Exploring Prompt-based Multi-task Learning for Multimodal Dialog State Tracking and Immersive Multimodal Conversation

Yirong Chen[1]*, **Ya Li**[2]†, **Tao Wang**[2], **Xiaofen Xing**[1], **Xiangmin Xu**[3,4],
Quan Liu[2,5], **Cong Liu**[2,6], **Guoping Hu**[2,5]
[1]Guangdong Provincial Key Laboratory of Human Digital Twin, School of EE.,
South China University of Technology, Guangzhou, China
[2]iFLYTEK Research, Hefei, China [3]Pazhou Lab, Guangzhou, China
[4]School of Future Technology, South China University of Technology, Guangzhou, China
[5]State Key Laboratory of Cognitive Intelligence, Hefei, China [6]National Engineering
Research Center of Speech and Language Information Processing, Hefei, China
eeyirongchen@mail.scut.edu.cn, {yali8, taowang49}@iflytek.com
{xfxing, xmxu}@scut.edu.cn, {quanliu, congliu2, gphu}@iflytek.com

Abstract

With the rise of the metaverse, immersive multimodal conversation has attracted more and more researchers' attention. Multimodal contexts will become more important for human-computer interaction in the metaverse, especially in shopping domain. Unlike traditional conversation tasks, immersive multimodal conversation has challenges such as multimodal ambiguous candidate identification and multimodal coreference resolution, which makes it more difficult to dialog state tracking and response generation, as described in SIMMC 2.1 challenge, a part of DSTC11. In particular, as the number of objects in the scene increases, the difficulty will increase dramatically. We proposed PMTLED (**P**rompt-based **M**ulti-**T**ask **L**earning **E**ncoder-**D**ecoder), in which different subtasks use different prompts to make the model tend to focus on the current subtask. We achieve the winner in ambiguous candidates indentification and runner-up in multimodal coreference resolution (MM-Coref), multimodal dialog state tracking (MM-DST) and assistant response generation. Our code and model are made publicly available at https://github.com/scutcyr/dstc11-simmc2.1-scut-bds-lab.

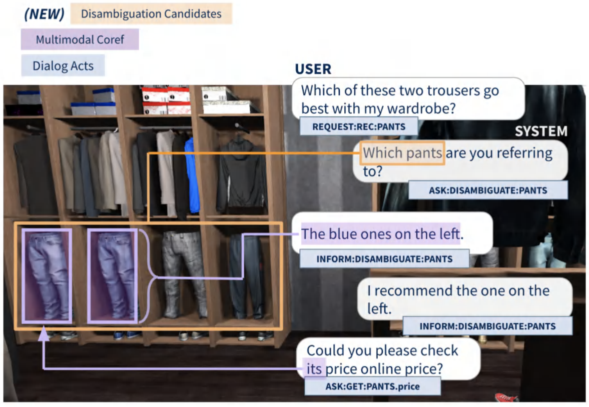

(NEW) Disambiguation Candidates
Multimodal Coref
Dialog Acts
USER
Which of these two trousers go best with my wardrobe?
REQUEST:REC:PANTS
SYSTEM
Which pants are you referring to?
ASK:DISAMBIGUATE:PANTS
The blue ones on the left.
INFORM:DISAMBIGUATE:PANTS
I recommend the one on the left.
INFORM:DISAMBIGUATE:PANTS
Could you please check its price online price?
ASK:GET:PANTS.price

Figure 1: Overview of SIMMC2.1 challenge. The user ambiguously uses 'these two trousers', so that the system needs to identify all the ambiguous objects on the scene to help further disambiguation and coreference resolution.

1 Introduction

With the rise of the metaverse(Mystakidis, 2022) and virtual reality (VR), immersive multimodal conversation has attracted more and more researchers' attention. Unlike traditional conversation tasks, immersive multimodal conversation has challenges such as multimodal ambiguous candidate identification and multimodal coreference resolution, which makes it difficult to dialog state tracking and response generation. To this end, Facebook released multimodal conversational dataset (SIMMC 2.1) and proposed the SIMMC 2.1 challenge (Kottur et al., 2021), in which the virtual assistant shares the same scene with the user.

