Amirkabir University of Technology
(Tehran Polytechnic)

Mathematics and Computer Science Faculty
Computer Science Department

B.Sc. Thesis

Multi-domain Multi-Modal Task-Oriented Dialogue System

By
Mohammadreza Ardestani

Supervisor
Dr. Mohammad Akbari

Advisor
Dr. Mahdi Ghatee

April – 2022

Abstract

Humans tend to convey their intentions with different modalities, namely text, image, and audio, each of which is reach of effective information facilitating conversations. Current SOTA Task-oriented dialogue (TOD) systems are wanting robust multi-modality; we bolster them with this feature. Unlike many TOD systems that support single-domain datasets, we use the MMConv dataset, which has five domains, to increase the versatility of our system. MMTOD is our proposed system which adapts a GPT-2 decoder for managing dialogue and text-generation parts and a dual-encoder for its zero-shot image retrieval module. MMTOD outperforms baselines in image match rate by 49 percent and covers more multimodal queries that baseline methods cannot cover.

Key Words: Multi-Domain Dialogue, Multi-Domain Dialogue, Conversational AI, task-oriented dialogue systems, Chabot.

Contents

1) Introduction

Intelligent conversational agents have been humans' lofty dream for a long time and found paramount importance as they were used in Turing Test [Wikipedia: Turing test]. Their implemented versions started from ELIZA in the last century [Hussain et al, 2019] and continued to improve to the current successful systems, namely Xiaoice by Microsoft which has around 660 million users in the world [Fu et al, 2022].

Despite recent advances in methods and the enormous amount of data available for training systems, conversational agents are way far from human performance and there is no version of them capable of having general conversations (both task-oriented and non-task-oriented) or having coherent long conversations like humans. Current systems try to find a narrowed task and perform near humans in medium-length dialogues on that task. Since knowing the exact type of a dialogue system helps users utilize its all capability and guide the system developers for the technical standpoints, we demonstrated the type of our system in Fig 1, adapted from [Hussain et al, 2019].

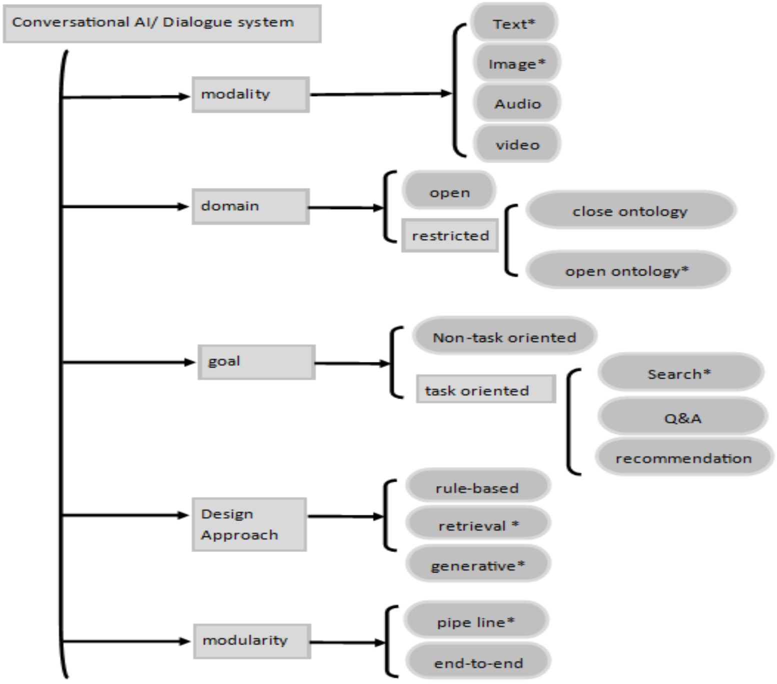


Fig1 Broad classification of dialogue systems: Our system type is marked with *. There are hybrid methods and rarely more modalities are used. Usually "Chabot" refers to open domain systems aiming to engage and inform users while "TOD system" refers to restricted-domain systems aimed at doing a specific task.

Our proposed system support text and image and performs on five domains with open ontology (implications of open ontology and having multi-domains has discussed later). Our model does not have an end-to-end design (Fig2. Visualize the design) since we use two separated modules for language modeling and image retrieval which language model module is generative and the image retrieval module is retrieval-based. Although dialogue systems might have the same agent-side modality, they can be very different in terms of architecture whether they use generative or retrieval-based methods for that specific modality.

