Knowledge Grounded Multimodal Dialog Generation in Task-oriented Settings

Deeksha Varshney, Anushkha Singh, Asif Ekbal
Department of Computer Science and Engineering
Indian Institute of Technology Patna
{1821cs13, anushkha_1901me72, asif}@iitp.ac.in

Abstract

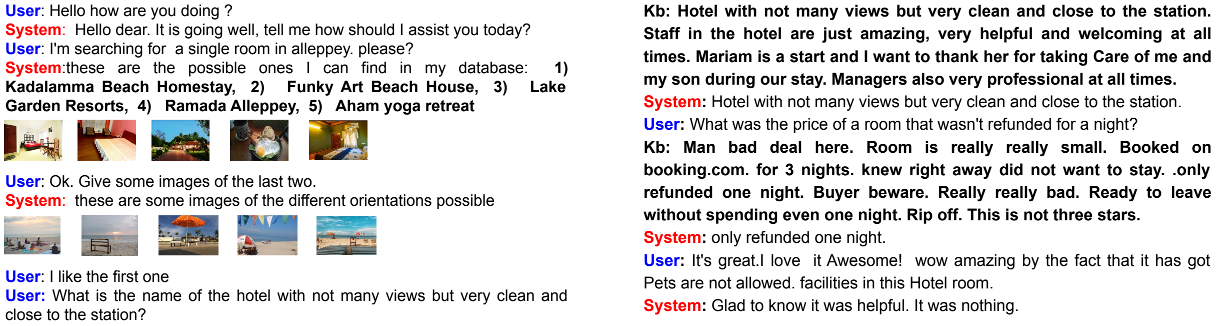
Knowledge-grounded dialogue generation is the process of formulating an informed response based on both the conversation context and external knowledge. Multi-modality in dialogue systems has paved the way for more robust conversational bots. Any multimodal system seeks to bridge the gap between language and vision by combining information from image, audio, video, and text, all of which are often complementary. In every task-oriented dialog system, product or service attributes are crucial for satisfying the user's needs. We propose the task of knowledge grounded response generation in a multimodal task-oriented dialog setting. We employ a multi-modal knowledge grounded incremental transformer network to generate responses that combine information from context, visuals, and external knowledge. We prepare a Knowledge Grounded Multi-Modal Dialog (KGMMD) dataset to help users in more efficient decision making. Dialog utterances, accompanying images, and knowledge in the form of hotel reviews from the hotel domain are all included in the dataset. The newly formed KGMMD dataset has been subjected to both quantitative and qualitative analysis. Evaluation results on the KGMMD datasets show that the proposed methodology outperforms the baseline models for knowledge grounded multimodal response generation.

1 Introduction

Conversational agents now have more possibilities thanks to advances in artificial intelligence (AI). Human-machine interaction is an important application of artificial intelligence that assists people in their daily lives. AI advancements have resulted in the invention of personal assistants such as Apple's Siri, Amazon's Alexa, and Microsoft's Cortana, which aid people in their daily tasks. Growing demands in numerous industries such as travel, entertainment, shopping, and so on necessitate conversational agents to communicate by merging information from various modalities to develop a robust system.

A multimodal dialogue system allows for interdisciplinary research by bridging the gap between visual and text information. Data from several dimensions, such as text, image, audio, and video, has been shown to offer all of the details required to build effective end-to-end dialogue systems (Le et al., 2019a; Saha et al., 2018). Multimodal conversational systems supplement existing dialogue systems by supplying information that unimodal systems lack. Several recent efforts on multimodal dialogue systems (Chauhan et al., 2019; Das et al., 2017; Le et al., 2019b) have spurred interest in this field by combining data from a variety of sources, including texts, and images.

Many natural language generation (NLG) tasks require semantic comprehension, especially when it is supplemented by real-world data. Building a good robust dialog system requires integrating the necessary subjects, entities, and relations mentioned in the user's statement while also providing an acceptable answer (Wu et al., 2019; Agarwal et al., 2018b). It is critical to obtain precise information about the service requested by the user in order to accomplish



User: Hello how are you doing ?
System: Hello dear. It is going well, tell me how should I assist you today?
User: I'm searching for a single room in alleppey. please?
System:these are the possible ones I can find in my database: **1) Kadalamma Beach Homestay, 2) Funky Art Beach House, 3) Lake Garden Resorts, 4) Ramada Alleppey, 5) Aham yoga retreat**
User: Ok. Give some images of the last two.
System: these are some images of the different orientations possible
User: I like the first one
User: What is the name of the hotel with not many views but very clean and close to the station?