In the past, the SIMMC 1.0 (Moon et al., 2020) challenge provided the controllable and sanitized multimodal contexts, while the SIMMC2.0 (Kottur et al., 2021) challenge provied the cluttered and closer-to-real-world multimodal contexts, in which there is no clear correspondence between textual context and objects in the scene. On the basis of the SIMMC 1.0 and SIMMC 2.0, SIMMC 2.1 proposed a new subtask (ambiguous candidate identification) in order to attract more attention on the key challenge of fine-grained visual disambiguation, as shown in Figure 1. Different from the SIMMC 2.0, SIMMC 2.1 has been annotated with additional labels (i.e. identification of all possible

*Work done during an internship at iFLYTEK Research.
†Corresponding author. Email: yali8@iflytek.com

Proceedings of The Eleventh Dialog System Technology Challenge, pages 1–8
September 11, 2023 ©2023 Association for Computational Linguistics

referent candidates given ambiguous mentions). In particular, the number of objects in a scene ranges from 6 to 141. Each object is different due to its visual or non-visual attributes. There are multiple objects with one same attribute in the scene, but the number of user-mentioned objects is less than the count in the scene, which makes the virtual assistant unable to identify the specific object that the user refers to, thus causing ambiguity.

The SIMMC2.1 challenge has four different subtasks, which are ambiguous candidates indentification (ACI), multimodal coreference resolution (MM-Coref), multimodal dialog state tracking (MM-DST) and assistant response generation. Due to the introduction of ACI, the other three subtasks are affected accordingly. In addition, ambiguous candidates and multimodal coreferential objects do not appear at the same time based on statistical analysis of the dataset.

The state-of-the-art method in the previous SIMMC challenge (Lee et al., 2022a,b) adopt BART and joint-learning approach on all subtasks. However, the performance of the original framework will be limited after adding the ambiguous candidates indentification task. In addition, The model also ignores the ground truth APIs in terms of response generation performance. Modeling MM-DST tasks as sequential prediction tasks is also easy to cause instability in prediction. For example, the model may forget to generate "<EOB>" tokens or ")" in some test samples. In order to consider both the performance and robustness of the model, we propose a prompt-based multi-task learning Transformer framework to address the above problems. In particular, ACI, MM-Coref and MM-DST share the same prompt, while the assistant response generation task uses another set of prompts alone, which will make full use of ground-truth APIs when generating responses.

Our model was declared as the **winner** of the subtask 1 (ambiguous candidates indentification) with 70.50% ambiguous object identification F1. Moreover, our model was declared as the **runner-up** in the official evaluation on all other subtasks, in which we achieved 80.28% coreference F1, 92.66% slot F1, 97.75% intent F1 (rank #1) and 0.3650 BLEU-4.

2 Related Work

2.1 Multi-task Learning for Task-oriented Dialog System

Task-oriented dialog systems have explicit goals (e.g. request to compare, request to get, etc.), making dialog understanding important before generating response. When using multi-task learning, the subtasks related to dialogue understanding and dialogue generation in task-oriented dialog system can be modeled into one model. In the Encoder-Decoder (e.g. Transformer (Vaswani et al., 2017)) or UniLM (Dong et al., 2019) framework, the classification tasks can be further modeled based on the output of Encoder, and the generation tasks can be further modeled based on Decoder. In recent years, there has been an increasing amount of literature on multi-task learning for task-oriented dialog system (Zhao et al., 2022; Su et al., 2022). Recently, a unified dialog model named SPACE-3 (He et al., 2022) has verified that the performance of various tasks can be significantly improved by conducting multi-task joint pretraining on large-scale task-oriented dialog corpus, in which the authors proposed 5 tasks, including span masked language modeling, understanding semantic modeling, semantic region modeling, policy semantic modeling and response generation modeling. In the SIMMC 2.1 challenge of DSTC-10, multi-task learning has been verified to be effective for task-oriented dialog system (Lee et al., 2022a; Nguyen et al., 2022).

2.2 Prompt Learning

With the rise of large-scale pre-trained language models (Devlin et al., 2019; Han et al., 2021), prompt learning has recently been widely studied by the NLP community, e.g. AutoPrompt (Shin et al., 2020), Prefix-Tuning (Li and Liang, 2021) and etc. Both discrete and continuous prompts are widely used to solve downstream tasks with pre-trained models. Some prompt-based works are proposed for task-oriented dialog system, such as UniTranSeR (Ma et al., 2022), Cins (Mi et al., 2022), SPACE-3 (He et al., 2022) and etc. SPACE-3 employed two kinds of prompts to extract semantics with three subtasks for helping pass the task-flow in a task-oriented dialog system. We use prompts to model MM-DST task as several different prompt-based classification tasks and used the ground-truth APIs to design the prompts for response generation.

