Methods	**MHRED**	33.56	28.74	25.23	21.68	0.7980	0.8859	0.9345
	KMD*	-	-	-	-	0.9198	0.9552	0.9755
	UMD(Ours)	**42.78**	**33.69**	**28.06**	**23.73**	**0.9796**	**0.9980**	**0.9990**

Table 3: Subjective comparison between the responses of UMD and other baselines according to four evaluation factors.

	Fluency				**Relevance**			
Opponent	Win	Loss	Tie	Kappa	Win	Loss	Tie	Kappa
UMD vs. SEQ2SEQ	**12.9%**	12.2%	74.8%	0.59	**17.0%**	7.5%	75.5%	0.46
UMD vs. HRED	**25.2%**	9.2%	65.6%	0.38	**20.1%**	7.5%	72.4%	0.40
UMD vs. MHRED	**84.0%**	5.1%	10.9%	0.60	**64.3%**	9.2%	26.5%	0.46

	Logical Consistency				**Informativeness**			
Opponent	Win	Loss	Tie	Kappa	Win	Loss	Tie	Kappa
UMD vs. SEQ2SEQ	**17.3%**	16.0%	66.7%	0.43	**30.3%**	24.5%	45.2%	0.48
UMD vs. HRED	**19.7%**	16.3%	63.9%	0.36	18.0%	**39.5%**	42.5%	0.50
UMD vs. MHRED	**64.3%**	10.2%	25.5%	0.49	**74.8%**	6.5%	18.7%	0.67

similarity among product images definitely plays a key role in the selection of the best images.

4.4 Subjective Evaluation

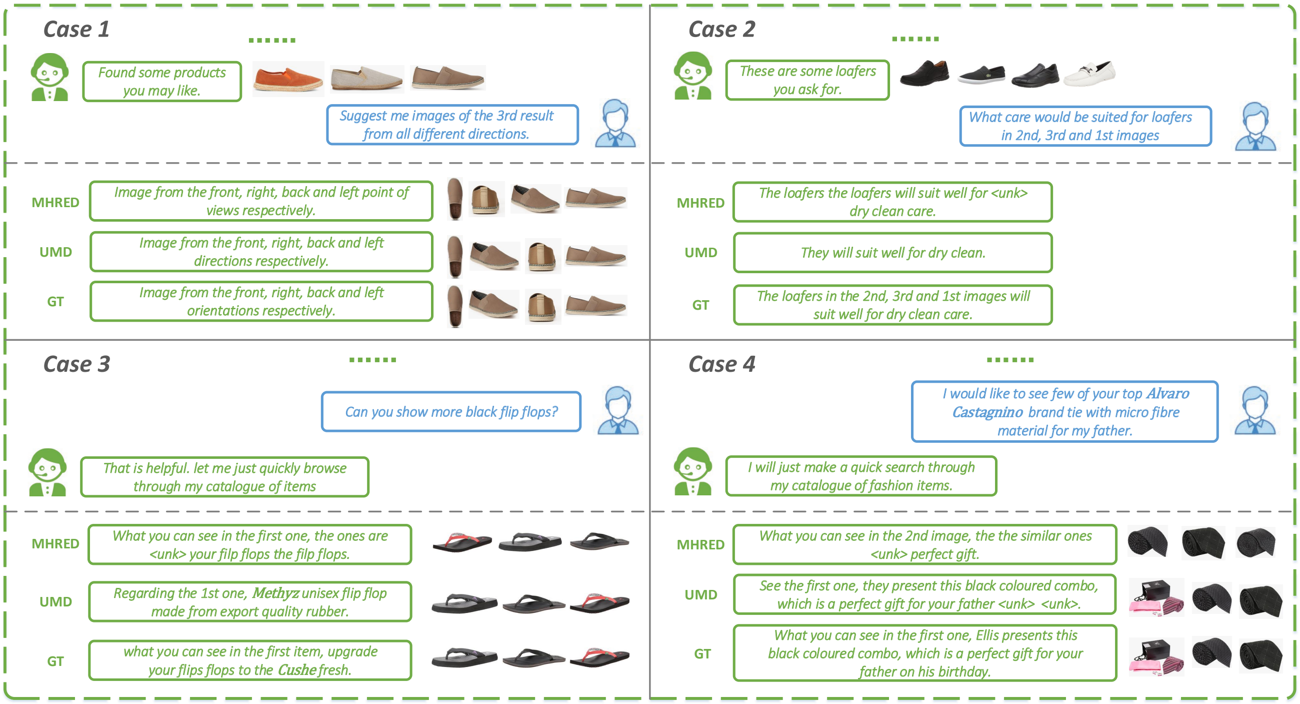
Considering that sometimes the objective metrics are not completely accurate to evaluate the responses [19], we also designed the subjective evaluation. 1,000 samples with the multimodal context are randomly chosen from the testing data, and then their contexts are fed into UMD and three baselines to generate the textual responses and select the visual images. Thereafter, the 1,000 multimodal responses of UMD are compared with the corresponding responses generated by the three baselines. In this way, we obtained 3,000 pair-wise responses. And then we carried out the subjective evaluation by the following disciplines: 1) Three annotators compared the pair-wise responses from four perspectives independently: fluency, informativeness, relevance, and logical consistency. 2) The annotators were required to choose one option from "win", "loss" and "tie", indicating "the first response is better", "the first response is worse" and "it is hard to tell which is better", respectively. Notably, the order of the responses is shuffled randomly. 3) When collecting the statistic results, we calculated the averaged values of three annotators and their kappa scores. As a result, the kappa scores [24] indicate that the annotators reached a moderate agreement with the quality of the responses. Ultimately, the results of subjective evaluation are presented in Table 3.