(a)

Kb: Hotel with not many views but very clean and close to the station. Staff in the hotel are just amazing, very helpful and welcoming at all times. Mariam is a start and I want to thank her for taking Care of me and my son during our stay. Managers also very professional at all times.
System: Hotel with not many views but very clean and close to the station.
User: What was the price of a room that wasn't refunded for a night?
Kb: Man bad deal here. Room is really really small. Booked on booking.com. for 3 nights. knew right away did not want to stay. .only refunded one night. Buyer beware. Really really bad. Ready to leave without spending even one night. Rip off. This is not three stars.
System: only refunded one night.
User: It's great.I love it Awesome! wow amazing by the fact that it has got Pets are not allowed. facilities in this Hotel room.
System: Glad to know it was helpful. It was nothing.

(b)

Figure 1: An example conversation from the KGMMD dataset.

the user's goals or objectives. For example, in the hotel domain, the conversational agent must know the pricing range, hotel location, number of people, and so on in order to provide the best accessible alternative to the user. We observe that these extra information can be provided by hotel reviews.

In Figure 1, we show some samples of dialogs from the newly constructed KGMMD dataset that are related to the hotel domain. The figure demonstrates that graphics as well as textual information, are equally vital in any task-oriented system. From the example, we may note that incorporation of customer reviews leads to a more diverse response and engages the user in a long-term conversation.

In summary, the contributions of our current work are as follows:

1. We present a novel system for knowledge grounded dialog generation in a goal-oriented multimodal setting.
2. We create a large Knowledge Grounded Multi-Modal Dialogue (KGMMD) dataset comprising of dialogs with textual and image information. This dataset contains conversations from the hotel domain with every conversation having images and review based external knowledge to facilitate efficient response generation.
3. We propose to use a multi-modal knowledge grounded incremental transformer framework for generating responses conditioned on the both images and external knowledge.
4. We observe from the automatic and manual evaluation metrics that our proposed framework can effectively capture the information from the associated knowledge and images provided in the newly created KGMMD dataset.

2 Related Work

With the rapid growth of Artificial Intelligence (AI), there is a growing trend toward designing multimodal dialogue systems that use images, audio, and video in addition to text. Deep neural models are highly good at modelling dialogues, as demonstrated in (Shang et al., 2015; Vinyals and Le, 2015). The authors of (Sordoni et al., 2015) presented a hierarchical structure capable of maintaining prior information to capture the context of the user's previous queries. Similarly, the hierarchical encoder-decoder framework was examined (Serban et al., 2016; Serban et al., 2017) in order to preserve the dependencies among utterances in a dialogue. In (Xu et al., 2019), the authors enhanced the hierarchical encoder-decoder framework by including a latent variable for interpreting the goals of discussions in a task-oriented dialogue system. Hierarchical pointer networks (Golchha et al., 2019) have also been used to generate responses in task-oriented dialogues.

In order to create a robust system, researchers are concentrating more towards merging several kinds of information, such as audio, images, video, and text. The study detailed in (Das et al., 2017; De Vries et al., 2017; Gan et al., 2019; Mostafazadeh et al., 2017) has helped to bridge the gap between vision and language. With the release of the DSTC7 dataset in (Le et al., 2019b), which used a multimodal transformer in order to represent the combined information from multiple modalities, such as video and text. Similarly, the DSTC7 dataset was used for creation in (Le et al., 2019a; Lin et al., 2019; Alamri et al., 2018) by combining audio and visual elements. For building multimodal dialogue systems, recently a Multimodal dialogue (MMD) dataset (Saha et al., 2018), containing dialogues in the fashion domain was released. On the MMD dataset (Agarwal et al., 2018a; Agarwal et al., 2018b; Liao et al., 2018) used the hierarchical encoder-decoder paradigm by using information from text, images, and external knowledge bases. (Chauhan et al., 2019) recently proposed position-aware and attribute-aware attention for textual response generation. The authors of (Cui et al., 2019) generated responses on the MMD dataset using a hierarchical attention technique.

Memory networks (Sukhbaatar et al., 2015) were first employed to handle dialogue state and knowledge together (Bordes et al., 2017; Meng et al., 2018), and subsequently they were enhanced to include a copying mechanism (Madotto et al., 2018; Wu et al., 2019). For task-oriented dialogue generation, knowledge-based end-to-end memory networks have been established (Raghu et al., 2019; Reddy et al., 2019; Chen et al., 2019; Wang et al., 2020) using multi-level, working, and dynamic types of memory. In DDMN (Wang et al., 2020), a memory manager is used to dynamically track the flow of history information during conversations to retain the important parts from both the dialogues and KB.

The authors of (Wu et al., 2019) employed a encoder and a local decoder to exchange external knowledge in task-oriented dialogue settings. (Tian et al., 2019) examines a memory-augmented architecture capable of extracting critical information during training for better response generation. (Zhang et al., 2021) create a knowledge-grounded dialog system by making use of state-of-the-art pretrained GPT-2 language model. (Agarwal et al., 2018b) built Text-only HRED (T-HRED) encoder-decoder model with context and knowledge extensions. They devised a mechanism to ground textual responses on knowledge (KB).

