Tackling Situated Multi-Modal Task-Oriented Dialogs with a Single Transformer Model

Haeju Lee[1*], Oh Joon Kwon[1*], Yunseon Choi[1*], Jinhyeon Kim[1], Youngjune Lee[2], Ran Han[3], Yoonhyung Kim[3], Minho Park[3], Kangwook Lee[4], Haebin Shin[4], Kee-Eung Kim[1 2]

[1] Kim Jaechul Graduate School of AI, KAIST, Daejeon, Republic of Korea
[2] School of Computing, KAIST, Daejeon, Republic of Korea
[3] Electronics Telecommunications Research Institute (ETRI), Daejeon, Republic of Korea
[4] Samsung Research, Seoul, Republic of Korea
{hjlee, ojkwon, yschoi, jhkim, yjlee}@ai.kaist.ac.kr, {ran.han, yhkim1127, roger618}@etri.re.kr,
{kw.brian.lee, haebin0.shin}@samsung.com, kekim@kaist.ac.kr

Abstract

The Situated Interactive Multi-Modal Conversations (SIMMC) 2.0 track in the Dialog System Technology Challenge 10 (DSTC10) aims to create virtual shopping assistants that can accept complex multi-modal inputs, i.e. visual appearances of objects and user utterances. It consists of four subtasks, multi-modal disambiguation (MM-Disamb), multi-modal coreference resolution (MM-Coref), multi-modal dialog state tracking (MM-DST), and response retrieval and generation. While many task-oriented dialog systems usually tackle each subtask separately, we propose a jointly learned encoder-decoder that performs all four subtasks at once for efficiency. Moreover, we handle the multi-modality of the challenge by representing visual objects as special tokens whose joint embedding is learned via auxiliary tasks. Finally, we won in the MM-Coref and response retrieval subtasks and nominated runner-up for the remaining subtasks using a single unified model. In particular, our model achieved 81.5% MRR, 71.2% R@1, 95.0% R@5, 98.2% R@10, and 1.9 mean rank in response retrieval task along with competitive results in all subtasks, setting a high bar for the state-of-the-art result in SIMMC 2.0.

Introduction

A task-oriented dialog system aims to assist users accomplish certain tasks, such as executing actions or retrieving specific information, with natural language conversations. To build a successful task-oriented dialog system that can satisfy the user's need and handle unexpected circumstances, the system should be able to understand the user's intent, track the flow of dialog, and then respond appropriately given the dialog context. Recent advances in machine learning have led to such systems being deployed as actual products (Bordes and Weston 2016; Joshi, Mi, and Faltings 2017).

With the rising interest and ubiquity of virtual reality (VR), the next generation of task-oriented virtual assistants is expected to handle conversations in a multi-modal context. For instance, a multi-modal dialog agent may help the user navigate a virtual clothing store and look for an object meeting the user's criteria. In such cases, not only will the assistant's capability to understand dialog but also its ability to parse and understand other modalities, namely scene objects, will be crucial in creating a successful multi-modal agent.

To this end, SIMMC 1.0 (Moon et al. 2020) was proposed to reflect a situated multi-modal context in the form of co-observed scene. However, it did not fully capture the complexity of multi-modal conversations as it only provided simplistic, sanitized scene contexts. The new SIMMC 2.0 (Kottur et al. 2021) lifts these limitations by providing a more realistic scene set in VR stores to incorporate the complexity of multi-modal task-oriented dialogs. The additional subtasks, MM-Disamb and MM-Coref, intend to test the capability of virtual agents to identify the need for disambiguating reference mentions and to ground them to the objects in the scene. While challenging, these subtasks are essential to building a successful multi-modal task-oriented dialog agent.

In this paper, we present our end-to-end, joint-learning approach to address this challenge.[1] We adopt BART (Lewis et al. 2020) and attach task-specific heads so that the model can make predictions on all subtasks. To be more specific, our model performs MM-Disamb, MM-Coref, and response retrieval by the encoder and MM-DST and response generation in a string format by the decoder. We also integrate multi-modality into the model by treating scene objects as unique object tokens and coreference sentinel tokens. Our model is jointly trained on all subtasks and a few auxiliary objectives to help the model align object tokens to its attributes. For retrieval, we use in-batch negative samples for contrastive metric learning instead of creating a pool of separate training samples.

Our model was ranked at the first place for MM-Coref and

[*] These authors contributed equally.

[1] The code is available at https://github.com/KAIST-AILab/DSTC10-SIMMC

response retrieval with 75.8% coreference F1, 81.5% MRR, 71.2% R@1, 95.0% R@5, 98.2% R@10, and 1.9 mean rank in the official evaluation. Moreover, our model was nominated runner-up for all other subtasks, in which we achieved 93.8% disambiguation accuracy, 90.3% slot F1, 95.9% intent F1, and 0.295 BLEU-4. The results were obtained with only a single model and consistent with the results on the devtest (i.e. validation) set, demonstrating a robust, common representation on all subtasks learned by the model.