Total # dialogs	11,244
Total # utterances	117,236
Total # scenes	3,133
Avg # words per user turns	12
Avg # words per assistant turns	13.7
Avg # utterances per dialog	10.4
Avg # objects mentioned per dialog	4.7
Avg # objects in scene per dialog	19.7
Avg # candidates per ambiguous turn	5.6

Table 1: Summary of SIMMC 2.1 dataset statistics.

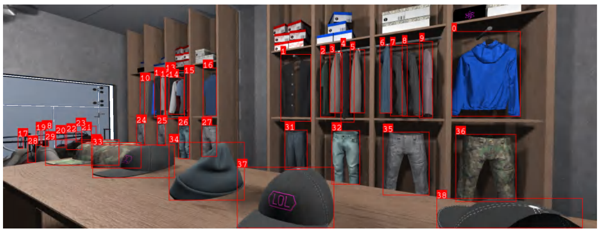


Figure 2: An example of the scene.

3 SIMMC2.1 Dataset

SIMMC 2.1[1] is a dataset for studying immersive multimodal conversations in the form of a co-observed image or virtual reality (VR) environment(Kottur et al., 2021) in two shopping domain: furniture (4k dialogs) and fashion (7.2k dialogs). Different from other task-oriented dialog datasets. SIMMC 2.1 is highly structured with abundant annotation information, including dialog act, slot values, coreference objects, ambiguous candidate objects, scene images, metadata of all objects, bounding boxes and etc. The statistics of SIMMC 2.1 is shown in Table 1. For the SIMMC 2.1 challenge, the dataset is split into 4 sets: train, dev, devtest and teststd according to the ratio of 6.4:0.5:1.5:1.5. Each turn of dialog has a corresponding scene image (e.g., Figure 2), the objects in which have corresponding canonical ID(s) for subtasks ACI or MM-Coref. In particular, all the subtasks are prohibited from using visual metadata of the objects, but non-visual metadata is allowed to use. Both subtask ACI and MM-Coref formally look for subsets from a set of objects and return them as prediction results. The number of ambiguous candidates ranges from 2 to count of all objects in a scene image, while the number of coreference objects is commonly less than 4.

4 Method

In this section, we demonstrate the implementation of the proposed PMTLED for solving each subtask. The architecture for subtasks 1, 2 and 3 is shown in Figure 3 (a), and the architecture for subtask 4 is shown in Figure 3 (b). We first introduce the input of the model and the model architecture. Then we elaborate on objectives of each subtask and auxiliary tasks.

4.1 Model Architecture

Similar to (Lee et al., 2022a), We choose BART (Lewis et al., 2020) as our backbone model for both dialog understanding tasks and dialog generation task. BART contains a Transformer Encoder and a Transformer Decoder, which is particularly effective when fine tuned for both dialog understanding and response generation. Let $\mathcal{D}$ denotes a dialog with L_{turn} turns in SIMMC2.1 dataset. $\mathcal{D}$ can be defined as:

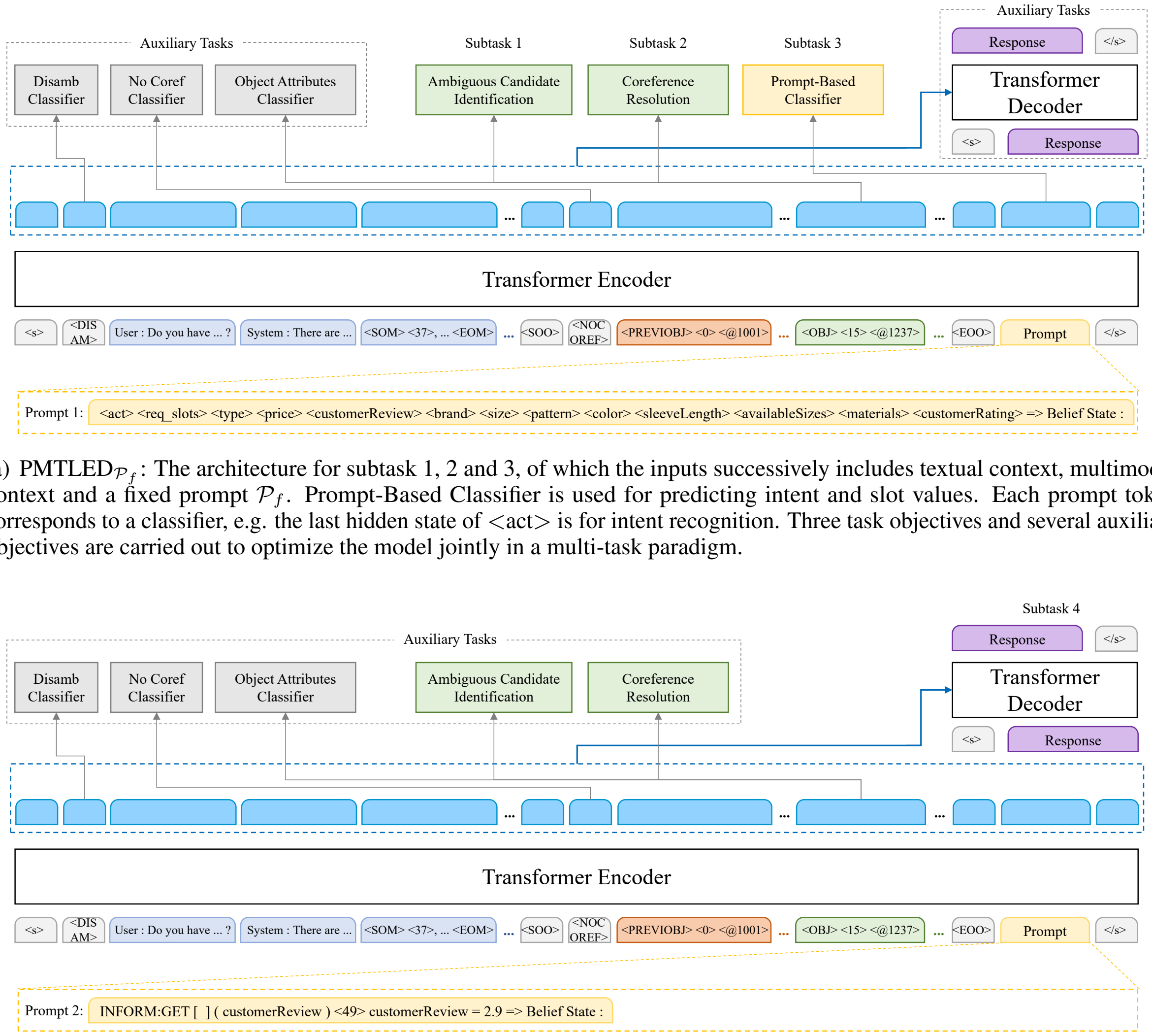
$$\mathcal{D} := \{(U_i, A_i, M_i, B_i^u, B_i^a, S_i)\}_{i=1}^{L_{turn}} \quad (1)$$

where U_i and A_i are the user and system utterances at turn i, M_i is the multimodal context that consists of a set of object indices mentioned by the system, B_i^u and B_i^a are the user and system belief states, S_i is the scene context including both the scene image and all objects in it.

4.2 Input Representation

As shown in Figure 3, the input is concatenated as in Equation 2. We adopt the same separator tokens and other special tokens described in (Lee et al., 2022a), including "<DISAM>", "<NOCOREF>", "<SOM>", "<EOM>", "<SOO>", "<EOO>", "<OBJ>", "<PREVIOBJ>" and etc. The objects in the scene is represented by their canonical object ID tokens with the form of “<obj_scene_index>" (e.g. “<1>") and their unique meta ID with the form of “<obj_meta_id>" (e.g. “<@1001>"). In addition, in order to obtain the better representation of scene context S_i, the embedding of the objects is added with the representation of the bounding box information and all the embedding of non-visual attributes (*customerReview*, *brand*, *price*, *size*, *materials*). We follow the object box embedding method used in (Lee et al., 2022a) to obtain representation of the bounding box through the encoding process $(x_1/w - 0.5, y_1/h - 0.5, x_2/w - 0.5, y_2/h - 0.5, (x_2 - x_1)(y_2 - y_1)/(h \cdot w))$ and then pass it

[1] https://github.com/facebookresearch/simmc2



(a) PMTLED$_{\mathcal{P}_f}$: The architecture for subtask 1, 2 and 3, of which the inputs successively includes textual context, multimodal context and a fixed prompt $\mathcal{P}_f$. Prompt-Based Classifier is used for predicting intent and slot values. Each prompt token corresponds to a classifier, e.g. the last hidden state of <act> is for intent recognition. Three task objectives and several auxiliary objectives are carried out to optimize the model jointly in a multi-task paradigm.