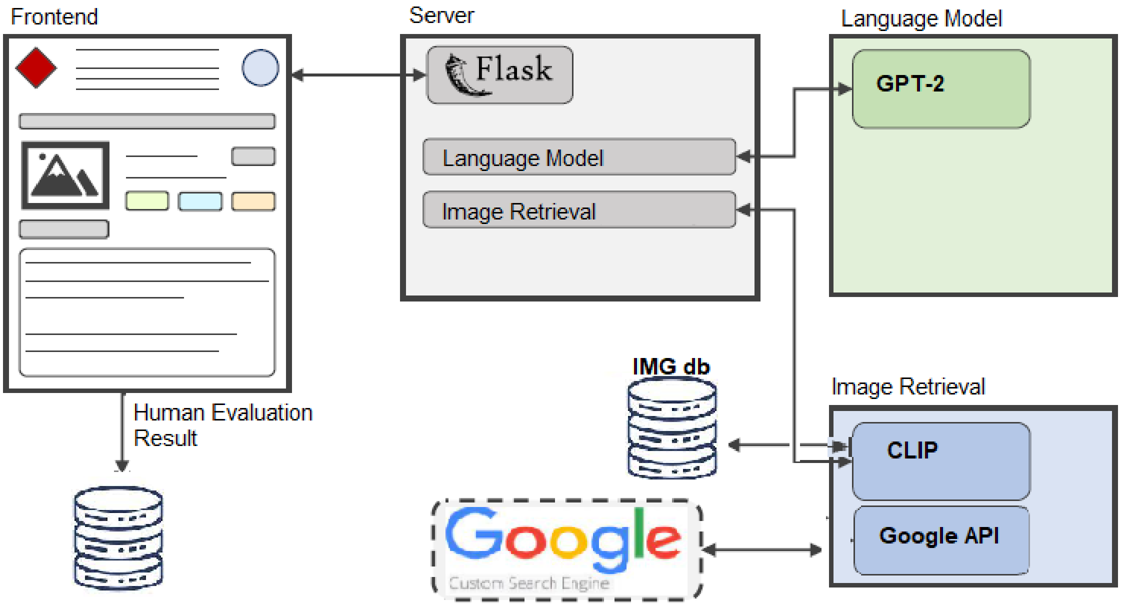


Fig2: Design of the proposed system

The highlights of our work are summarized as follows:

- MMTOD is the first model, to the best of our knowledge, achieves SOTA image match rate performance on the MMConv dataset [liao et al, 2021].
- A robust multi-modal TOD system in the absence of a large image dataset with external API image retrievals.
- Supporting more multi-modal involved scenarios than base lines.
- Discovered inconsistency and noisy labeling in the MMConv dataset [liao et al, 2021]. and provided a clean version of it at github.com/MMTOD.

2) Related Works

First, we investigate different related datasets and distinguish our dataset type and its conversation settings from others then discuss other methods on datasets related to ours. Since modality is a pivotal point in our comparison we have formulated agent-user interaction in Fig3 and used this later.

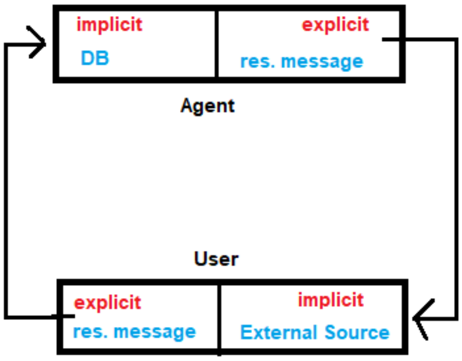


Fig3: Modality formulation. Users' implicit modalities do not need to enter into our equations since it does not affect the system architecture and it is up to users to handle it. CoDraw users [kim et al,2019], for example, describe a picture to the agent verbally; hence, we only need to consider handling textual response.