Related Works

Task-Oriented Dialog Systems

The traditional approach for building task-oriented (also called goal-oriented) dialog systems adopts a modular pipeline architecture that integrates natural language understanding (NLU) module that identifies user's intent (Liu and Lane 2016), dialog state tracking (DST) module that extracts values for slots (Henderson, Thomson, and Young 2013; Mrksic et al. 2017), dialog policy management (POL) module that decides system action (Wen et al. 2017), and natural language generation (NLG) module that generates appropriate system utterance according to system action (Wen et al. 2015).

Recent works on task-oriented dialog systems suggested leveraging pretrained language models (LM) that can perform all these tasks in an end-to-end, auto-regressive manner (Ham et al. 2020; Hosseini-Asl et al. 2020; Yang, Li, and Quan 2021). Given a dialog context, such systems sequentially generates belief state, system action, and response, making predictions based on decisions made by previous modules in the form of tokens. Moreover, some of these systems incorporate knowledge base (KB) so that the system can either explicitly retrieve relevant information or implicitly instill KB into its latent representation (Yang, Zhang, and Erfani 2020) to generate even more accurate responses.

Multi-Modal Models

Building cross-modal models has recently gained a lot of attention, especially in the domain of vision and language (VL). Many recent works develop the cross-modal model on the top of the transformer-based (Vaswani et al. 2017) pretrained LM, focusing on aligning visual features to linguistic contents. They propose various pretraining methods to learn the aligned embedding space for VL tasks and achieve state-of-the-art performance in downstream tasks as shown in UNITER (Chen et al. 2020) and OSCAR (Li et al. 2020). Many of these models focus on solving question answering problems with visual observations. In this context, a number of visual question answering (VQA) tasks (Antol et al. 2015; Johnson et al. 2017) have been developed to accelerate research in this area.

Conversational Recommendation

A conversational recommender system (CRS) should assist the user who looks for objects in a free form text, i.e. learn the user preference from dialogs and recommend the object based on external knowledge of objects. As a recent example, Zhou et al. (2020) present CRS that integrates word-oriented knowledge graph (KG) and object-oriented KG to produce more informative response. In this paper, we focus on understanding objects (i.e. shopping items) appearing in a visual scene. Based on the objects in a scene, the dialog system needs to recommend objects or provide information of objects in the response.

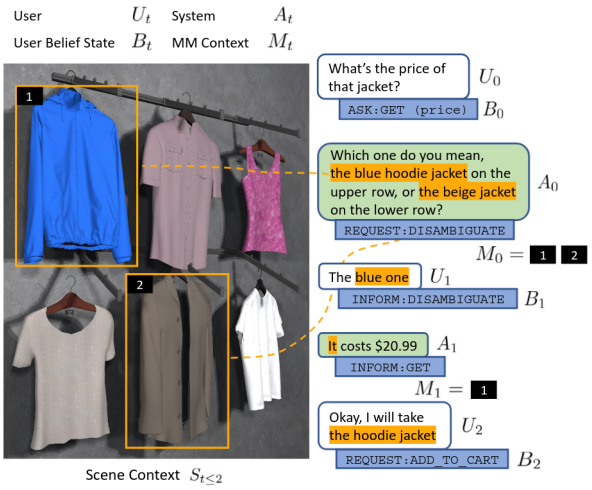

Figure 1: An instance of dialog and the corresponding scene in SIMMC 2.0. Here, the assistant asks the user to disambiguate between *the blue hoodie jacket* (indexed as 1) and *the beige jacket* (indexed as 2), grounding its mentions to the scene via multi-modal context $M_0 = \{1, 2\}$. Once the user chooses *the blue one*, the system retrieves the information on the disambiguated object. The multi-modal context in this case would be $M_1 = \{1\}$.

Task Descriptions

SIMMC 2.0 Dataset

SIMMC 2.0 (Kottur et al. 2021) follows the setting of SIMMC 1.0 (Moon et al. 2020), which assumed conversations occurring between a user and an assistant in a situated, co-observed VR scene. This newer iteration of the dataset lifts the limitations of SIMMC 1.0 by further capturing the complexity of multi-modal conversations; whereas SIMMC 1.0 had at most three objects in a simple, sanitized scene, SIMMC 2.0 provides a far richer visual context with 19.7 objects on average that are often occluded, cluttered, or even out of view. An example dialog is shown in Figure 1.