From Table 3, we can summarize the following conclusions: 1) UMD outperforms the baselines in most comparisons, especially in the logical consistency. 2) Most responses of SEQ2SEQ, HRED and UMD are fluent so that the annotators chose lots of "ties" in their comparisons. 3) HRED is inclined to generate more informative responses. Nevertheless, they are limited in maintaining the relevance and logical consistency. 4) MHRED performs the worst in the subjection evaluation. The repeated phases and syntax errors heavily hurt its fluency and logical consistency. 5) The comparable performance of text-only methods demonstrates our former conclusion again: although the visual features may help to generate the detailed product information, the textual response generation task depends more on the textual features.

4.5 Discussion

4.5.1 Case Study. Figure 7 lists four cases sampled from the test data, and only the responses generated by UMD and MHRED are shown due to the space limitations. Notably, MHRED actually performs better than other baselines in these samples. From Figure 7, we can have the following observations:

- The general responses are well predicted, such as "Image from the front, right, back and left orientations respectively." in Case 1 and "let me just quickly browse through my catalogue." in Case 3. It is because that they are frequent in the training data and their features are quite easy to recognize.
- Taking Case 2 and 3 for example, many responses generated by UMD are quite different from the ground truth while they are also reasonable. In fact, considering the taxonomy and attributes of products, UMD may convey other useful information to users by textual utterances.
- UMD generates more informative responses than MHRED by exploring more distinguishable visual features and more representative textual features. And MHRED tends to generate repeated phrases, which heavily decrease the



Case 1

Found some products you may like.

Suggest me images of the 3rd result from all different directions.

MHRED

Image from the front, right, back and left point of views respectively.

UMD

Image from the front, right, back and left directions respectively.

GT

Image from the front, right, back and left orientations respectively.

Case 2

These are some loafers you ask for.

What care would be suited for loafers in 2nd, 3rd and 1st images

MHRED

The loafers the loafers will suit well for <unk> dry clean care.

UMD

They will suit well for dry clean.

GT

The loafers in the 2nd, 3rd and 1st images will suit well for dry clean care.

Case 3

Can you show more black flip flops?

That is helpful. let me just quickly browse through my catalogue of items

MHRED

What you can see in the first one, the ones are <unk> your filp flops the filp flops.

UMD

Regarding the 1st one, *Methyz* unisex flip flop made from export quality rubber.

GT

what you can see in the first item, upgrade your flips flops to the *Cushe* fresh.

Case 4

I would like to see few of your top *Alvaro Castagnino* brand tie with micro fibre material for my father.

I will just make a quick search through my catalogue of fashion items.

MHRED

What you can see in the 2nd image, the the similar ones <unk> perfect gift.

UMD

See the first one, they present this black coloured combo, which is a perfect gift for your father <unk> <unk>.

GT

What you can see in the first one, Ellis presents this black coloured combo, which is a perfect gift for your father on his birthday.

Figure 7: Case Study. Each case includes the context, the ground truth responses (GT), and the responses generated by MHRED and UMD. The dash lines divide the context and its corresponding responses into two parts. The ellipsis implies that there are more conversational interactions before the current utterances, whereas they are omitted due to the space limitations.

fluency of responses but may produce higher BLEU scores, such as Case 2 and 3.