In this work, we demonstrate how appropriate knowledge can improve the task of multimodal dialogue generation. Our work focuses on creating knowledge grounded replies in a multimodal context using input from both text and images using a neural conversational model named MKGITN. Our method allows use of a large-scale knowledge base and exploits visual features to facilitate both the understanding of an utterance and the generation of a response with a novel multi-modal knowledge grounded incremental transformer model.

3 Methodology

3.1 Problem Statement

The aim of this work is to incorporate relevant reviews of hotels as knowledge base and associated images with the input utterances into the multi-turn conversations to generate contextually relevant textual responses. Let a whole dialogue containing K utterances be represented by $\mathbf{U} = \mathbf{u}^{(1)},...,\mathbf{u}^{(k)},...,\mathbf{u}^{(K)}$. $\mathbf{u}^{(k)}$ is the k-th utterance containing I words $(u_1^{(k)},...,u_i^{(k)},...,u_I^{(k)})$. There is a knowledge base $\mathbf{b}^{(k)}$ associated with every utterance containing J words $(b_1^{(k)},...,b_j^{(k)},...,b_J^{(k)})$. Similarly, for every utterance $\mathbf{u}^{(k)}$, there are images also associated with the k-th utterance $\mathbf{g}^{(k)} = g_1^{(k)},...,g_m^{(k)},...,g_M^{(k)}$ which represents the image base related to the k-th utterance containing M images. Our aim is to generate a response $\mathbf{u}^{(k+1)}$ for the task of grounded multimodal conversations given the previous k utterances $\mathbf{U}_{\leq k}$ with related knowledge base $\mathbf{B}_{\leq k}$ and image base $\mathbf{G}_{\leq k}$.

As a result, the likelihood of generating target response $\mathbf{u}^{(k+1)}$ is computed as:

$$\mathrm{P}(\mathbf{u}^{(k+1)}|\mathbf{U}_{\leq k},\mathbf{B}_{\leq k},\mathbf{G}_{\leq k};\theta) = \prod_{i=1}^{I} \mathrm{P}(u_i^{(k+1)}|\mathbf{U}_{\leq k},\mathbf{B}_{\leq k},\mathbf{G}_{\leq k},\mathbf{u}_{<i}^{(k+1)};\theta)$$

3.2 Self-Attention based utterance and knowledge encoder

We use a self-attention based encoder (Vaswani et al., 2017) to encode the knowledge sentence and current utterance independently. We use multi-head attention to estimate the features of the knowledge base, $\mathbf{b}^{(k)}$. The encoder receives a sequence of knowledge base word embeddings along with positional embedding as input.

$$\mathbf{In}_b^{(k)} = [b_1^{(k)}, ..., b_J^{(k)}] \tag{1}$$

$$b_j^{(k)} = \mathbf{e}_{b_j} + PE(j) \tag{2}$$

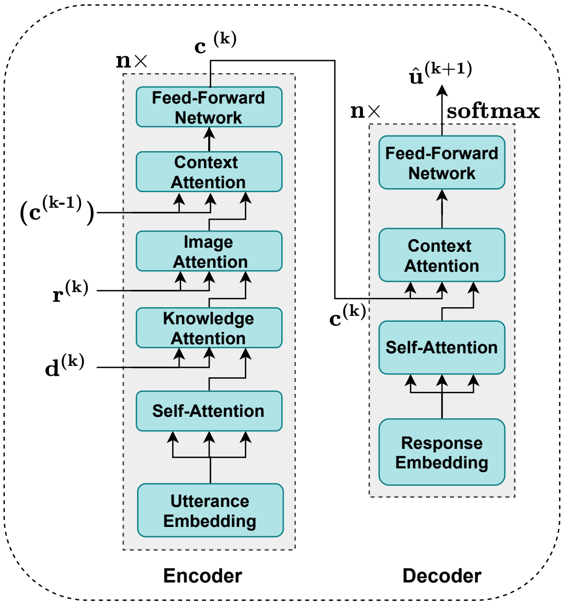


Figure 2: MKGITN architecture

where $\mathbf{e}_{b_j}$ is the word embedding of $b_j^{(k)}$ and PE(.) denotes the positional encoding function. There are two sublayers in each layer. A multi-head self-attention is the initial sub-layer. The second sub-layer is a fully connected feed-forward network. This FFN is made up of two linear transformations separated by a ReLU activation.