			Modality				
Dataset	Type	Domain	User res.	Agent res.	Agent DB	Open ontology	Reference
ViDA-MAN	Vis conv .search	Bank	a	v	t	idk	[shen et al, 2021]
MCIM	Vis Q&A	Assembly line	a,v	a,i	t,i,	idk	[chen et al, 2021]
MMConv	Conv. search	5 domains in travel	t,i	t,i	t,i	YES	[liao et al, 2021]
UniMF	Conv. search	Restaurant	t	t	t,i	idk	[yang et al, 2021]
MMD	Image search	Fashion	t	t,i	t,i	No	[saha et al,2018]
CoDraw	Image drawing	Common objects	t	t,i	t,i	N/A	[kim et al, 2019]
VisDial	Vis Q&A	Common objects	t	t	t,i	idk	[das et al, 2016]
MultiWoZ	Conv. search	Hotel, Taxi, etc.	t	t	t	idk	[Budzianowski et al, 2018]
Facebook	Conv. Rec	Movie	t	t	t	idk	[dodge et al, 2016]
SIMMC	Conv. search	Furniture, Fashion	t	t,i	t,i	idk	[moon et al, 2020]

Fig4: Broad dataset categorization. (A: audio). There is no consensus about the type of listed datasets. For exact functionality details about each one, we need reading original papers.

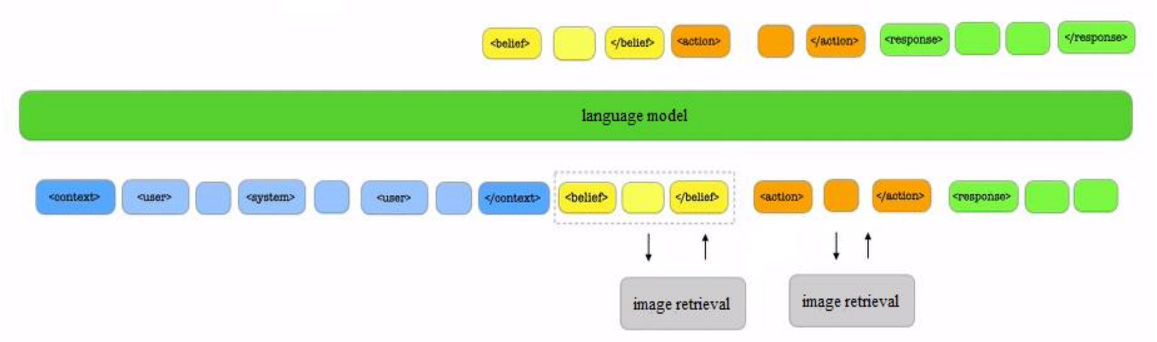
We present ten TOD datasets that are representative of a category in Fig4 which shows how different our dataset and the proposed method are from other datasets and their methods being proposed.

Based on the modality, type, and being multi-domain UniMF and MMD are the most similar dataset to the dataset we have used (MMConv). UniMF is not published by its authors. So, only use SOTA methods on MMConv, SimpleTOD, and MARCO, as our baseline. “User Attention-guided Multimodal Dialog Systems” [cui et al, 2019], implemented on the MMD dataset, can be compared to our method, although it does not support open ontology.

3) Proposed Method

TOD systems can be decomposed into three tasks; dialogue state tracking (DST), system action prediction, and producing system response. Modern end-to-end designs try to cast these steps into one sequence-to-sequence language modeling. We adapt GPT-2 [Radford et al, 20], inspired by [hosseini-asl et al, 2021], for this sequence prediction task, yet in a non-end-to-end manner in general.

Pre-trained Language Modeling We feed GPT-2 with context that encompasses all previous user and agent responses and gets a delexicalized agent response at the end. There are also two intermediary sequences that guide response generation: belief state sequence, and action sequence. Belief states are the distillation of all previous utterances and action sequences including actions that should be taken by the agent in the next utterance and response is a human-understandable format of actions predicted by the system in the English language. Due to our multimodal setting, whenever we have “image inform” or “image request” in the belief state sequence or the action sequence we need to call the image retrieval module to return the result and replace the correct values into belief and actions currently being produced. This separated image retrieval module makes our proposed architecture.



Zero-shot image retrieval Many image retrieval approaches, namely [cui et al, 2019], use classification to find the relevant images in multi-modal dialogue settings. This approach works fine when we have predefined and certain classes, in our terms “fixed ontology”, while these methods fall short in open-ontology settings, like ours. To solve this issue, we adopt SOAT pre-trained zero-shot classifiers CLIP [redford et al, 2021] that suits our open-ontology travel-domain image retrieval since CLIP is also trained on similar domain images from social media and webpages on the internet.