The SIMMC 2.0 dataset consists of 11,244 dialogs split into train (65%), dev (5%), devtest (15%), and teststd (15%) sets. Each dialog consists of multiple turns where each turn has grounded multi-modal context and an accompanying scene with referential indices. We shall denote a SIMMC dialog with r rounds as $\mathcal{D} := \{(U_t, A_t, M_t, S_t, B_t)\}_{t=1}^{r}$, where U_t is user utterance, A_t system utterance, M_t multi-modal context, S_t scene context, and B_t user belief state at turn t (refer to Figure 1 for a concrete example). Here, M_t is a set of object indices mentioned by the system and S_t are all objects in a scene. User belief state B_t is com-

bination of dialog act and slot, where dialog act reflects intention of user utterance, slot indicates attributes of objects user is interested in, and request slot indicates the information user wants to know about interesting objects. We also define the dialog history at some turn $T \leq r$ as $H_T := \{U_0, A_0, M_0, \ldots, U_{T-1}, A_{T-1}, M_{T-1}\}$. The assistant needs to make predictions conditioned on history H_T, current user utterance U_T, and the scenes up to the current turn $S_{t \leq T}$. The object set of SIMMC 2.0. is composed of fashion domain and furniture domain, where each domain has 288 and 57 items respectively. The system is allowed to look up which item is present in a scene at all time. As a side information, the metadata of each object such as color, type, brand, size, and price are provided, but looking up the visual attribute (e.g. color, pattern, materials, sleeve length) is prohibited at inference time so as to make the agent need to process visual information from the scene image.

SIMMC 2.0 Subtasks

Multi-modal disambiguation (MM-Disamb) The first subtask is to identify whether the assistant should disambiguate mentions in the next turn given the dialog and multi-modal context. For instance, given user utterance "*How much is the pair on the left?*", there may be more than two pairs of pants on the left. In this case, ambiguity in reference should be resolved. This can be cast into a binary classification task, and the performance is measured by accuracy.

Multi-modal coreference resolution (MM-Coref) The second subtask is to map the referential mentions of the user utterance to the object indices in the scene. These mentions should be resolved through the linguistic context and the multi-modal context. The performance is measured by object slot F1 score.

Multi-modal dialog state tracking (MM-DST) The third subtask extends the traditional uni-modal DST to ground user belief state on the multi-modal objects. This will measure the assistant's understanding throughout each dialog, which includes disambiguation and coreference resolution. The performance is measured by the F1 score for dialog act and request slots.

Response retrieval & generation The last subtask is to retrieve or generate appropriate system utterance. Response generation is evaluated with BLEU-4 (Papineni et al. 2002). For response retrieval, the system is expected to choose the most relevant response from a pool of 100 candidate responses. Recall@k ($k \in \{1, 5, 10\}$), mean rank, and mean reciprocal rank (MRR) are used for retrieval evaluation.

Methods

Even though the setting of the dataset is similar to that of VQA where finetuning the pretrained VL models are prevalent, we chose to work with LM, representing objects by tokens. There are several reasons behind this choice. First, the vision models are usually pretrained on natural images (Lin et al. 2014; Krishna et al. 2017), so finetuning them requires a relatively large number of training samples of 3D rendered images that are aligned properly with text. Second, in a realistic scenario where the assistant is deployed in a VR environment, the object metadata and scene graphs would be readily available as a part of the system. In this case, using a vision backbone model would be an unnecessary overhead. Lastly, we can still easily provide additional supervision signals at train time for modality alignment by looking up the object metadata. For this, we represent multi-modal objects as the concatenation of their referential indices in the scene (canonical object ID) and their absolute attribute (unique object token).

We note that all of the subtasks are related to each other. Recent works (Ham et al. 2020; Hosseini-Asl et al. 2020; Yang, Li, and Quan 2021) predict user belief states then system response end-to-end so that the model can generate appropriate response conditioned on user belief state. We further this idea for the SIMMC 2.0 subtasks, which are closely related to each other and suggest compelling reason for joint learning of all subtasks. For example, if the assistant decides that the user utterance needs to be disambiguated, then the appropriate system action is to respond along the line of "*Which one are you referring to?*". We expect that the latent representation of the multi-modal dialog learned from other subtasks will translate readily to other subtasks.

Among the four SIMMC 2.0 subtasks, we identify MM-

Common Input (x)	
U_{T-1}	<USR> What are the good hoodies around here?
A_{T-1}	<SYS> I advise you consider the solid green one.
M_{T-1}	<SOM> <56> <EOM>
U_T	<USR> I do like solid colors, but I'm looking for something with excellent ratings.
$S_{t<T}$	<SOO> <PREV_OBJ> <12> <fashion_142> <PREV_OBJ> <13> <fashion_058>
S_T	<OBJ> <56> <fashion_269> <OBJ> <85> <fashion_007> <EOO>
Generation Target	
B_T	<SOB> INFORM:GET <customerReview> good <pattern> plain <type> hoodie <EOB>
A_T	In fact, that green hoodie is very highly rated.
Response Candidate	<SYS> In fact, that green hoodie is very highly rated.

Table 1: Example input representations for our model. We show only up to last 1 turn due to space limit. The common input x is a concatenation $[H_T; U_T; S_{t \leq T}]$ where $H_T = \{U_{T-1}, A_{T-1}, M_{T-1}\}$. Here, we separate the scene history $S_{t<T}$ to show how we handle out-of-view objects. The generation target is a concatenation $[B_T; A_T]$, which is used by the decoder. The response candidate is A_T with speaker identifier <SYS> prepended.

