MultiDM-GCN: Aspect-guided Response Generation in Multi-domain Multi-modal Dialogue System using Graph Convolutional Network

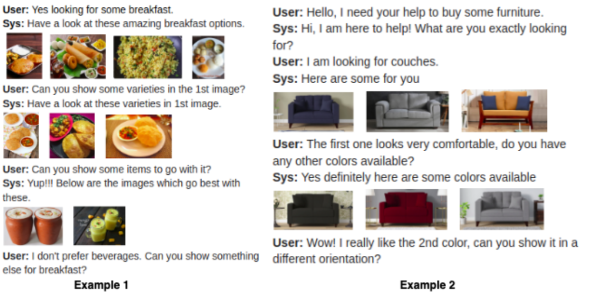
Mauajama Firdaus*, Nidhi Thakur*, Asif Ekbal
Department of Computer Science and Engineering
Indian Institute of Technology Patna
Patna, India
`(mauajama.pcs16,asif)@iitp.ac.in`

Abstract

In the recent past, dialogue systems have gained immense popularity and have become ubiquitous. During conversations, humans not only rely on languages but seek contextual information through visual contents as well. In every task-oriented dialogue system, the user is guided by the different aspects of a product or service that regulates the conversation towards selecting the product or service. In this work, we present a multi-modal conversational framework for a task-oriented dialogue setup that generates the responses following the different aspects of a product or service to cater to the user's needs. We show that the responses guided by the aspect information provide more interactive and informative responses for better communication between the agent and the user. We first create a Multi-domain Multi-modal Dialogue (MDMMD) dataset having conversations involving both text and images belonging to the three different domains, such as restaurants, electronics, and furniture. We implement a Graph Convolutional Network (GCN) based framework that generates appropriate textual responses from the multi-modal inputs. The multi-modal information having both textual and image representation is fed to the decoder and the aspect information for generating aspect guided responses. Quantitative and qualitative analyses show that the proposed methodology outperforms several baselines for the proposed task of aspect-guided response generation.

1 Introduction

Conversational systems have become ubiquitous in our everyday lives. Previous research suggests that the conversational agents need to be more interactive and informative for building engaging systems (Takayama and Arase, 2019; Shukla et al., 2019). These research indicates that engaging conversations include visual cues (e.g., a video or images) or audio cues (e.g., tone, the pitch of the speaker). Information contained in these cues is often integral for the conversation. In Figure 1, we show an example of a conversation where the visual cues in the form of images are crucial for better understanding and interactive dialogue between the agent and the user. The appropriate responses to the user queries are highly dependent on the visual information pertaining to the different aspects of the various images in the conversation. Thus, it is natural to conclude that a conversational agent would be more effective if the visual information were part of its underlying conversational model. Multi-modality in goal-oriented dialogue systems (Saha et al., 2018) for the fashion domain has established the significance of visual information for effective communication between the user and agent. Inspired by their works, we take a step forward by creating a multi-modal aspect guided response framework for a multi-domain goal-oriented dialogue system. From Figure 1, it can be observed that visual information of the aspects encourages improved communication and informative response generation by the agent with regards to the user queries.



User: Yes looking for some breakfast.
Sys: Have a look at these amazing breakfast options.
User: Can you show some varieties in the 1st image?
Sys: Have a look at these varieties in 1st image.
User: Can you show some items to go with it?
Sys: Yup!!! Below are the images which go best with these.
User: I don't prefer beverages. Can you show something else for breakfast?

User: Hello, I need your help to buy some furniture.
Sys: Hi, I am here to help! What are you exactly looking for?
User: I am looking for couches.
Sys: Here are some for you
User: The first one looks very comfortable, do you have any other colors available?
Sys: Yes definitely here are some colors available
User: Wow! I really like the 2nd color, can you show it in a different orientation?

Figure 1: Examples from the Multi-domain Multi-modal Dialogue(MDMMD) dataset

In this paper, we propose the task of generating

[*] First two authors have contributed equally

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 2318–2328
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

informative and interactive responses guided by the aspect information in a multimodal dialogue system. Firstly, we create a high quality multi-modal conversational dataset. Thereafter, we present a multi-modal graph convolutional network (GCN) that incorporates information from both textual and visual modalities to generate the aspect-guided responses. We aim to create a generalized response generation framework for a multi-domain multi-modal dialogue system that is informative, interesting, aspect-guided, and logical. Hence, the main contributions of this work are: (i) We propose the task of aspect-guided response generation for the interactive and informative responses in a multi-modal dialogue system. This is the first attempt to incorporate aspect information in the multi-modal dialogue systems to the best of our knowledge. (ii) We create a Multi-domain Multi-modal Dialogue (MDMMD) dataset comprising both text and images having conversations belonging to the three different domains, namely restaurant, electronics, and furniture. (iii) We propose a multi-modal graph convolutional framework for response generation while explicitly providing aspect information to the decoder to generate aspect-guided responses. (iv) The proposed model for both automatic and human evaluation shows its effectiveness over several baselines.