$$\mathbf{A}^{(n)} = MultiHead(\mathbf{In}_b^{(k)}, \mathbf{In}_b^{(k)}, \mathbf{In}_b^{(k)}) \quad (3)$$

$$\mathbf{B}^{(n)} = FFN(\mathbf{A}^{(n)}) \quad (4)$$

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (5)$$

$$\mathbf{d}^{(k)} = \mathbf{B}^{(n)} \quad (6)$$

3.3 Image Encoder

To extract image features for all of the images in a given dialogue turn, a pre-trained VGG-19 (Simonyan and Zisserman, 2014) model is employed. To produce a global image context representation, the concatenation of single image features is fed into a single linear layer.

$$Img_m^{(k)} = \text{VGG}(g_m^{(k)}) \quad (7)$$

$$Img_c = \text{concat}(Img_1^{(k)}, Img_2^{(k)}, ..., Img_m^{(k)}) \quad (8)$$

$$Img_f = \text{ReLU}(\mathbf{W}(Img_c) + \textbf{bias}) \quad (9)$$

$$\mathbf{r}^{(k)} = Img_f \quad (10)$$

where **W** and **bias** are the trainable weight matrix and biases, respectively. Since there are only 5 photos in a single turn, zero vectors are evaluated in the absence of images.

3.4 Multi-modal knowledge grounded incremental transformer network

The multi-modal knowledge grounded incremental transformer network is a unified transformer encoder that encodes multi-turn utterances using an incremental encoding approach. We use a transformer encoder to encode the multi-turn knowledge and image grounded utterances. The knowledge features, image information, and context are incorporated into the current utterance's features using multi-head attention. The architecture is shown in Figure 2.

Each layer is divided into five sub-layers. A multihead self-attention on the output of the last layer is the first sub-layer:

$$\mathbf{C}^{(n)} = MultiHeadAttn(\mathbf{S}^{(n-1)}, \mathbf{S}^{(n-1)}, \mathbf{S}^{(n-1)}) \quad (11)$$

where, $\mathbf{S}^{(0)} = \mathbf{In}_u^{(k)}$. We apply a multi-head attention on the knowledge base representation in the second sub-layer. The third sub-layer is multi-head attention over the image representation. A multi-head context attention is applied in the fourth sub-layer:

$$\mathbf{D}^{(n)} = MultiHeadAttn(\mathbf{C}^{(n)}, \mathbf{d}^{(k)}, \mathbf{d}^{(k)}) \quad (12)$$

$$\mathbf{E}^{(n)} = MultiHeadAttn(\mathbf{D}^{(n)}, \mathbf{r}^{(k)}, \mathbf{r}^{(k)}) \quad (13)$$

$$\mathbf{F}^{(n)} = MultiHeadAttn(\mathbf{E}^{(n)}, \mathbf{c}^{(k-1)}, \mathbf{c}^{(k-1)}) \quad (14)$$

where $\mathbf{c}^{(k-1)}$ is the representation of the previous utterances. The fifth sub-layer is a position-wise fully connected feed-forward network:

$$\mathbf{S}^{(n)} = FFN(\mathbf{F}^{(n)}) \quad (15)$$

$$\mathbf{c}^{k} = \mathbf{S}^{n} \quad (16)$$

3.5 Decoder

Similarly, the transformer decoder is used to decode the responses. The encoder and decoder layers are nearly identical, but the decoder layer now incorporates two multi-head attention layers for self-attention and encoder-decoder attention, respectively.

$$\mathbf{In}_r^{(k+1)} = [u_1^{(k+1)}, ..., u_{i-1}^{(k+1)}] \tag{17}$$

$$\mathbf{P}^n = MultiHeadAttn(\mathbf{Q}^{(n-1)}, \mathbf{Q}^{(n-1)}, \mathbf{Q}^{(n-1)}) \tag{18}$$

where, $\mathbf{Q}^{(0)}$ = $\mathbf{In}_r^{(k+1)}$.

$$\mathbf{R}^{(n)} = MultiHeadAttn(\mathbf{P}^{(n)}, \mathbf{c}^k, \mathbf{c}^k) \tag{19}$$

$$\mathbf{Q}^{(n)} = FFN(\mathbf{R}^{(n)}) \tag{20}$$

To predict the next word, we utilize softmax to get the probabilities of the words decoded by the decoder.

$$P(\hat{\mathbf{u}}^{(k+1)}) = softmax(\mathbf{Q}^{(n)}) \tag{21}$$

3.6 Training Loss

$$Loss = -\sum_{i=1}^{m} logP(\hat{\mathbf{u}}^{(k+1)}) \tag{22}$$

4 Datasets and Experiments

We give the details of the datasets used in our experiments, as well as a full summary of the experimental parameters.