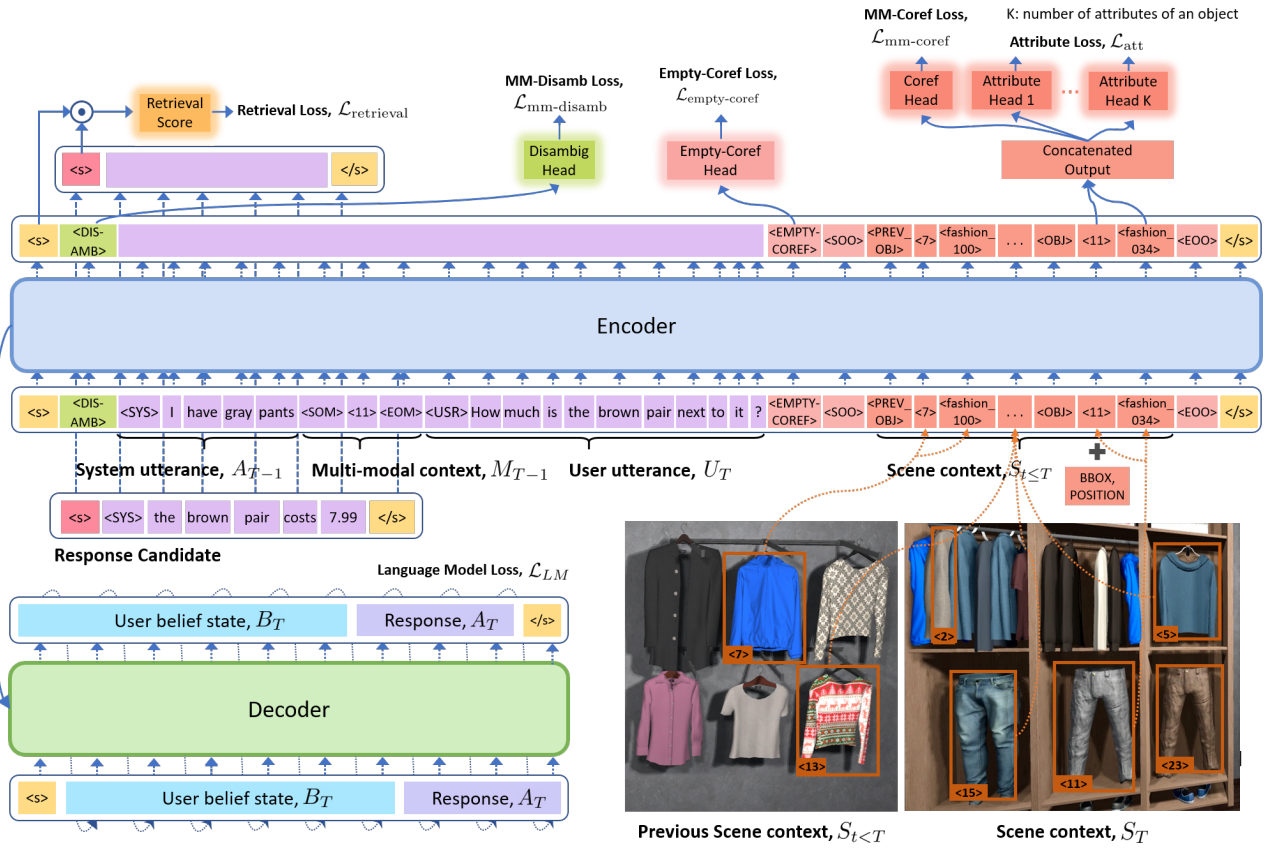

Figure 2: Overview of the jointly learned multi-tasking BART. For H_T, we show only the last turn without user utterance due to space limit. The details on the loss functions are provided in model specifics. Each scene object is represented by the concatenation of scene canonical object ID token (e.g. <11>) and unique object token (e.g. <fashion_123>). It is then passed through MM-Coref and attribute classification head. MM-DST and response generation subtasks are approached in terms of auto-regressive LM.

Coref to be the most challenging one. In this subtask, the assistant should be able to understand both the natural language utterances and the object attributes to choose the right objects from the scene. We view this as a type of set prediction, where joint learning of set cardinality and state distribution was shown to be effective (Rezatofighi et al. 2018). Hence, we define an auxiliary empty coreference target set prediction, a simplified cardinality prediction that outputs whether the current user utterance has no coreference targets. Moreover, we conduct a supervised learning on object attributes to help align object-language modalities.

In order to harness the power of NLU/NLG capabilities demonstrated by pretrained transformer encoder-decoder, we adopt BART (Lewis et al. 2020) as the pretrained language backbone. We attach classification heads for MM-Disamb and MM-Coref subtasks at the encoder and LM head for MM-DST and response generation at the decoder. We also perform retrieval by computing the dot product between representation vectors of response candidates and multi-modal dialog context. The overview of the model is provided in Figure 2.