2 Related Work

Uni-modal Dialogue Systems The effectiveness of deep learning has shown significant progress in dialog generation. Deep neural frameworks, as shown in the (Vinyals and Le, 2015; Shang et al., 2015), are very effective in modeling conversations. The hierarchical encoder-decoder system was studied in (Sordoni et al., 2015; Serban et al., 2016, 2017; Xu et al., 2019) to preserve the dependencies among the utterances in dialogue. Recently, memory networks (Madotto et al., 2018; Raghu et al., 2018; Reddy et al., 2019; Tian et al., 2019; Wu, 2019; Chen et al., 2019b; Lin et al., 2019b) have been investigated to capture the contextual information in dialogues for generating responses. In task-oriented dialogues, hierarchical pointer networks (Raghu and Gupta) have been used to generate the responses. With the release of the task-oriented dialog dataset, such as MultiWoz , a few works (Budzianowski and Vulić, 2019; Chen et al., 2019a) have emerged that operate in a multi-domain dialogue setting. The meta-learning approach (Mi et al., 2019; Qian and Yu, 2019) has been implemented on the various datasets to improve the domain adaptability for generating responses.

Multi-modal Dialogue Systems Recently, research on the dialog system has shifted towards integrating various modalities, such as images, audio, and video, along with text, to obtain the information to build a robust framework. The research reported in (Das et al., 2017; Mostafazadeh et al., 2017; De Vries et al., 2017; Gan et al., 2019) has been effective in narrowing the gap between vision and language. Similarly in (Le et al., 2019; Alamri et al., 2018; Lin et al., 2019a), DSTC7 dataset has been used for response generation by incorporating audio and visual features. The release of the Multi-modal Dialog (MMD) dataset (Saha et al., 2018), having conversations on the fashion domain with the information from both texts and images, has facilitated the research on response generation (Agarwal et al., 2018b,a; Liao et al., 2018; Chauhan et al., 2019; Cui et al., 2019) in a multi-modal setup. Our newly designed framework is different from these existing ones, as our focus here is on creating aspect guided multi-modal dialogue dataset that contains the information of three different domains. Our present work distinguishes from the prior works of multi-modal dialog systems in the sense that we aim at generating responses conditioned on a particular aspect of the product or service in accordance with the conversational history.

Our research is novel concerning the following two aspects *viz.* (i). our research is focused on the task of aspect controlled dialog generation in a multi-modal setup; and (ii). we create a high-quality dataset that includes conversations belonging to multiple domains having both textual and image information.

3 Dataset

In this section, we describe the procedure of creating multi-modal dialogue data.

3.1 Data Creation Process

We come up with the following two top-level principles for domain selection after closer review and extensive discussions: (i). it encompasses a broad group of task-oriented frameworks used by industries/service providers and is likely to build user interfaces; (ii). for deeper comprehension and clarification of the services, the domains need visual details. Therefore, we choose to curate conversa-

Domain	Aspect Category	Aspect Terms
Electronics	Model_Type, Shapes, dimensions, brand, color,	Varies in each product(Mobiles, laptop, AC, TV, Fridge, Washing Machine; Rectangle, circle, square, cylindrical, oval; length, breadth, height; Samsung, Apple, LG; red, black, silver
Restaurant	Quantity, Cuisine_Type, Restaurant_ Type, Meal_Type, Course_type, Meal_course, Beverage_type, Dessert_type	One serving, Two serving; Indian, Italian, Chinese; Fast food, casual dining, fine dining; Breakfast, Lunch, Dinner; 2-course, 3-course, 4-course; appetizers, starters, dessert; juice, soft drinks, alcoholic drinks; chocolates, puddings, sweets;
Furniture	Furniture_style, Brand, Material, Room_type, Living_ftype	Contemporary, Modern, Traditional; CasaCraft, Amberville; Living room, Bedroom, Kids Room; Plywood, Veneer, Plastic, Stainless steel, Copper, Wrought Iron, Wood; Sofa, Chair, Table

Table 1: Aspect information per domain

tions belonging to three distinct domains in our newly established large-scale MDMMD dataset[1], namely restaurants, electronics, and furniture. With the cooperation of a dedicated team of 15 domain experts corresponding to each domain, the multi-modal dyadic dialogue aggregation was achieved. Given the various aspects of a product or service, the professionals from each domain demonstrated several dialogue flow during the selection and procurement of a specific product. The importance of various aspects in the sale of a product was established, whereafter these domain details were integrated with different chat sessions to make the conversations seamless and free-flowing. The creation of data concerns with the following key steps: (1). Data gathering; (2). Building a large-scale multimodal conversation includes both text and images, thus integrating the domain's information into the interaction; and (3). Aspect Annotation.

1. Data Gathering Method: As a consequence of the experts' interactions, we recognize the nuances of different styles in a natural conversation for every domain, guided by the background knowledge both the domain experts and the customer use these style information in their conversation. The necessary steps followed in this process are the following: (i). We crawl approximately 1 million products belonging to the different domains, such as food items, restaurants, electronics, furniture from the different websites together with the images of the products, and semi/un(structured) information; (ii). The domain experts manually inspected the unstructured data according to the domain information and parsed the free text in a structured format; (iii) Each domain selected was closely observed first. Then, the aspect categories were listed to mark the aspect information. The different aspect categories, along with the associated aspect terms belonging to the different domains, are listed in Table 1.