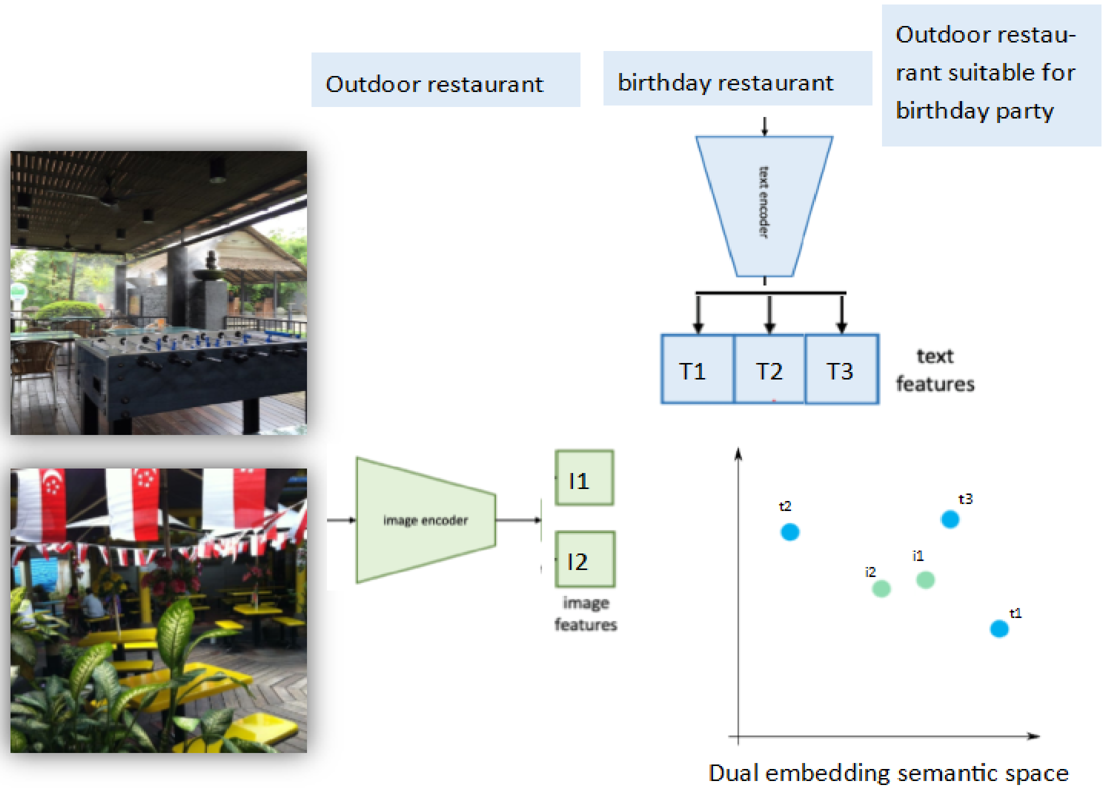


Fig 6 CLIP text and image encoders. (pictures from MMConv image dataset and the 2 D dual embedding semantic space is generated with initial 512 D feature vectors by PCA method [wikipedia: PCA] in sklearn library[1])

/https://scikit-learn.org [1]

As demonstrated in Fig 6, we use CLIP dual encoders, heavily inspired by haltakov[2] in our implementation, to map images and textual inquiries in a dual embedding semantic space in which we can retrieve images by a textual query, and vice versa.

Architecture In our multi-modal setting there are several image-involved scenarios that affect the architecture and are listed as follows:

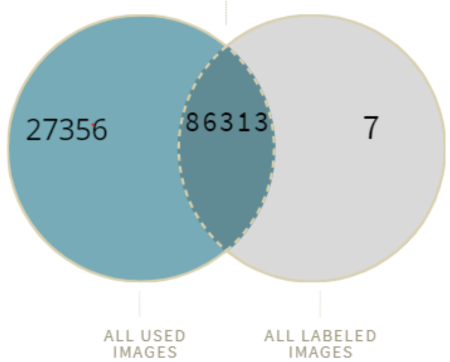
1. Agent sends the most relevant image after every recommendation.
2. Agent sends the most relevant image when the user asks (user can ask different questions like the venue's foods, drinks, night view, etc.).
3. Agent can also send an image at their own discretion.
4. User sends an image and asks the agent to find the "venue name" of its place or to find the concept (name/label) of its object (generally food).
5. User asks the agent to find a venue with the ambiance of the attached image. (eg: the user sends an image of an open field or river-side place and asks the agent to find the with these qualities)

The first and last items are extra scenarios supported by our proposed system and the rest already exist in our dataset. The first scenario is covered by adding "image inform" after every recommendation in our train dataset so that we fully utilize the multi-modality of our system. The fifth scenario, which mostly all classification-based image retrieval approaches fail to fully answer it, is handled in our system by adding encoding image and textual query and then search through image dataset with this enriched query vector. The system approaches other scenarios similarly.