Input representation

For all of the subtasks, we define our input to be a simple concatenation $x := [H_T; U_T; S_{t\leq T}]$ with separators. We define H_T to be the dialog history up to 2 turns to limit the length of input, i.e. $\{U_{T-2}, A_{T-2}, M_{T-2}, U_{T-1}, A_{T-1}, M_{T-1}\}$. SIMMC 2.0 assumes that utterances may mention objects that are not in the current scene S_T but in the previously observed scene $S_{t<T} \neq S_T$. Hence, our model integrates the objects from the previous scene that are not in the current scene. We find that our scene representation by enumerating all objects is a simple yet effective method for the model to understand the multi-modal context. An example input is provided in Table 1.

Canonical object ID token A canonical object ID token takes the form of <\d+> (e.g. <32>). This provides a relational context of the object within the scene, grounding each object to its scene object index provided in the dataset. This scheme was also used in the baseline code for SIMMC 2.0 (Kottur et al. 2021), but without any association to its attributes. In our method, this token intends to provide con-

textual information of the object alongside its absolute attributes (unique object token), allowing the assistant to make connections between dialog mentions and its multi-modal attributes.

Unique object ID token Unique object ID token takes the form of <{domain}_\d+> (e.g. <fashion_123>, <furniture_028>). The digits following the domain specifier denote index of the unique object in that domain. This encodes the visual (e.g. type, color, material) and non-visual (e.g. price, customer rating) attributes unique to each object.

Separator tokens We define several separator tokens to delimit different components of the multi-modal dialogs. We use <SOM>, <EOM> for the start and the end of multi-modal context and <SOO>, <EOO> for the start and the end of scene objects. Within the scene context, <OBJ> token is used as a separator token between objects, which are represented by the concatenation of a canonical object ID token and a unique object token. We also mark the objects from the previous scene with <PREV_OBJ> instead of <OBJ>. For generation target, we mark the start and the end of the user belief state with <SOB>, <EOB>.

Encoding object locations For the assistant to understand the spatial relation among objects within the scene, we must incorporate encoded representation of location of each object. We follow the commonly used techniques in VL models (Li et al. 2020; Chen et al. 2020; Zhang et al. 2021) for encoding object locations with the bounding box information. Given a bounding box represented by its upper-left and lower-right vertices, (x_1, y_1) and (x_2, y_2), with height h and width w, we encode its location as tuple $(x_1/w - 0.5, y_1/h - 0.5, x_2/w - 0.5, y_2/h - 0.5, (x_2 - x_1)(y_2 - y_1)/(h \cdot w))$. This is passed through a location embedding layer (a fully-connected layer followed by layer norm) to be added with the canonical object ID token encodings.

Model specifics

Binary prediction for MM-Disamb and MM-Coref We formulate MM-Disamb as a binary classification on the pooled output of the encoder, from the pooling token <DISAMB>. The binary head for MM-Disamb should predict true if the current user utterance U_T needs to be disambiguated and false otherwise.

For MM-Coref, we make binary predictions on all objects in $S_{t \leq T}$. We do so by passing the concatenated canonical object (e.g. <11>) and unique object ID (e.g. <fashion_001>) encoder output of each object through a binary classification head. The MM-Coref head will predict true if the current user utterance mentions that object and false otherwise. We use a simple cross-entropy loss for both MM-Disamb and MM-Coref, denoted $\mathcal{L}_{\text{mm-disamb}}$ and $\mathcal{L}_{\text{mm-coref}}$.

Auto-regressive LM for MM-DST and generation We also approach MM-DST and response generation subtasks with auto-regressive LM following the recent approaches in end-to-end dialog systems. For MM-DST and response generation, we use the standard left-to-right LM loss (Bengio et al. 2003).

$$\mathcal{L}_{\text{LM}} = \sum_{i=1}^{L} -\log P(\omega_i \mid \omega_1, \ldots, \omega_{i-1}),$$

where ω_i is the i-th target token and L the total length of the target.

In-batch negative samples for retrieval For response retrieval task, we make use of in-batch negative samples for contrastive learning on similarity metrics. We treat the system responses of the other samples in the batch formatted according to Table 1 as in-batch negatives. We then pool the encoder outputs of the input and the response candidates with BART bos token, i.e. <s>, to compute their dot product, so that the correct scene-response candidate pair stays close and the incorrect pairs stay apart. In order to improve retrieval performance, we utilize as large a training batch size as the hardware can manage. We use multi-class cross-entropy loss applied to dot-product similarities, i.e.

$$\mathcal{L}_{\text{retrieval}} = -\log \frac{\exp(\mathbf{x} \cdot \mathbf{a}^+)}{\sum_{\mathbf{a}^- \in B^-(\mathbf{x}) \cup \{\mathbf{a}^+\}} \exp(\mathbf{x} \cdot \mathbf{a}^-)},$$

where $\mathbf{a}^+$ is the positive response sample of the input $\mathbf{x}$ and $B^-(\mathbf{x})$ the set of in-batch negative responses (assume $\mathbf{x}$, $\mathbf{a}^+$, and $\mathbf{a}^-$ are pooled representations from the encoder). We formulate the task loss $\mathcal{L}_{\text{task}}$ as a linear combination of losses from each subtask.