2. Creating user-agent Dialogues: The domain experts who had detailed knowledge of the respective domains along with crowd-sourced workers were employed to build goal-oriented multimodal conversations using a Wizard-of-Oz (WOZ) approach. For every conversation belonging to a particular domain, the domain experts assume the role of a system agent while the workers act like the customer agents. Different criteria for creating the conversations, such as the minimum length of the conversation, number of aspect categories, number of images in response, number of goals, number of complex requests, etc, were specified to increase the conversation diversity. At the implementation level for dialogue creation, we establish a web interface for the experts and the workers that display the instructions and different aspect categories along with the aspect terms belonging to a particular domain next to the ongoing dialogue creation. This assists the participants in creating good conversations while referring to the guidelines and the different aspects information pertaining to a domain without stopping the conversation. Though we follow a known approach (Wizard-of-oz) for data creation as done in the existing works (Budzianowski et al., 2018; Peskov et al., 2019; Saha et al., 2018), our MDMMD dataset constitutes of more varied responses belonging to the multiple domains and having both textual and visual modalities.

To the best of our knowledge, this dataset is novel in the sense that it is created in full supervision of the experts and we explicitly monitor and guide the workers to participate in the process to create engaging, informative, and diverse conversations while focusing on the different aspects of a particular product/service. For example, in the restaurant domain, participants were advised to pretend that they were either interested in ordering food or looking for a fine place to dine. The different aspects associated with this domain like the type of cuisine (Chinese, Italian, Indian, etc), type of restaurant (cafes, lounges, etc), ambience, the meal type (dinner, breakfast, etc,), type of food (desserts, snacks, appetizers, etc) are provided for creating diverse conversations. They were asked to change their preferences in between the conversations (e.g. from Chinese they could shift into the Italian foods) for making it more challenging, real, and complex. Similarly, in case of the other domains, participants were instructed to follow the guidelines and make use of the different aspect categories for creating diverse, interesting, and en-

[1] The dataset is available in https://www.iitp.ac.in/~ai-nlp-ml/resources.html#mdmmd

DATASET STATISTICS	TRAIN	VALID	TEST
Number of modalities	T+I	T+I	T+I
Number of dialogues	99813	11081	21105
Number of utterances	2086091	217187	436873
Number of image responses	1151321	107331	210997
Avg. turn per dialogue	20.9	19.6	20.7
Avg. word in textual response	12.07	11.7	11.74
Total Aspect category	85	22	45
Aspect Terms	258	87	125
Vocabulary size	45,453	-	-

Table 2: Dataset statistics of the MDMMD dataset

gaging responses.

3. Aspect Annotation: Our dataset has two kinds of annotations: Aspect category [AC] and aspect terms [AT]. Intuitively, the aspect category for a particular domain can be constant for a group of utterances but the aspect terms in every utterance may or may not be consistent. For example, the *cuisine* is the aspect category but *Chinese* is the aspect term that according to the user could change into *Mexican, Japanese* in the remaining utterances of a particular dialogue. Therefore, the labeling of both the aspect category and aspect term is essential for the generation of aspect guided responses to learn the subtle differences between the different aspect terms within the same category. By exploring the numerous internet sources used for data crawling, we compile a predefined list of aspect categories. The aspect terms for a particular category are also listed for every domain. Crowd members and experts were instructed to label the aspect categories in the interface provided for the creation of the dialogues from the predefined list along with the aspect terms contained in each utterance. The utterances with no aspect information, e.g. the starting and ending utterances of the dialogues, were marked with the *None* label to signify the absence. A group of 6 annotators was selected to verify the annotations done by the experts and the crowd workers on a set of 1500 dialogues. We observe the multi-rater Kappa agreement ratio of approximately 75%, which may be considered as a reliable estimate. Hence, from the survey, it can be concluded that the annotation done by the experts and crowd workers for both the aspect category and aspect terms were correct.

3.2 Dataset Statistics

The statistics of the complete dataset having all the three domains are provided in Table 2. The dataset is divided into train, test, and validation with 75%, 15%, and 10% conversations in each, respectively.

4 Methodology

For the proposed task, we assume that the aspect term information will be provided for the response to be generated. As different aspects are extremely subjective in a goal-oriented system, hence the responses are majorly dependent upon the respondent. Therefore, there can be several potential responses possible for a given input. Because of this subjectivity in goal-oriented systems, we like to focus on solving the task of generating responses with the desired aspect information.

Problem Definition: Our current work addresses the task of aspect guided response generation in a multi-modal goal-oriented dialog system conditioned on the conversational history having both textual and visual information. To be more specific, given an utterance $U_k = (w_{k,1}, w_{k,2}, ..., w_{k,n})$, a set of images I_k = $(i_{k,1}, i_{k,2}, ..., i_{k,j})$, and a conversational history $H_k = ((U_1, I_1), (U_2, I_2), ..., (U_{k-1}, I_{k-1}))$ and the aspect term V_a the task is to generate the next textual response $Y = (y_1, y_2, ..., y_{n'})$, where n and n' are the given input utterance and response length, respectively.