(b) PMTLED$_{\mathcal{P}_d}$: the architecture for subtask 4, of which the inputs successively includes textual context, multimodal context and a dynamic prompt $\mathcal{P}_d$. For the system response generation task, the system's belief state is designed as prompt. Three task objectives and several auxiliary objectives are carried out to optimize the model jointly in a multi-task paradigm.

Figure 3: Overview of multi-task jointly fine-tuning and prompt-tuning for each subtask.

to a fully-connected layer followed by LayerNorm by its upper-left vertex (x_1, y_1), lower-right vertex (x_2, y_2), height h and width w.

$$input_i = [(U, A, M)_{i-\ell:i}, U_i, S_i, P_i] \tag{2}$$

where $(U, A, M)_{i-\ell:i}$ are the previous user utterances, system utterances and multimodal context with dialog history up to ℓ turns to limit the length of input. P_i is the prompt designed at turn i.

In this work, we designed the fixed prompt $\mathcal{P}_f$ for MM-DST and the dynamic prompt $\mathcal{P}_d$ for response generation.

4.3 Fixed Prompt for MM-DST

In SIMMC 2.0, the MM-DST is universally considered as an auto-regressive language modeling task (Lee et al., 2022a; Nguyen et al., 2022), which makes the model unstable in the inference process. Therefore, we adopt a fixed prompt $\mathcal{P}_f$= "<act> <req_slots> <type> <price> <customerReview> <brand> <size> <pattern> <color> <sleeveLength> <availableSizes> <materials> <customerRating>", the output of which are used for classification for MM-DST. For example, the last hidden state of the prompt token "<act>" is passed into a classifier for intent recognition. All the prompt-based classification tasks can be expressed as:

$$\boldsymbol{\mathcal{C}_{p_t}} = \boldsymbol{W_{p_t}} \boldsymbol{h^{\ell}_{p_t}} + \boldsymbol{b_{p_t}}, p_t \in \mathcal{P}_f \tag{3}$$

where $h^{\ell}_{p_t}$ denotes the last hidden state of the prompt token p_t, $\boldsymbol{W_{p_t}} \in \mathbb{R}^{n \times d}$ and $\boldsymbol{b_{p_t}} \in \mathbb{R}^n$ are the trainable parameters. n is the number of categories for the classification task on the prompt token p_t. d is the dimension of the hidden states.

We denote the prompt-based MM-DST classification loss $\mathcal{L}_{dst}$ for all prompt token in $\mathcal{P}_f$ by

$$\mathcal{L}_{dst} = -\sum_{p_t \in \mathcal{P}_f} \sum_{c \in \mathcal{C}_{p_t}} \mathbb{L}\{c = y_{p_t}\} \log P(c) \quad (4)$$

where $\mathcal{C}_{p_t}$ is the set of all classes on the classification task of prompt token p_t, y_{p_t} the label, and $\mathbb{L}$ is an indicator function.

4.4 Dynamic Prompt for Response Generation

Since the annotation information *system_transcript_annotated* of current turn is allow to use in subtask response generation, we use *act*, *slot_values* and *request_slots* to design the dynamic prompt $\mathcal{P}_d$ in both the training and testing phases. Key-value pairs similar to *"Object ID: 36": {"customerReview": 4.3}* will be characterized as "*<36> customerReview = 4.3*", which makes the representation of the objects consistent with that of scene context S_i. In the end, the process of response generation is split into two phases: (i) dynamic prompt generation driven by *system_transcript_annotated* and (ii) response generation driven by both the multimodal context and dynamic prompt. Then the cross-entropy loss is employed to calculate the subtask loss:

$$\mathcal{L}_{lm} = -\sum_{i=1}^{L} \log P(t_i | t_1, t_2, \cdots, t_{i-1}) \quad (5)$$

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

4.5 Classification for ACI and MM-Coref

Both subtsk 1 and subtask 2 can be formulateed as a binary classification on all objects in the scene image. For the obj_i in the scene, the last hidden states of the canonical object ID token and its unique meta ID token is concatenated to obtain the joint representation h_{obj_i} as follows:

$$\boldsymbol{h_{obj_i}} = concat(\boldsymbol{h^{\ell}_{pos_i}}, \boldsymbol{h^{\ell}_{pos_i+1}}) \quad (6)$$

where pos_i is the position of the canonical object ID token in the input sequence, $h^{\ell}_{pos_i}$ the last hidden state of the canonical object ID token, $h^{\ell}_{pos_i+1}$ the last hidden state of the unique meta ID.