$$\begin{aligned}\mathcal{L}_{\text{task}} = \; & \lambda_{\text{LM}}\mathcal{L}_{\text{LM}} + \lambda_{\text{mm-disamb}}\mathcal{L}_{\text{mm-disamb}} \\ & + \lambda_{\text{mm-coref}}\mathcal{L}_{\text{mm-coref}} + \lambda_{\text{retrieval}}\mathcal{L}_{\text{retrieval}}\end{aligned} \tag{1}$$

Auxiliary tasks

Binary prediction for empty coreference set We define an additional empty coreference prediction task, in which the assistant predicts whether the current dialog turn has MM-Coref targets. This can be seen as a simpler version of set cardinality prediction. We find this additional signal for coreference resolution, denoted $\mathcal{L}_{\text{empty-coref}}$, is advantageous in boosting MM-Coref performance, a type of set prediction task. For this, we use <EMPTY_COREF> for pooling. At inference time, we override any MM-Coref predictions if the empty coreference prediction is true (i.e. there is no coreference target). We use a binary cross-entropy loss for $\mathcal{L}_{\text{empty-coref}}$.

Encoding object attributes We encode object attributes by providing additional supervision signal during training. We do so by simply training to classify each object to its corresponding visual and non-visual attributes such as color, price, and customer ratings. Each object is represented as a concatenation of its canonical object ID and unique object token as in MM-Coref (refer to Figure 2). Each attribute head predicts a categorical class for each corresponding object, for example, if <fashion_001> is a grey jacket, the color-attribute head predicts the class of grey and the type-attribute head predicts the class of jacket.

Let $\mathcal{O}_{t \leq T}$ be the set of objects in the scene history, $S_{t \leq T}$. We denote attribute multi-class classification loss $\mathcal{L}_{\text{att}}$ for all

Models	#1 Disamb.	#2 MM-Coref	#3 MM-DST		#4-1 Res. Retrieval					#4-2 Res. Gen.
	Accuracy (↑)	Obj. F1 (↑)	Slot F1 (↑)	Act. F1 (↑)	MRR (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)	M. Rank (↓)	BLEU-4 (↑)
GPT-2 Baseline	73.8%	36.6%	81.7%	94.5%	8.8%	2.6%	10.7%	18.4%	38.0	0.192
MTN Baseline	-	-	74.8%	93.4%	-	-	-	-	-	0.217
`bart-large`	92.7%	74.3%	89.2%	96.2%	80.7%	71.1%	94.4%	98.3%	1.93	0.314

Table 2: Overall and ablation study results on the devtest set. GPT-2 and MTN are the baselines provided by the organizers, which are separately trained on each subtask. The MTN baseline performs only MM-DST and response generation.

Entry ID	#1 Disamb.	#2 MM-Coref	#3 MM-DST		#4-1 Res. Retrieval					#4-2 Res. Gen.
	Accuracy (↑)	Obj. F1 (↑)	Slot F1 (↑)	Act. F1 (↑)	MRR (↑)	R@1 (↑)	R@5 (↑)	R@10 (↑)	M. Rank (↓)	BLEU-4 (↑)
1	-	52.1%	89.1%	96.3%	53.5%	42.8%	65.4%	74.9%	11.9	0.285
2	89.5%	42.2%	87.8%	96.2%	61.2%†	49.6%†	74.7%†	84.5%†	6.6†	0.256
3 (Ours)	93.9%†	**75.8%**	90.3%†	95.9%†	**81.5%**	**71.2%**	**95.0%**	**98.2%**	**1.9**	0.295†
4	93.8%†	56.4%	89.3%	96.4%	32.0%	19.9%	41.8%	61.2%	12.9	**0.322**
5	**94.7%**	59.5%	**91.5%**	**96.0%**	-	-	-	-	-	-
6	93.1%	57.3%	-	-	-	-	-	-	-	-
7	93.1%	68.2%	4.0%	41.4%	-	-	-	-	-	0.297†
8	-	73.3%†	-	-	-	-	-	-	-	-
9	93.6%†	68.2%	87.7%	95.8%	-	-	-	-	-	**0.327**

Table 3: The official leaderboard of DSTC10 on the teststd set. The subtask winners are bold-faced and runner-ups are marked with †. "-" means that the entry did not participate in that subtask.

objects in $\mathcal{O}_{t\leq T}$,

$$\mathcal{L}_{\text{att}} = \sum_{j\in\mathcal{O}_{t\leq T}} \sum_{k=1}^{K} \sum_{c\in\mathcal{C}_k} -\mathbb{1}\{c = y_{jk}\} \log P(c),$$

where K is the number of attributes, $\mathcal{C}_k$ the set of all classes of the k-th attribute, y_{jk} the label of the k-th attribute of the j-th object, and $\mathbb{1}\{\cdot\}$ is an indicator function.