4.1 Background

Graph convolutional networks (GCNs) work on a graph structure and compute representations for the graph nodes by looking at the node's neighbourhood. Precisely, let $\mathcal{G} = (V, E)$ denote a directed graph, where V is the set of nodes (let $|V|$ = i) and E is the set of edges. The input feature matrix having i nodes is represented by $\mathcal{X} \in \mathbb{R}^{i \times j}$, whereas each node n_k ($k \in V$) is denoted by an i-dimensional feature vector. By stacking m layers of GCNs, we can account for the neighbours that are m-hops away from the current node. The hidden representation of a 1-layer GCN is a matrix $\mathcal{H} \in \mathbb{R}^{i \times p}$ where each p-dimensional representation of a node captures the interaction with its 1-hop neighbors. Multiple layers of GCNs can be stacked together to seize interactions with nodes that are several hops away. In particular, node v representation after the m^{th} layer of GCN can be formulated as:

$$h_v^{m+1} = RELU\left(\sum_{k \in \mathcal{N}(v)} (W^m_{dir(k,l)} h^m_k + b^m_{dir(k,l)})\right) \tag{1}$$

Here, h^m_k is the representation of the k^{th} node in the $(m-1)^{th}$ GCN layer and $h^1_k = n_k$; and $dir(k, l)$

illustrates whether the information flows from k to l, l to k or $k = l$; $\forall\ v \in \mathcal{V}$.

4.2 Model Description

1. **Utterance Encoder** For a given utterance U_k, we employ a bidirectional Gated Recurrent Units (Bi-GRU) (Cho et al., 2014) to encode each word $w_{k,i}$, where $i \in (1, 2, 3, ...n)$ having d-dimensional embedding vectors into the hidden representation $h_{U_k,i}$. We concatenate the last hidden representation from both the unidirectional GRUs to form the final hidden representation of a given utterance as follows:

$$\begin{aligned} \overrightarrow{h_{U_k,i}} &= GRU_{U,f}(w_{k,i}, \overrightarrow{h_{U_k,i-1}}) \\ \overleftarrow{h_{U_k,i}} &= GRU_{U,b}(w_{k,i}, \overleftarrow{h_{U_k,i-1}}) \\ h^{txt}_{U_k,i} &= [\overrightarrow{h_{U_k,i}}, \overleftarrow{h_{U_k,i}}] \end{aligned} \quad (2)$$

Now, consider the dependency parse tree of the current utterance denoted by $\mathcal{T}_G = (V_G, E_G)$. We use an utterance-specific GCN to operate on $\mathcal{T}_G$, which takes $\{h^{txt}_{U_k,i}\}^{|G|}_{i=1}$ as the input to the first GCN layer. The node representation in the m^{th} hop of the utterance specific GCN is computed as:

$$U_v^{m+1} = RELU\left(\sum_{k \in \mathcal{N}(v)} (W^m_{dir(k,l)} U^m_k + b^m_{dir(k,l)})\right) \quad (3)$$

$\forall v \in \mathcal{V}$. Here, $W^m_{dir(k,l)}$ and $b^m_{dir(k,l)}$ are the edge direction specific utterance-GCN weights and biases for the m^{th} hop and $U^1_k = U_k$.

2. **Image Encoder** A pre-trained VGG-16 (Simonyan and Zisserman, 2015) having a 16-layer deep convolutional neural network (CNN) trained on more than millions of images present in the ImageNet dataset is used for encoding the images. As a result, the network can learn rich features from a wide range of images. Here, it is also used to extract the local image representation for all the images in the dialogue turns and concatenate them together. The concatenated image vector is passed through the linear layer to form the global image context representation as given below:

$$\begin{aligned} I_{k,i} &= VGG(I_{k,i}) \\ T_k &= Concat(T_{k,1}, T_{k,2}, \ldots, T_{k,j}) \\ h^{img}_{I,k} &= ReLU(W_I T_k + b_I) \end{aligned} \quad (4)$$

where, W_I and b_I are the weight matrix and biases, respectively, which are the trainable parameters. In every turn, the maximum number of images $i \leq 6$, so in-case of only text, vectors of zeros are considered in place of image representation.

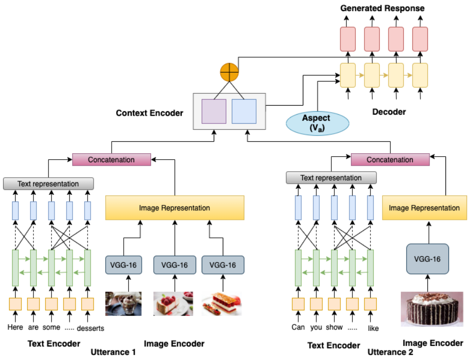


Figure 2: Architectural diagram of the proposed framework for aspect guided response generation

3. **Context Encoder** As shown in Figure 2, the final hidden representations from both image and text encoders are concatenated together for each turn and given as input to the context level GRU. A hierarchical encoder is built to model the conversational history that is placed over the text and image encoders. The decoder GRU is initialized by the final hidden state of the context encoder.

$$h^{ctx}_{c,k} = GRU_c([U_v^{m+1}; h^{img}_{I,k}], h_{c,k-1}) \quad (5)$$

where $h^{ctx}_{c,k}$ is the final hidden representation of the context for a given turn.