4.1 Dataset

One of the primary contributions of this work, as previously noted, is a large-scale Knowledge Grounded Multimodal Dialog dataset (KGMMD). The dataset contains dialogues with textual and visual information, as well as content related knowledge in the sequence of interactions. To assist customers in making informed judgments, relevant visuals relating to the products are used. The underlying knowledge comprises of reviews associated with every product mentioned in the conversation. This dataset will aid in the development of powerful multimodal goal-oriented discussion systems. KGMMD is an extension of Multi-domain Multi-modal Dialogue (MDMMD) dataset (Firdaus et al., 2021; Firdaus et al., 2020) having text-based and image-based conversations. Figure 1 in the introduction section provides a few samples from our freshly constructed dataset. In the following section, we provide the details of corpus creation, its statistics, and quality analysis, accompanied by references to the current goal-oriented conversation to demonstrate the use and relevance of our work.

4.2 Process for Data Collection

Clients reserving hotels for various events such as excursions, conferences, parties, and so on are the focus of conversations in the hotel domain. The customer will need to see images of the rooms in order to choose the best one. When choosing a room, customers see at the different facilities such as a T.V., fan and an air cooler etc. This information is immediately evident through the photographs, which can assist the consumer in selecting the best accommodations for themselves.

In addition to the images, we augment the conversation with a grounded pair of utterance and response. We initially extract all of the hotel names from the dataset and then crawl their reviews[1]. We then use a pre-trained question creation technique to construct questions for all of the collected reviews[2]. All of the generated queries are transformed into user utterances. We use cosine similarity to select the most comparable sentence from the review to the produced question and assign it as the response for the related system response. Lastly, we give the matching review to the utterance-response pair as knowledge base (kb). In total, two such utterance-response pairs have been added to the original dataset. Finally, we use three human annotators to see if the augmented utterance-response combinations were out of context from the previous dialogue.

4.3 Dataset Statistics

The statistics for the complete dataset are shown in Table 1. In the table, we show the total number of dialogues in the entire dataset, which includes both written and visual responses. The dataset is divided into three sections: train, test, and validation, comprising of 80%, 10%, and 10% of the conversation in each segment. Table 1 shows the details of the number of utterances, average turn per dialogue, as well as the average length of utterances.

---

[1] tripadvisor.com

[2] https://github.com/patil-suraj/question_generation

	Train	**Valid**	**Test**
#Conversation	38,361	4,795	4,794
#Utterances	9,80,420	1,22,551	1,22,531
Avg. length of utterances	14.820	14.693	14.725
Avg. # of Turns	10.779	10.782	10.769

Table 1: Dataset details

4.4 Experimental setup and Evaluation metrics

In this section, we describe the implementation details, baselines and the evaluation metrics.

4.5 Implementation Details

We have used the existing pytorch based code framework by (Agarwal et al., 2018a) for our experiments. The encoder and decoder layers are both set to six, with 512-dimensional hidden states and a dropout of 0.1. In the feed-forward hidden layers, there are 8 multi-head attention heads and 2048 nodes. The word embedding dimension is empirically set to 512. For optimization, we use Adam (Kingma and Ba, 2014). VGG-19's 4096-dimensional FC6 layer image representations were used.

4.6 Baseline Models

To demonstrate the utility of our model, we compare it to the following benchmarks:

1. **HRED:** The hierarchical encoder-decoder model (Serban et al., 2015; Serban et al., 2016) is used to establish this baseline. The words of the utterances are encoded by the encoder RNN, while the conversation history is encoded by the context RNN which is used to decode a response using another decoder RNN network.
2. **MHRED:** The second fundamental model is an extension of the first baseline model in which multi-modal information, such as images, is combined to provide a cohesive response. (Agarwal et al., 2018a).
3. **HRED(+kb):** This model is an extension of the HRED baseline (Agarwal et al., 2018b), in which we leverage review knowledge, which is encoded by the kb encoder before going through the decoder to generate suitable responses, in addition to the context encoder for encoding conversation history.
4. **MHRED(+kb):** The MHRED model is extended to MHRED(+kb) in which along with combining image output along with encoder output as in MHRED framework to include multimodal data, we also use review knowledge which is encoded by kb encoder before passing through the decoder to generate relevant responses.
5. **MHRED(+attn):** To capture relevant words in the input sequence, we apply word-level attention (Luong et al., 2015) on the encoder side of the MHRED model.
6. **MHRED(+kb)(+attn):** We utilise the knowledge base encoded by kb encoder and word level attention mechanism as in MHRED(+attn) model which is used by the decoder to generate contextual responses.
7. **Transformer:** This is a model based on the transformer encoder by (Li et al., 2019) which incrementally applies multi-head attention to build up the representation of the relevant conversation context.
8. **MTransformer:** We extend the first baseline of transformer model by including multi-modal information like images for generating logical responses. We apply a multi-head image attention layer and multi-head context attention to build up the representation of relevant context and image knowledge.
9. **Transformer(+kb):** Lastly, we again modify the Transformer model to accommodate the knowledge base by utilizing an extra multi-head knowledge attention layer over the already established multi-head context attention layer.