Then, the joint representation h_{obj_i} is passed to a ACI classifier (see Equation 7) and MM-Coref classifier (see Equation 8) to predict true or false. Both the ACI classifier and MM-Coref classifier are fully-connected layers. Then, the ACI loss $\mathcal{L}_{aci}$ and MM-Coref loss $\mathcal{L}_{coref}$ can be calculated by using cross-entropy loss.

$$\boldsymbol{C_{ACI}} = \boldsymbol{W_{ACI} h_{obj_i}} + \boldsymbol{b_{ACI}} \quad (7)$$

$$\boldsymbol{C_{coref}} = \boldsymbol{W_{coref} h_{obj_i}} + \boldsymbol{b_{coref}} \quad (8)$$

where $\boldsymbol{W_{ACI}} \in \mathbb{R}^{2\times 2d}$, $\boldsymbol{b_{ACI}} \in \mathbb{R}^2$, $\boldsymbol{W_{coref}} \in \mathbb{R}^{2\times 2d}$ and $\boldsymbol{b_{coref}} \in \mathbb{R}^2$ are the trainable parameters.

4.6 Auxiliary tasks

Similar to (Lee et al., 2022a), we adopt three same auxiliary tasks for both the model of subtask 1 to 3 and the model of subtask 4: (i) Binary prediction for disambiguation with task objective $\mathcal{L}_{disamb}$, (ii) Binary prediction for empty coreference set with task objective $\mathcal{L}_{nocoref}$ and (iii) Prediction for object attributes with task objective $\mathcal{L}_{attr}$ as shown in Equation 9.

$$\mathcal{L}_{attr} = -\sum_{j \in \mathcal{O}_s} \sum_{n=1}^{N} \sum_{c \in \mathcal{C}_n} \mathbb{L}\{c = y_{jn}\} \log P(c) \quad (9)$$

where $\mathcal{O}_s$ is the set of objects in the scene history, N the number of attributes, $\mathcal{C}_n$ the set of all classes of the n-th attribute, y_{jn} the label of the n-th attribute of the j-th object, and $\mathbb{L}$ is an indicator function.

The auxiliary objective can be calculated by

$$\begin{aligned}\mathcal{L}_{aux} = &\lambda_{attr}\mathcal{L}_{attr} + \lambda_{disamb}\mathcal{L}_{disamb} \\ &+ \lambda_{nocoref}\mathcal{L}_{nocoref}\end{aligned} \quad (10)$$

where λ_{attr}, λ_{disamb} and $\lambda_{nocoref}$ are the hyperparameters.

4.7 Training Objective

As shown in Figure 3, the model for subtask 1 to 3 and the model for subtask 4 has different multi-task join objective. To sum up, the overall training objective of the model for subtask 1 to 3 $\mathcal{L}_{subtask123}$ can be defined as follows:

$$\begin{aligned}\mathcal{L}_{subtask123} = &\lambda_{aci}\mathcal{L}_{aci} + \lambda_{coref}\mathcal{L}_{coref} \\ &+ \lambda_{dst}\mathcal{L}_{dst} + \lambda_{lm}\mathcal{L}_{lm} + \mathcal{L}_{aux}\end{aligned} \quad (11)$$

where λ_{aci}, λ_{coref}, λ_{dst} and λ_{lm} are the hyperparameters.

Toward the model for subtask 4, the overall training objective is defined as:

$$\begin{aligned}\mathcal{L}_{subtask4} = &\lambda_{aci}\mathcal{L}_{aci} + \lambda_{coref}\mathcal{L}_{coref} \\ &+ \lambda_{lm}\mathcal{L}_{lm} + \mathcal{L}_{aux}\end{aligned} \quad (12)$$

Team ID	Models	Subtask #1	Subtask #2	Subtask #3		Subtask #4
		Object F1	Object f1	Slot F1	Act. F1	BLEU-4
0(baseline)	GPT2	18.00%	26.50%	73.50%	93.00%	0.192
	GPT-2(MM)	43.20%	-	-	-	-
	BERT(MM)	43.90%	-	-	-	-
	MTN	-	-	75.00%	94.30%	0.210
1	Longformer	67.26%†	-	-	-	-
	Longformer	-	**94.29%**	-	-	-
	Longformer	-	-	**94.24%**	95.98%	-
	OFA	-	-	-	-	**0.4093**
2	CoCondenser	65.17%	-	-	-	-
3	ALBEF	63.84%	75.85%	-	-	-
	BART	-	-	90.48%	96.77%	0.3029
4	COMBINER	-	-	-	-	0.2519
5(ours)	PMTLED$_{\mathcal{P}_f}$	**70.50%**	80.28%†	92.66%†	**97.30%**	-
	PMTLED$_{\mathcal{P}_d}$	-	-	-	-	0.3650†

Table 2: The official leaderboard of DSTC11 SIMMC 2.1 Challenge on the teststd set. The subtask winners are bold-faced and runner-ups are marked with †. "-" means that the model did not participate in that subtask.