4. **Decoder** In the decoding section, we build another GRU for generating the response in a sequential manner based on the context hidden representation of the hierarchical encoder (context GRU), and the words decoded previously. We use the input feeding decoding along with the attention (Luong et al., 2015) mechanism for enhancing the performance of the model. Using the decoder state $h^{dec}_{d,t}$ as the query vector, we apply self-attention on

the hidden representation of the context-level encoder. The decoder state and the context vector are concatenated and used to calculate a final distribution of the probability over the output tokens.

$$\begin{aligned} h_{d,t}^{dec} &= GRU_d(y_{k,t-1}, h_{d,t-1}) \\ c_t &= \sum_{i=1}^{k} \alpha_{t,i} h_{c,k}^{ctx}, \\ \alpha_{t,i} &= softmax(h_{c,k}^{ctx\,T} W_f h_{d,t}) \\ \tilde{h}_t &= tanh(W_{\tilde{h}}[h_{d,t}; c_t]) \\ P(y_t/y_{<t}) &= softmax(W_V \tilde{h}_t) \end{aligned} \tag{6}$$

where, W_f, W_V and $W_{\tilde{h}}$ are the trainable weight matrices.

For generating responses with the specified aspects as shown in Figure 2, we provide the aspect term embedding V_a as input during decoding at every decoder time-step. In order to include the aspect vector in the decoder, we modify Equation (6) to incorporate the aspect information for the generation of responses and the modified equation is as follows:

$$h_{d,t}^{dec} = GRU_d(y_{k,t-1}, [h_{d,t-1}, V_a]) \tag{7}$$

5. **Training and Inference** We employ commonly used teacher forcing (Williams and Zipser, 1989) algorithm at every decoding step to minimize the negative log-likelihood on the model distribution. We define $y^* = \{y_1^*, y_2^*, \ldots, y_m^*\}$ as the ground-truth output sequence for a given input by:

$$\mathcal{L}_{ml} = -\sum_{t=1}^{m} \log p(y_t^* | y_1^*, \ldots, y_{t-1}^*) \tag{8}$$

We apply uniform label smoothing(Szegedy et al., 2016) to alleviate the common issue of low diversity in dialogue systems, as suggested in (Jiang and de Rijke, 2018).

4.3 Baseline Models

Model 1 (HRED): The first baseline is a simple hierarchical encoder-decoder framework that makes use of only textual information for generating the responses.

Model 2 (MHRED): The second baseline model is the extension of the HRED framework, where we incorporate the multi-modal information i.e., the images for the generation of coherent responses.

Model 3 (HRED + Aspect): In this model at the decoder side, instead of only textual conversational information we add the desired aspect at the decoder side for generating aspect controlled responses.

Model 4 (MHRED + Aspect): To learn the aspect information at the decoder we provide the aspect information to the decoder along with the text and the visual representation.

5 Experimental Details

In this section we present the details of the experimental setup and evaluation metrics.

Implementation details All the implementations were done using the PyTorch[2] framework. For all the models including baselines, the batch size is set to 32. The utterance encoder is a bidirectional GRU with 600 hidden units in each direction. We use the dropout(Srivastava et al., 2014) with probability 0.45. During decoding, we use a beam search with beam size 10. The model is initialized with the parameters chosen randomly using a Gaussian distribution with the Xavier scheme (Glorot and Bengio, 2010). The hidden size for all the layers is 512. AMSGrad (Reddi et al., 2019) is used as the optimizer for model training to mitigate the slow convergence issues. We use uniform label smoothing with $\epsilon = 0.1$ and perform gradient clipping when the gradient norm is above 5. We use 300-dimensional word-embedding initialized with Glove (Pennington et al., 2014) embedding pre-trained on Twitter. We consider the previous 2 turns for the dialogue history, and the maximum utterance length is set to 50. For image representation, FC6(4096 dimension) layer representation of the VGG-19 (Simonyan and Zisserman, 2015), pre-trained on ImageNet is used.

Automatic evaluation metrics To evaluate our proposed framework at the content level we report Perplexity (Chen et al., 1998). Lesser perplexity scores signify that the generated responses are grammatically correct and fluent. We also report the results using the standard metrics like BLEU-4 (Papineni et al., 2002) and Rouge-L (Lin, 2004) to measure the quality of the generated response for capturing the correct information.