4.7 Evaluation Metrics

Automatic Evaluation: For evaluation, we choose BLEU (Papineni et al., 2002), F1[3] and Embedding-based metrics[4] (Liu et al., 2016) such as Greedy Matching, Vector Extrema and Embedding Average. BLEU and the unigram F1-score are used to calculate the word overlap between the ground truth and predicted response. Word-matching-based statistics have been replaced by embedding-based

---

[3] https://github.com/facebookresearch/ParlAI/blob/master/parlai/core/metrics.py

[4] https://github.com/Maluuba/nlg-eval

Models	BLEU-1%	BLEU-2%	BLEU-3%	BLEU-4%	F1%	Embedding Average	Vector Extrema	Greedy Matching
HRED	49.597	47.426	46.241	45.408	74.399	0.948	0.795	0.896
MHRED	49.664	47.522	46.338	45.496	74.423	0.947	0.793	0.895
MHRED(+attn)	50.133	48.076	46.901	46.050	75.697	0.948	0.803	0.903
Transformer	50.498	48.426	47.244	46.394	75.910	0.950	0.806	0.902
MTransformer	51.626	49.743	48.614	47.781	77.597	0.952	0.819	0.911
HRED(+kb)	52.000	50.317	49.352	48.641	78.890	0.954	0.824	0.915
MHRED(+kb)	51.380	49.549	48.510	47.756	78.082	0.951	0.818	0.911
MHRED(+attn)(+kb)	52.009	50.153	49.060	48.249	78.627	0.953	0.822	0.914
Transformer(+kb)	53.096	51.365	50.310	49.514	79.570	0.955	0.830	0.918
MKGITN	**53.271**	**51.607**	**50.582**	**49.803**	**79.846**	**0.955**	**0.832**	**0.919**

Table 2: Automatic and human evaluation metrics for baseline and our proposed (MKGITN) model. Leading results for each metrics are indicated with a bold face.

Models	Fluency	Adequacy	Knowledge Existence	Knowledge Correctness	Knowledge Relevance
Transformer	1.83	1.64	1.25	1.125	1.125
Transformer (+kb)	1.83	1.68	1.5	1.44	1.44
MTransformer	1.84	1.69	1.88	1.69	1.69
MKGITN	**1.85**	**1.73**	**1.94**	**1.75**	**1.81**

Table 3: Human assessment results for transformer based baseline models along with our proposed (MKGITN) model on the KGMMD dataset. The results in bold reflect the best value for the metric.

statistics. These metrics assign a vector to each word, as specified by word embedding, in order to perceive the intended sense of the predicted sentence. On the newly created KGMMD dataset, we test our models.

Human Evaluation: Aside from the automatic evaluation, we chose 100 samples at random from the KGMMD dataset. We engage two experts, each with a post-graduate degree and relevant expertise, to serve as annotators for human judgment using the metrics below:

(i) Fluency: It's a metric for measuring the grammatical correctness of a sentence. (ii) Adequacy: We evaluate the coherence of the generated response in relation to the conversation context. (iii) Knowledge Existence (KE): This statistic determines whether or not the response is knowledgeable. (iv) Knowledge Correctness (KC): It's used to see if the knowledge in the generated responses is correct. (v) Knowledge Relevance (KR): This metric is used to determine whether or not the knowledge is right and relevant to the situation.

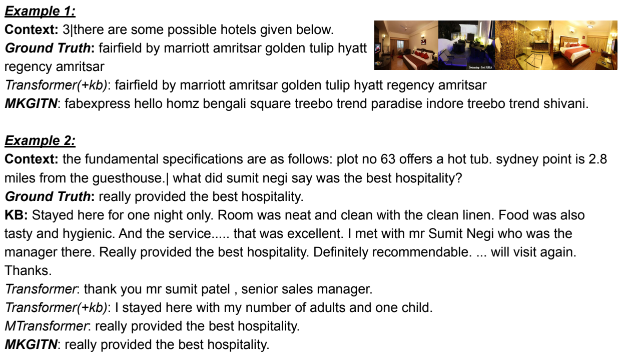
Each response is given a score of 0 to 2 (representing ”incorrect,” ”moderately correct,” and ”perfect”) by the annotators. The annotators' agreement is calculated using Fleiss' kappa(Fleiss, 1971).