5 Experiments

5.1 Settings and Hyperparameters

The BART-large[2] (Lewis et al., 2020) model is used as the backbone. The whole implementation was based on the *Huggingface Transformers* (Wolf et al., 2020)[3]. The maximum length of dialog turns L_{turn} is set to 6 and the max sequence length is 1,024. The model is finetuned for 30 epochs with an initial learning rate 5e-5 and a batch size of 16 per GPU with AdamW optimizer (Loshchilov and Hutter, 2018). Besides, the linear warmup schedule with warmup ratio of 0.1 and clip gradient norms of 1.0 are equipped. The weight decay and dropout rate are set to 0.1. The value of λ_{aci}, λ_{coref}, λ_{dst} and λ_{lm} are set to 1.0, while the value of λ_{attr}, λ_{disamb} and $\lambda_{nocoref}$ are set to 0.1 based on the performance in validation set. We finetuned all models on a Linux server with Centos and 4 GPU of NVIDIA Tesla A100.

5.2 Results and Analysis

The results on the teststd set is shown in Table 2. The proposed model mt-bart-dstcla was declared as the winner of the subtask 1 with 70.50% ambiguous object identification F1, and was declared as the runner-up of the subtask 2 and 3 with 80.28% coreference F1 in MM-Coref, 92.66% slot F1 and 97.30% Act. F1 in MM-DST. The proposed dynamic prompt-based multi-task model mt-bart-sys-ensemble was declared as the runner-up with 0.3650 BLEU-4 in response generation. For comparison, the Entry #1 used separate models Longformer (Beltagy et al., 2020) for subtask 1, 2 and 3 and adopted the mode of jointly multi-task training of current task and auxiliary tasks. In particular, for Task 3, Entry #1 adopted a fixed prompt for MM-DST task, similar to our approach. Different from us, Entry #1 has designed two different fixed prompts and adopted different classifiers for two different domain: furniture and fashion. This has resulted in higher Act. F1 (1.32%↑) for our proposed model and higher Slot F1 (1.58%↑) for the model of Entry #1. Torward subtask 4, both Entry #1 and us bsed the *system_transcript_annotated* to design the dynamic prompt for response generation. However, we do not use the mentioned objects of the current turn, while Entry #1 use the mentioned objects and their non-visual attributes, which is one of the core factors that they get a higher BLEU-4 (0.0443↑). As for subtask 1, Entry #1 only use the samples with *disambiguation_label*=1 to finetuned the ACI classifier, in which other subtasks do not participate in joint fine-tuning process. Therefore, we finally achieved the higher ambiguous object identification F1 (3.24%↑).

[2] https://huggingface.co/facebook/bart-large

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

Models	Subtask #1	Subtask #2	Subtask #3		Subtask #4
	Object F1	Object f1	Slot F1	Act. F1	BLEU-4
baseline (Lee et al., 2022a)	-	75.9%	90.%	97.4%	0.3193
PMTLED$_{\mathcal{P}_f}$	68.5%	78.4%	92.0%	97.5%	-
w/o $\mathcal{P}_f$	65.7%	77.3%	88.4%	96.6%	-
PMTLED$_{\mathcal{P}_d}$	-	-	-	-	0.3874
w/o $\mathcal{P}_d$	-	-	-	-	0.3433

Table 3: Results on the devtest set. $\mathcal{P}_f$: fixed prompt. $\mathcal{P}_d$: dynamic prompt.

5.3 Ablation Study

We conduct ablation experiments to answer the questions: (i) Wherther the fixed prompt and dynamic prompt are useful? Table 3 provides detailed results on the devtest for our ablation study. The second row of results is obtained according to the hyperparameters setting of (Lee et al., 2022a). As shown in the row 3 and 4, the performance of the model without $\mathcal{P}_f$ degrades largely on subtask 1, 2 and 3, demonstrating the effectiveness of using the fixed prompt. According to the row 5 and 6, BLEU-4 significantly deteriorates if we discard the dynamic prompt $\mathcal{P}_d$, suggesting that dynamic prompt generated by *system_transcript_annotated* is crucial component for the response generation task.