Human evaluation metrics From the generated responses we randomly take 700 responses from

[2] https://pytorch.org/

the test dataset for qualitative evaluation. For a given input along with aspect information, three annotators with post-graduate exposure were assigned to evaluate the correctness, relevance, domain and aspect consistency of the generated responses by the different approaches for the following four metrics: (i) Fluency (F): This metric is used to measure the grammatical correctness of the generated response. It checks that the response is fluent and does not contain any errors; (ii). Relevance (R): It is used to judge whether the generated response is relevant to the conversational history; (iii). Aspect Appropriateness (AP): For this metric, we take care of the fact that the response generated is in consonance to the specified aspect (e.g. cuisine, color, type, etc) and is also coherent to the conversational history; (iv). Domain Consistency (DC): This metric is used to measure the consistency of the generated response in accordance with the domain being discussed. For the human evaluation metrics, we calculate the Fleiss' kappa (Fleiss, 1971) to determine the inter-rater consistency. For fluency and relevance, the kappa score is 0.75, and for aspect appropriateness and domain consistency is 0.77, indicating substantial agreement.

6 Results and Discussion

In this section we report the evaluation results along with the necessary analysis and discussions on these.

Automatic evaluation results: Evaluation results using automatic evaluation metrics are provided in Table 3. From the table, it is clear that the proposed approach outperforms all the baseline models and these improvements are statistically significant [3].

Model Description		**Perplexity**	**BLEU-4**	**Rouge-L**
Baseline Approaches	HRED	1.0385	0.5078	0.5155
	MHRED	1.0274	0.5236	0.5387
	HRED + Aspect	1.0249	0.5195	0.5298
	MHRED + Aspect	1.0211	0.5308	0.5419
Proposed Approach	**T-GCN**	**1.0186**	**0.5687**	**0.5712**
	M-GCN	**1.0137**	**0.5871**	**0.5925**
	M-GCN + Aspect	**1.0112**	**0.6014**	**0.6105**

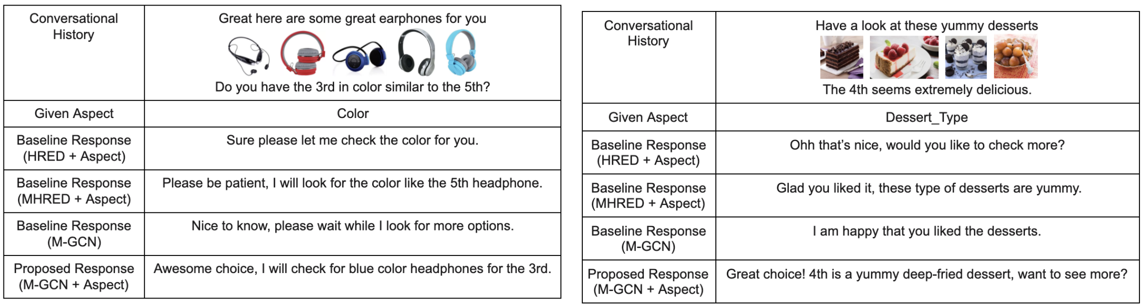
Table 3: Results of different baselines and the proposed model on the MDMMD dataset

As lower the perplexity better is the generated responses, hence, it is visible that the perplexity scores of the proposed *M-GCN + Aspect* model are the lowest among all the baseline models. As opposed to the text-based models, multi-modal frameworks, such as *MHRED* and *M-GCN* have lower scores for the perplexity exhibiting improvement in performance. The reduction in perplexity scores for the aspect guided models both text-based and multimodal frameworks further to ensure the robustness of these models for generating better responses. In the case of the BLEU-4 metric, we see that the proposed model *M-GCN + Aspect* having the ability to generate responses according to the specified aspect information achieves higher scores with an improvement of *6.2%* from the *MHRED + Aspect* baseline model. The superior performance establishes the fact that the proposed model generates correct responses while preserving the information present in the ground-truth response as BLEU-4 compares the generated response to the ground-truth. Similarly, in the case of Rouge-L, there is an increase of *6.12%* in comparison to the multimodal HRED framework. The significant jump in the performance entitles the fact that images play a crucial role in generating contextually correct responses. As our research focus is on aspect-guided response generation in multi-modal dialogue systems, we see that the frameworks having aspect information outperforms the other baseline models.

Human evaluation results: Along with the automatic evaluation, human evaluation is also essential for assessing the quality of the responses. Hence, for our specified task of generating responses in a multi-modal setup, we evaluate the baseline and our proposed model with the human evaluation metrics as mentioned. In Table 4, we present the results of human evaluation for all the baselines and the proposed model. The fluency scores of the baseline *HRED* model are the lowest for grammatically correct responses due to repetition and incomplete responses. The current work revolves around the aspect, hence the generated responses are assessed according to the specified aspects. It is evident from the results that the proposed framework generates responses that are appropriate to the specified aspects with an improvement of *8.46%* from the *MHRED + Aspect* based baseline.