5 Results and Analysis

5.1 Automatic evaluation results

The results of utilizing automatic evaluation metrics on the KGMMD dataset are shown in table 2. On most measures, MKGITN outperforms the baseline models, highlighting the value of combining external knowledge with image features for the job of document-grounded multimodal dialogue generation. On KGMMD, MKGITN achieves a considerable increase in F1 score and BLEU-4 when compared to the previous best HRED-based models, achieving a new state-of-the-art. The results are all statistically significant (t-test with p-value at 0.05 significance level).

We saw a significant improvement in the automatic evaluation metrics from our models that incorporate multimodality, such as MHRED and MTransformer, indicating efficient use of visual information provided with the input utterances. MKGITN, on the other hand, is able to capture the dialogue context and successfully apply external information to generate proper responses, as evidenced by the significant improvement in all models when knowledge is used. Rather than relying on word overlaps, embedding-based metrics compute the cosine similarity between the predicted and reference sentences' embeddings. All three metrics have greatly improved, suggesting that our methodology is successful and outperforms the strongest baseline, MHRED(+kb)(+attn), by a significant margin.



Example 1:

Context: 3|there are some possible hotels given below.

***Ground Truth*:** fairfield by marriott amritsar golden tulip hyatt regency amritsar

Transformer(+kb): fairfield by marriott amritsar golden tulip hyatt regency amritsar

MKGITN: fabexpress hello homz bengali square treebo trend paradise indore treebo trend shivani.

Example 2:

Context: the fundamental specifications are as follows: plot no 63 offers a hot tub. sydney point is 2.8 miles from the guesthouse.| what did sumit negi say was the best hospitality?

***Ground Truth*:** really provided the best hospitality.

KB: Stayed here for one night only. Room was neat and clean with the clean linen. Food was also tasty and hygienic. And the service..... that was excellent. I met with mr Sumit Negi who was the manager there. Really provided the best hospitality. Definitely recommendable. ... will visit again. Thanks.

Transformer: thank you mr sumit patel , senior sales manager.

Transformer(+kb): I stayed here with my number of adults and one child.

MTransformer: really provided the best hospitality.

MKGITN: really provided the best hospitality.

Figure 3: Predictions from our proposed MKGITN model.

5.2 Human evaluation results

Table 3 shows the findings of the human evaluation. Because manual evaluation is costly, we only compare our proposed model to the four transformer-based models on KGMMD. MKGITN has improved fluency and outperforms baseline models on both adequacy and knowledge-related criteria, showing high correlation with the automatic evaluation results. The annotators' for human evaluation have kappa scores greater than 0.75, indicating that they all agree with each other.

In Figure 3, we present a few example conversations as predicted by the proposed (*MKGITN*) and baseline models (*Transformer*, *MTransformer* and *Transformer(+kb)*) on test set of KGMMD dataset. In the first example, we may note that the proposed model (*MKGITN*) efficiently utilizes the image information associated with the conversation. As seen in the second example, the model correctly decodes the response using more detailed information from the review based knowledge.

5.3 Error Analysis

Using the generated responses, we conduct a thorough examination of our proposed model and classify the errors it encounters into the following categories:

1. Incorrect values: Because multiple products may have similar qualities, the model occasionally generates inaccurate values in the predicted utterance, such as hotel pricing, when compared with the gold utterance: Gold: *Rs 5,305 3 nights for Rs 15,915*; Predicted: *Rs 36,480*. v 2. Incorrect knowledge: The model focuses on incorrect knowledge with respect to the conversational history since sometimes the knowledge contains extra information not related to the context. Gold: *if you are a golfer - or even if you are not - staying at seaview is truly a treat*; Predicted: *if you are not staying at seaview is truly a great hotel - a long - term or stay*; Knowledge Base: *We started staying at this hotel several years back for a junior golf tournament - we thought it was beautiful then. ... If you are a golfer - or even if you are not - staying at Seaview is truly a treat. Old school hospitality, beautiful furnishings, a quiet setting and just a short drive from AC...love this place!*.

6 Conclusion

This research proposes a novel knowledge-driven paradigm for generating knowledge-based discussion in task-oriented settings. We employ both an image and an external knowledge base to support conversation understanding and generation. The results of the assessments on the KGMMD dataset show that a multi-modal knowledge grounded incremental transformer network may provide factually correct and informative responses by combining grounded knowledge and multimodal information linked with input utterances.

The dataset, codes and models used to replicate our findings are available at https://github.com/deekshaVarshney/KGMMD.git. We plan to leverage the sentiment values linked with the reviews in our knowledge base in the future to expand our work to sentiment and aspect based multimodal dialogue generation based on knowledge.

Acknowledgments

Authors duly acknowledge the support from the Project titled "Sevak-An Intelligent Indian Language Chatbot", Sponsored by SERB, Govt. of India (IMP/2018/002072)

References

Shubham Agarwal, Ondrej Dusek, Ioannis Konstas, and Verena Rieser. 2018a. Improving context modelling in multimodal dialogue generation. In *11th International Conference of Natural Language Generation 2018*, pages 129–134. Association for Computational Linguistics.