As a result, the auxiliary loss $\mathcal{L}_{\text{aux}}$ is defined as the weighted sum of attribute loss and empty-coreference prediction loss:

$$\mathcal{L}_{\text{aux}} = \lambda_{\text{att}}\mathcal{L}_{\text{att}} + \lambda_{\text{empty-coref}}\mathcal{L}_{\text{empty-coref}} \quad (2)$$

In summary, we minimize the total loss $\mathcal{L}_{\text{total}}$, which is the sum of the task loss $\mathcal{L}_{\text{task}}$ from Equation 1 and the auxiliary loss $\mathcal{L}_{\text{aux}}$ from Equation 2.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{aux}}$$

Experiments

Experimental setup Our model is built on top of 24-layer BART from HuggingFace (`facebook/bart-large`) (Wolf et al. 2019).[2] We finetune the model for 10 epochs with an initial learning rate of 5e-5 and a batch size of 16 with AdamW optimizer (Loshchilov and Hutter 2018). We also use linear warmup schedule with 8000 warmup steps and clip gradient norms at 1.0. For joint learning coefficients, we choose $(\lambda_{\text{LM}}, \lambda_{\text{mm-disamb}}, \lambda_{\text{mm-coref}}, \lambda_{\text{retrieval}}, \lambda_{\text{att}}, \lambda_{\text{empty-coref}}) = (1.0, 0.1, 0.8, 0.4, 0.1, 0.1)$. For decoding, we use top-p sampling (Holtzman et al. 2020) with $p = 0.9$ to generate the user belief state and system response. We choose the best checkpoint evaluated at every 1000 steps on the devtest set.

Baselines The challenge organizers provided two baseline models: an end-to-end GPT-2 (Radford et al. 2019) and multi-modal transformer networks (MTN) (Le et al. 2019). The baseline models do not explicitly use object attributes and model each subtask separately, except for MM-Coref, MM-DST, and response generation. GPT-2 baseline generates the user belief state, coreference objects (in the form of canonical object IDs), and response in an end-to-end manner. MTN baseline conditions on the scene image and dialog history then generate the user belief state and response using a multi-model transformer. MM-DST and response generation are not implemented in the GPT-2 baseline.

Results The results on the devtest (validation) and teststd (test) splits are shown in Table 2 and 3, respectively. On devtest set, our proposed model outperforms the baselines by a large margin. Our proposed model based on `bart-large` and was ranked at the first place with 75.8% coreference F1 in MM-Coref. This demonstrates that our method of injecting object attributes to the model was effective, providing a richer context about the scene and its objects to the assistant. Furthermore, our model was declared winner in the response retrieval subtask with 71.2% R@1, 95.0% R@5, 98.2% R@10, and 1.9 mean rank. This is a remarkable performance compared to existing methods such as bi- and poly-encoders (Humeau et al. 2020), despite the fact that we only used a single encoder built into the model to encode both the dialog context and candidates.

Our method of representing scene and learning joint embedding between dialog and scene successfully captured

[2] https://github.com/huggingface/transformers

fine-grained information on the scene objects. This allows for the model to attend and focus on objects that are being mentioned in the conversation, learning to choose the right response most of the time. Moreover, our model showed competitive performance and was declared runner-up in all remaining sub-tasks, in which we achieved 93.8% disambiguation accuracy, 90.3% slot F1, 95.9% intent F1, and 0.295 BLEU-4 with a single model. We also observe that the results on the teststd split does not deviate much from the results on the devtest split, demonstrating a robust representation learned by the model.

Conclusion

In this paper, we propose a multi-modal task-oriented dialog system based on BART that can perform all SIMMC 2.0 subtasks at once. Our model overcomes the challenge of adopting occluded, 3D rendered images to vision models by integrating multi-modal objects as special tokens. In addition to joint learning of all subtasks, we introduce empty coreference set prediction and attribute classification to directly align objects to their corresponding attributes. We observe that these additional subtasks are crucial in building a successful multi-modal assistant for SIMMC 2.0. Our model is able to perform competitively in all of the subtasks with a single model, ranking first place for MM-Coref and response retrieval and runner-up for the remaining subtasks in DSTC10.

Despite the success in SIMMC 2.0, our approach has a few limitations. Most notably, our approach cannot be applied to cases with novel objects at inference, i.e. the objects that don't appear in the database at training. As such, it relies on latent object features learned from linguistic description for retrieving the requested object attributes. Our method also does not fully capture the semantic locality of objects within the scene (e.g. on the table, in the closet, etc.). We believe that these limitations can be addressed by training with a larger amount of data and including visual features in the multi-modal context as part of the input to the transformer.

Acknowledgements

This work was supported by the Ministry of Science and Information communication Technology (MSIT) of Korea (IITP No. 2019-0-00075)and the ETRI(Contract No. 21ZS1100).

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433.