The improvement in the proposed model with aspect information provided additionally is significantly higher compared to the other methods. This is majorly due to the following facts: very precise and fine-grained information in the form of aspects of the products and/or services, better memory re-

[3] we perform statistical significance t-test (Welch, 1947) and it is conducted at 5% (0.05) significance level



Conversational History	Great here are some great earphones for you Do you have the 3rd in color similar to the 5th?
Given Aspect	Color
Baseline Response (HRED + Aspect)	Sure please let me check the color for you.
Baseline Response (MHRED + Aspect)	Please be patient, I will look for the color like the 5th headphone.
Baseline Response (M-GCN)	Nice to know, please wait while I look for more options.
Proposed Response (M-GCN + Aspect)	Awesome choice, I will check for blue color headphones for the 3rd.

(a) Example 1

Conversational History	Have a look at these yummy desserts The 4th seems extremely delicious.
Given Aspect	Dessert_Type
Baseline Response (HRED + Aspect)	Ohh that's nice, would you like to check more?
Baseline Response (MHRED + Aspect)	Glad you liked it, these type of desserts are yummy.
Baseline Response (M-GCN)	I am happy that you liked the desserts.
Proposed Response (M-GCN + Aspect)	Great choice! 4th is a yummy deep-fried dessert, want to see more?

(b) Example 2

Figure 3: Generated examples from different models

Model Description		**Fluency**			**Relevance**		**Aspect Appropriateness**		**Domain Consistency**	
		0	**1**	**2**	**0**	**1**	**0**	**1**	**0**	**1**
Baseline Approaches	HRED	34.36	35.83	29.81	55.93	44.07	42.20	57.80	47.73	52.27
	MHRED	32.11	36.71	31.18	52.56	47.44	40.14	59.86	44.14	55.86
	HRED + Aspect	31.85	35.43	32.72	54.33	45.67	39.88	60.12	44.39	55.61
	MHRED + Aspect	29.55	36.88	33.57	53.49	46.51	35.31	64.69	43.64	56.36
Proposed Approaches	**T-GCN**	**24.92**	**37.77**	**37.31**	**47.13**	**52.87**	**30.17**	**69.83**	**37.01**	**62.99**
	M-GCN	**20.87**	**39.17**	**39.96**	**44.72**	**55.28**	**27.97**	**72.03**	**34.28**	**65.72**
	M-GCN + Aspect	**19.64**	**39.65**	**40.71**	**43.56**	**56.44**	**26.85**	**73.15**	**33.85**	**66.15**

Table 4: Results of human evaluation

tention capability of the networks which generate responses that are consistent with the domain, and the multi-modal sources of information (text and image). From the human evaluation, it can be concluded that the generated responses are not only fluent and relevant but also consistent with the domain and the specified aspect information.

Error analysis: To gain better insights, we closely analyze the outputs generated from our proposed system, and observe the following error scenarios: (i). **Loss of information:** The uni-modal baselines such as HRED generate responses that lack complete information. Gold: *Here are the chairs in yellow color as in the 3rd image but not in round shapes as in the 5th image.*; Predicted: *The chairs are here but* $< unk >$ *not in the shape.* This indicates that the unavailability of multi-modal information (in this case, images) leads to the loss of information in the generated response. (ii). **Contextually wrong domain:** In some cases, our proposed framework generates the responses that are contextually incorrect with the domain. For example, with the aspect *color* the response generated belongs to the electronics domain, but the actual domain in the discussion is a restaurant. This type of error occurs due to the higher number of utterances with the color aspect belonging to the electronics domain in contrast to the restaurant domain. (iii). **Mistakes in image identification**: The baseline and proposed frameworks in some cases confuse the images being discussed leading to generating incorrect responses. As an example, Gold:*I have beverages to go with the 2nd image but it is similar to the 4th one.*; Predicted: *I have got you beverages to go with the 4th image but nothing like the 3rd one.* This indicates the model's inability to capture the correct positional information of the images. Also, the mention of different images in the contextual information confuses the model in selecting the correct images.

7 Conclusion and Future Work

Our current work emphasizes on the task of generating aspect-guided responses in a multi-modal dialogue system. We create a large scale task-oriented MDMMD dataset comprising of dyadic dialogues. The dataset comprises of three different domains, such as restaurant, electronics, and furniture. We develop a GCN based method to capture the textual representation, while we use VGG-19 for image representation. The context encoder captures the multi-modal information from the utterances. The representation from the context encoder along with the aspect vector is fed to the decoder for generating the aspect-guided responses. Experimental results show that our proposed methodology outperforms the baseline models in the case of both automatic and human evaluation metrics.

In future along with enhancing the architectural design of our proposed methodology, we would also like to investigate methods for image retrieval for complete multi-modal response generation. Furthermore, we would extend our method to deal with multiple aspects present in an utterance and generate the responses accordingly.

Acknowledgement

Authors duly acknowledge the support from the Project titled Sevak-An Intelligent Indian Language

Chatbot, Sponsored by SERB, Govt. of India (IMP/2018/002072). Asif Ekbal acknowledges the Young Faculty Research Fellowship (YFRF), supported by Visvesvaraya Ph.D. scheme for Electronics and IT, Ministry of Electronics and Information Technology (MeitY), Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

References

Shubham Agarwal, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018a. Improving context modelling in multimodal dialogue generation. In *Proceedings of the 11th International Conference on Natural Language Generation, Tilburg University, The Netherlands, November 5-8*, pages 129–134.