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

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

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

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 Annual Meeting of the Association for Computational Linguistics*, pages 5437–5447.

Xiuyi Chen, Jiaming Xu, and Bo Xu. 2019. A working memory model for task-oriented dialog response generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2687–2693, Florence, Italy, July. Association for Computational Linguistics.

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*, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2020. Multidm-gcn: Aspect-guided response generation in multi-domain multi-modal dialogue system using graph convolution network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 2318–2328.

Mauajama Firdaus, Nidhi Thakur, and Asif Ekbal. 2021. Aspect-aware response generation for multimodal dialogue system. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 12(2):1–33.

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

Zhe Gan, Yu Cheng, Ahmed El Kholy, Linjie Li, Jingjing Liu, and Jianfeng Gao. 2019. Multi-step reasoning via recurrent dual attention for visual dialog. *arXiv preprint arXiv:1902.00579*.

Hitesh Golchha, Mauajama Firdaus, Asif Ekbal, and Pushpak Bhattacharyya. 2019. Courteously yours: Inducing courteous behavior in customer care responses using reinforced pointer generator network. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 851–860.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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

Hung Le, Doyen Sahoo, Nancy F Chen, and Steven CH Hoi. 2019b. Multimodal transformer networks for end-to-end video-grounded dialogue systems. *arXiv preprint arXiv:1907.01166*.

Zekang Li, Cheng Niu, Fandong Meng, Yang Feng, Qian Li, and Jie Zhou. 2019. Incremental transformer with deliberation decoder for document grounded conversations. *arXiv preprint arXiv:1907.08854*.

Lizi Liao, Yunshan Ma, Xiangnan He, Richang Hong, and Tat-seng Chua. 2018. Knowledge-aware multimodal dialogue systems. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 801–809.

Kuan-Yen Lin, Chao-Chun Hsu, Yun-Nung Chen, and Lun-Wei Ku. 2019. Entropy-enhanced multimodal attention model for scene-aware dialogue generation. *arXiv preprint arXiv:1908.08191*.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, Austin, Texas, November. Association for Computational Linguistics.

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

Andrea Madotto, Chien-Sheng Wu, and Pascale Fung. 2018. Mem2seq: Effectively incorporating knowledge bases into end-to-end task-oriented dialog systems. *arXiv preprint arXiv:1804.08217*.

Fandong Meng, Zhaopeng Tu, Yong Cheng, Haiyang Wu, Junjie Zhai, Yuekui Yang, and Di Wang. 2018. Neural machine translation with key-value memory-augmented attention. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 2574–2580.

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. *arXiv preprint arXiv:1701.08251*.

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 on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Dinesh Raghu, Nikhil Gupta, and Mausam. 2019. Disentangling Language and Knowledge in Task-Oriented Dialogs. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1239–1255, Minneapolis, Minnesota, June. Association for Computational Linguistics.

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

Amrita Saha, Mitesh Khapra, and Karthik Sankaranarayanan. 2018. Towards building large scale multimodal domain-aware conversation systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2015. Hierarchical neural network generative models for movie dialogues. *arXiv preprint arXiv:1507.04808*, 7(8).

Iulian 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 AAAI Conference on Artificial Intelligence*, volume 30.

Iulian Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron Courville, and Yoshua Bengio. 2017. A hierarchical latent variable encoder-decoder model for generating dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. *arXiv preprint arXiv:1503.02364*.

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

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 on conference on information and knowledge management*, pages 553–562.

Sainbayar Sukhbaatar, Arthur Szlam, Jason Weston, and Rob Fergus. 2015. End-to-end memory networks. *Advances in Neural Information Processing Systems*, 2015:2440–2448.

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 Annual Meeting of the Association for Computational Linguistics*, pages 3816–3825.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

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

Jian Wang, Junhao Liu, Wei Bi, Xiaojiang Liu, Kejing He, Ruifeng Xu, and Min Yang. 2020. Dual dynamic memory network for end-to-end multi-turn task-oriented dialog systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4100–4110, Barcelona, Spain (Online), December. International Committee on Computational Linguistics.

Chien-Sheng Wu, Richard Socher, and Caiming Xiong. 2019. Global-to-local memory pointer networks for task-oriented dialogue. *arXiv preprint arXiv:1901.04713*.

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.

Weijie Zhang, Jiaoxuan Chen, Haipang Wu, Sanhui Wan, and Gongfeng Li. 2021. A knowledge-grounded dialog system based on pre-trained language models. *arXiv preprint arXiv:2106.14444*.