4.5.2 Error Analysis. To analyze the performance of UMD objectively, we collected the bad responses which "lose" in the subjective evaluation and tried to analyze the causes. We counted the proportion of the bad responses in each metric. Accordingly, the bad cases in fluency, relevance, informativeness and logical consistency occupy 16.2%, 14.7%, 43.2% and 25.9%, respectively. From that, we can conclude that: 1) Although UMD generates many fluent and relevant responses, a high percentage of bad responses lose in the comparison because they are not long enough and thereby decrease the informativeness. 2) The logical consistency of generated responses is still far from perfect due to the complexity of human language. These conclusions point out the disadvantages of UMD objectively and are definitely helpful to improve the model in the future.

5 CONCLUSION AND FUTURE WORK

In this work, we aim to build more intelligent multimodal dialog systems. To this end, we propose a hierarchical user attention-guided multimodal dialog system to learn the taxonomy-aware attribute-level visual representation and explore the user attention to products in the attribute level. From the high-level perspective, a bidirectional RNN model is applied to encode the utterance-level interaction between the user and chatbot. For the low-level perspective, the proposed multimodal encoder leverages a taxonomy-attribute combined tree and attention-based RNN to learn the multimodal utterance representation; whereas the multimodal decoder is designed to generate the textual responses and rank the visual images. We carry out extensive experiments on the MMD benchmark dataset and our proposed model yields promising performance in two tasks: the textual response generation and the best image response selection. Through the analysis of the experimental results, we can draw some conclusions: 1) Learning more distinguishable visual features by the multimodal encoder and incorporating the textual attributes into the multimodal decoder are very helpful in the selection of images. 2) Despite lots of responses of UMD are different from the ground truth, many of them are reasonable and convey meaningful information about the products. 3) As for the task of the textual response generation, it has a stronger dependence on textual features than visual features.

Although UMD performs well in the two basic tasks, we believe that there will be some other challenging issues in the real application. Therefore, we will proceed to promote the research work in the field from several aspects, such as the modeling of users' historical preference and the application of domain knowledge.

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 Tencent AI Lab Rhino-Bird Joint Research Program, No.: JR201805.

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] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. 2017. A survey on dialogue systems: Recent advances and new frontiers. *ACM SIGKDD Explorations Newsletter* 19, 2, 25–35.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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.

[10] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Xin-Shun Xu, and Mohan Kankanhalli. 2018. Multi-modal preference modeling for product search. In *2018 ACM Multimedia Conference on Multimedia Conference.* ACM, 1865–1873.

[11] 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.

[12] 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.

[13] Sungjin Lee and Maxine Eskenazi. 2013. Recipe for building robust spoken dialog state trackers: Dialog state tracking challenge system description. In *Proceedings of the 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue.* SIGDIAL, 414–422.

[14] 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.

[15] 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.

[16] Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016. Deep Reinforcement Learning for Dialogue Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* ACL, 1192–1202.

[17] 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.

[18] 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.

[19] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing.* ACL, 2122–2132.

[20] Meng Liu, Liqiang Nie, Xiang Wang, Qi Tian, and Baoquan Chen. 2019. Online data organizer: micro-video categorization by structure-guided multimodal dictionary learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.

[21] 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.

[22] Nikola Mrksic, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve J. Young. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.* ACL, 1777–1788.

[23] 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.

[24] J. Randolph. 2005. Free-Marginal Multirater Kappa (multirater free): An Alternative to Fleiss Fixed-Marginal Multirater Kappa. *Joensuu Learning and Instruction Symposium.*

[25] 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.

[26] 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.

[27] 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.

[28] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence.* AAAI Press, 3295–3301.

[29] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics on Natural Language Processing.* ACL, 1577–1586.

[30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556.*

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.*

[36] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence.* AAAI Press, 3351–3357.

[37] 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.

[38] 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.

[39] 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.

[40] Zhao Yan, Nan Duan, Peng Chen, Ming Zhou, Jianshe Zhou, and Zhoujun Li. 2017. Building Task-Oriented Dialogue Systems for Online Shopping.. In *Proceedings of the 31st AAAI Conference on Artificial Intelligence.* AAAI Press, 4618–4626.

[41] 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.

[42] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering. In *Proceedings of IEEE International Conference on Computer Vision.* IEEE, 1839–1848.

[43] Zheng Zhang, Lizi Liao, Minlie Huang, Xiaoyan Zhu, and Tat-Seng Chua. 2019. Neural Multimodal Belief Tracker with Adaptive Attention for Dialogue Systems. In *Proceedings of the 28th International Conference on World Wide Web.* ACM.