Shubham Agarwal, Ondřej Dušek, Ioannis Konstas, and Verena Rieser. 2018b. A knowledge-grounded multimodal search-based conversational agent. In *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31*, pages 59–66.

Huda Alamri, Chiori Hori, Tim K Marks, Dhruv Batr, and Devi Parikh. 2018. Audio visual scene-aware dialog (avsd) track for natural language generation in dstc7. In *DSTC7 at AAAI2019 Workshop*, volume 2.

Paweł Budzianowski and Ivan Vulić. 2019. Hello, it's gpt-2–how can i help you? towards the use of pretrained language models for task-oriented dialogue systems. *arXiv preprint arXiv:1907.05774*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Inigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5016–5026.

Hardik Chauhan, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Ordinal and attribute aware response generation in a multimodal dialogue system. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5437–5447.

Stanley Chen, Douglas H. Beeferman, and Ronald Rosenfeld. 1998. Evaluation metrics for language models.

Wenhu Chen, Jianshu Chen, Pengda Qin, Xifeng Yan, and William Yang Wang. 2019a. Semantically conditioned dialog response generation via hierarchical disentangled self-attention. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3696–3709.

Xiuyi Chen, Jiaming Xu, and Bo Xu. 2019b. A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2687–2693.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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 2019, Paris, France, July 21-25*, pages 445–454.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26*, pages 4466–4475.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Zhe Gan, Yu Cheng, Ahmed EI Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6463–6474.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2010, Chia Laguna Resort, Sardinia, Italy, May 13-15*, pages 249–256.

Shaojie Jiang and Maarten de Rijke. 2018. Why are sequence-to-sequence models so dull? understanding the low-diversity problem of chatbots. *Proceedings of the 2nd International Workshop on Search-Oriented Conversational AI, SCAI@EMNLP 2018, Brussels, Belgium, October 31, 2018*, pages 81–86.

Hung Le, S Hoi, Doyen Sahoo, and N Chen. 2019. End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In *DSTC7 at AAAI2019 workshop*.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26*, pages 801–809. ACM.

Chin-Yew Lin. 2004. Rouge: a package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*.

Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. 2019a. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. In *DSTC7 at AAAI2019 workshop*.

Zehao Lin, Xinjing Huang, Feng Ji, Haiqing Chen, and Ying Zhang. 2019b. Task-oriented conversation generation using heterogeneous memory networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7*, pages 4557–4566.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv:1508.04025*.

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1468–1478.

Fei Mi, Minlie Huang, Jiyong Zhang, and Boi Faltings. 2019. Meta-learning for low-resource natural language generation in task-oriented dialogue systems. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16*, pages 3151–3157.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 462–472.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1532–1543.

Denis Peskov, Nancy Clarke, Jason Krone, Brigi Fodor, Yi Zhang, Adel Youssef, and Mona Diab. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Hong Kong, China, November 3-7, 2019*, pages 4518–4528.

Kun Qian and Zhou Yu. 2019. Domain adaptive dialog generation via meta learning. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2639–2649.

Dinesh Raghu and Nikhil Gupta. Hierarchical-pointer generator memory network for task oriented dialog.

Dinesh Raghu, Nikhil Gupta, et al. 2018. Disentangling language and knowledge in task-oriented dialogs. *arXiv preprint arXiv:1805.01216*.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of adam and beyond. *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*.

Revanth Reddy, Danish Contractor, Dinesh Raghu, and Sachindra Joshi. 2019. Multi-level memory for task oriented dialogs. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3744–3754.

Amrita Saha, Mitesh M Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7*, pages 696–704.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*, pages 3776–3784.

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 Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 3295–3301.

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

Pushkar Shukla, Carlos Elmadjian, Richika Sharan, Vivek Kulkarni, Matthew Turk, and William Yang Wang. 2019. What should i ask? using conversationally informative rewards for goal-oriented visual dialog. *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6442–6451.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International Conference on Information and Knowledge Management, CIKM 2015, Melbourne, VIC, Australia, October 19 - 23*, pages 553–562. ACM.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 2818–2826.

Junya Takayama and Yuki Arase. 2019. Relevant and informative response generation using pointwise mutual information. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 133–138.

Zhiliang Tian, Wei Bi, Xiaopeng Li, and Nevin L Zhang. 2019. Learning to abstract for memory-augmented conversational response generation. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3816–3825.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Bernard L Welch. 1947. The generalization ofstudent's' problem when several different population variances are involved. *Biometrika*, 34(1/2):28–35.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Chien-Sheng Wu. 2019. Learning to memorize in neural task-oriented dialogue systems. *arXiv preprint arXiv:1905.07687*.

Haotian Xu, Haiyun Peng, Haoran Xie, Erik Cambria, Liuyang Zhou, and Weiguo Zheng. 2019. End-to-end latent-variable task-oriented dialogue system with exact log-likelihood optimization. *World Wide Web*, pages 1–14.