6 Conclusion and Future Work

In this paper, we proposed a prompt-based multi-task learning method for multimodal dialog state tracking and immersive multimodal conversation. The proposed model used fixed prompt for multi-modal dialog state tracking and dynamic prompt for immersive multimodal conversation based on Transformer framework. Experiments demonstrated that the proposed fixed prompt and dynamic prompt are effective on jointly fine-tuning the model for the SIMMC 2.1 Challenge. In the future, we will continue exploring the design of the prompt, promoting the model to apply more knowledge as much as possible, e.g. the mentioned objects of current turn of the assistant, the scene image and etc. In addition, the current version of the model depends on domain information, which makes it difficult to apply to other scenarios. Further research needs to solve the domain dependency.

7 Acknowledgement

This work was supported by the Science and Technology Project of Guangzhou (202103010002), the Natural Science Foundation of Guangdong Province (2022A1515011588), the National Key R&D Program of China (2022YFB4500600), the Science and Technology Project of Guangdong (2022B0101010003), the National Natural Science Foundation of China under Grant U1801262 and Guangdong Provincial Key Laboratory of Human Digital Twin (2022B1212010004).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *CoRR*, abs/2004.05150.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Li Dong, Nan Yang, Wenhui Wang, Furu Wei, Xiaodong Liu, Yu Wang, Jianfeng Gao, Ming Zhou, and Hsiao-Wuen Hon. 2019. Unified language model pre-training for natural language understanding and generation. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xu Han, Zhengyan Zhang, Ning Ding, Yuxian Gu, Xiao Liu, Yuqi Huo, Jiezhong Qiu, Yuan Yao, Ao Zhang, Liang Zhang, Wentao Han, Minlie Huang, Qin Jin, Yanyan Lan, Yang Liu, Zhiyuan Liu, Zhiwu Lu, Xipeng Qiu, Ruihua Song, Jie Tang, Ji-Rong Wen, Jinhui Yuan, Wayne Xin Zhao, and Jun Zhu. 2021. Pre-trained models: Past, present and future. *AI Open*, 2:225–250.

Wanwei He, Yinpei Dai, Min Yang, Jian Sun, Fei Huang, Luo Si, and Yongbin Li. 2022. Unified dialog model pre-training for task-oriented dialog understanding and generation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 187–200, New York, NY, USA. Association for Computing Machinery.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Jinhyeon Kim, Youngjune Lee, Ran Han, Yoonhyung Kim, Minho Park, Kangwook Lee, Haebin Shin, et al. 2022a. Tackling situated multi-modal task-oriented dialogs with a single transformer model. In *DSTC 10 Workshop @ AAAI 2022*.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, and Kee-Eung Kim. 2022b. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830, Seattle, United States. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597, Online. Association for Computational Linguistics.

I Loshchilov and F Hutter. 2018. Fixing weight decay regularization in adam, iclr 2018 conference. In *Vancouver, BC, Canada: Vancouver Convention Center.[Google Scholar]*.

Zhiyuan Ma, Jianjun Li, Guohui Li, and Yongjing Cheng. 2022. UniTranSeR: A unified transformer semantic representation framework for multimodal task-oriented dialog system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 103–114, Dublin, Ireland. Association for Computational Linguistics.

Fei Mi, Yasheng Wang, and Yitong Li. 2022. Cins: Comprehensive instruction for few-shot learning in task-oriented dialog systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11076–11084.

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Stylianos Mystakidis. 2022. Metaverse. *Encyclopedia*, 2(1):486–497.

Thanh-Tung Nguyen, Wei Shi, Ridong Jiang, and Jung jae Kim. 2022. Multimodal and joint learning generation models for simmc 2.0. In *DSTC 10 Workshop @ AAAI 2022*.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4222–4235, Online. Association for Computational Linguistics.

Yixuan Su, Lei Shu, Elman Mansimov, Arshit Gupta, Deng Cai, Yi-An Lai, and Yi Zhang. 2022. Multi-task pre-training for plug-and-play task-oriented dialogue system. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4661–4676, Dublin, Ireland. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Meng Zhao, Lifang Wang, Zejun Jiang, Ronghan Li, Xinyu Lu, and Zhongtian Hu. 2022. Multi-task learning with graph attention networks for multi-domain task-oriented dialogue systems. *Knowledge-Based Systems*, page 110069.