Bengio, Y.; Ducharme, R.; Vincent, P.; and Janvin, C. 2003. A Neural Probabilistic Language Model. *J. Mach. Learn. Res.*, 3: 1137–1155.

Bordes, A.; and Weston, J. 2016. Learning End-to-End Goal-Oriented Dialog. *CoRR*, abs/1605.07683.

Chen, Y.-C.; Li, L.; Yu, L.; El Kholy, A.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*, 104–120. Springer.

Ham, D.; Lee, J.-G.; Jang, Y.; and Kim, K.-E. 2020. End-to-end neural pipeline for goal-oriented dialogue systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 583–592.

Henderson, M.; Thomson, B.; and Young, S. J. 2013. Deep Neural Network Approach for the Dialog State Tracking Challenge. In *Proceedings of the SIGDIAL 2013 Conference, The 14th Annual Meeting of the Special Interest Group on Discourse and Dialogue, 22-24 August 2013, SUPELEC, Metz, France*, 467–471. The Association for Computer Linguistics.

Holtzman, A.; Buys, J.; Du, L.; Forbes, M.; and Choi, Y. 2020. The Curious Case of Neural Text Degeneration. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Hosseini-Asl, E.; McCann, B.; Wu, C.; Yavuz, S.; and Socher, R. 2020. A Simple Language Model for Task-Oriented Dialogue. In Larochelle, H.; Ranzato, M.; Hadsell, R.; Balcan, M.; and Lin, H., eds., *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Humeau, S.; Shuster, K.; Lachaux, M.; and Weston, J. 2020. Poly-encoders: Architectures and Pre-training Strategies for Fast and Accurate Multi-sentence Scoring. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Johnson, J.; Hariharan, B.; Van Der Maaten, L.; Fei-Fei, L.; Lawrence Zitnick, C.; and Girshick, R. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2901–2910.

Joshi, C. K.; Mi, F.; and Faltings, B. 2017. Personalization in Goal-Oriented Dialog. *CoRR*, abs/1706.07503.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In Moens, M.; Huang, X.; Specia, L.; and Yih, S. W., eds., *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, 4903–4912. Association for Computational Linguistics.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1): 32–73.

Le, H.; Sahoo, D.; Chen, N. F.; and Hoi, S. C. H. 2019. Multimodal Transformer Networks for End-to-End Video-Grounded Dialogue Systems. In Korhonen, A.; Traum, D. R.; and Màrquez, L., eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics,*

ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers, 5612–5623. Association for Computational Linguistics.

Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J. R., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, 7871–7880. Association for Computational Linguistics.

Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, 121–137. Springer.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, B.; and Lane, I. R. 2016. Attention-Based Recurrent Neural Network Models for Joint Intent Detection and Slot Filling. In Morgan, N., ed., *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, 685–689. ISCA.

Loshchilov, I.; and Hutter, F. 2018. Fixing weight decay regularization in adam.

Moon, S.; Kottur, S.; Crook, P. A.; De, A.; Poddar, S.; Levin, T.; Whitney, D.; Difranco, D.; Beirami, A.; Cho, E.; Subba, R.; and Geramifard, A. 2020. Situated and Interactive Multimodal Conversations. In Scott, D.; Bel, N.; and Zong, C., eds., *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, 1103–1121. International Committee on Computational Linguistics.

Mrksic, N.; Séaghdha, D. Ó.; Wen, T.; Thomson, B.; and Young, S. J. 2017. Neural Belief Tracker: Data-Driven Dialogue State Tracking. In Barzilay, R.; and Kan, M., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, 1777–1788. Association for Computational Linguistics.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I.; et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8): 9.

Rezatofighi, S. H.; Milan, A.; Shi, Q.; Dick, A. R.; and Reid, I. D. 2018. Joint Learning of Set Cardinality and State Distribution. In McIlraith, S. A.; and Weinberger, K. Q., eds., *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, 2018*, 3968–3975. AAAI Press.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in neural information processing systems*, 5998–6008.

Wen, T.; Miao, Y.; Blunsom, P.; and Young, S. J. 2017. Latent Intention Dialogue Models. In Precup, D.; and Teh, Y. W., eds., *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, 3732–3741. PMLR.

Wen, T.-H.; Gasic, M.; Mrksic, N.; Su, P.-H.; Vandyke, D.; and Young, S. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. *arXiv preprint arXiv:1508.01745*.

Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; and Brew, J. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *CoRR*, abs/1910.03771.

Yang, S.; Zhang, R.; and Erfani, S. M. 2020. GraphDialog: Integrating Graph Knowledge into End-to-End Task-Oriented Dialogue Systems. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, 1878–1888. Association for Computational Linguistics.

Yang, Y.; Li, Y.; and Quan, X. 2021. UBAR: Towards Fully End-to-End Task-Oriented Dialog System with GPT-2. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 14230–14238. AAAI Press.

Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; and Gao, J. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5579–5588.

Zhou, K.; Zhao, W. X.; Bian, S.; Zhou, Y.; Wen, J.-R.; and Yu, J. 2020. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 1006–1014.