Training Details We trained our We did not fine-tune CLIP and use the initial weights in our system, one account of computational resource constraints.

Dataset details There are four primary files dioaluge.json, image_classes.json, db.json, and image files.

https://github.com/haltakov/natural-language-image-search [2]



Considerable number of images used in dialogues has no label and this is problematic for the evaluation part. Furthermore, there are two conflict between labels used in dialogue file and the image_classes file (dialogue 5447, fourth and sixth turns), and the label “fried kway teow mee” is in capital words in the dialogue file, unlike the image_class file. These raise a question about the exact labeling process in the MMConv [ref] paper and the calculation of the image match rate.

There are a lot of attributes for venues in db file, for example, collected users’ views which can be utilized, however, we left this option for future works.

As described in the dataset paper, agents could respond to users freely and this flexibility in response might cause a problem for the language model to find a response pattern. In dialogue 4902, for instance, the agent sends images incoherent and inconsistent with that dialogue flow.

4) Evaluation

Evaluation We evaluate our model without adding extra image retrieval scenarios to match the evaluation settings of our baselines which are described in the dataset paper. There are six metrics in our evaluation that first five ones are conventional metrics and have to do with the accuracy of predicted actions and the quality of the generated text. Image Match rate, however, is vaguely introduced by MMConv paper which refers to the correctly predicted label of the image where the venue name is correctly predicted before it. Our model outperforms other baselines in image match rate and other metrics are not computed due to resource constraints.

Method/Metric	Join Accuracy	Inform Rate	Success Rate	Blue Score	Combined Score	Image Match
DS-DST	0.18	-	-	-	-	-
MARCO	-	88.7	82.4	17.09	102.64	0.17
SimpleTOD	0.28	14.6	9.2	20.30	32.30	0.02
MMTOD (ours)	-	-	-	-	-	**0.66***

Fig 7: * with exact image we reach 100% accuracy.

Result analysis Part of the reason that our image retrieval fails to retrieve 34 percent of time MMConv image dataset is labeled inadequately as demonstrated in Fig 8. This labeling is not only suitable for our image retrieval method but also is an ineffective way of evaluating base-line methods. There are some cases that our image retrieval fails due to inability to distinguish very similar foods, yet with different names. This issue can be solved by fine-tuning the image retrieval method.

Predicted label (yhat)	Dataset label (ytest)
coffee	drink
nightlife	clarke quay
night cityview	cityview at night
seafood noodle	japanese noodle (ramen/udon/soba)
artwork	Drawing

Fig 8: sample of cases in which image retrieval faults

Further discussion and Future works A valuable further development is to unify both language modeling and image retrieval modules to reach the first end-to-end multi-modal TOD. The way that these two modules already talk to each other is like rule-based models while this interaction can be implemented by the encoder-decoder approach.

Image match is not an appropriate measurement that is introduced with a classification-based image retrieval mindset while these approaches are rendered obsolete with the current pre-trained methods. In addition, many images, used in dialogue file, does not have a label that hinders the evaluation process.

5) Conclusion

In this paper, we presented a multi-modal multi-domain dialogue system, concentrated on five travel domains that are in high demand from travel agencies on account of high recreational need from elderly people and increasing employment wage in the service sector of the economy [bloom et al, 2016].

References

[Wikipedia: Turing test]. Wikipedia: Turing test. last edited on 1 April 2022. https://en.wikipedia.org/wiki/Turing_test

[Hussain et al, 2019] Hussain S., Ameri Sianaki O., Ababneh N. (2019) A Survey on Conversational Agents/Chatbots Classification and Design Techniques. In: Barolli L., Takizawa M., Xhafa F., Enokido T. (eds) Web, Artificial Intelligence and Network Applications. WAINA 2019. Advances in Intelligent Systems and Computing, vol 927. Springer, Cham. https://doi.org/10.1007/978-3-030-15035-8_93

[Fu et al, 2022]. Tingchen Fu and Shen Gao and Xueliang Zhao and Ji-rong Wen and Rui Yan. (2022) Learning towards conversational AI: A survey. https://doi.org/10.1016/j.aiopen.2022.02.001

[kim et al, 2019] Collaborative Drawing as a Testbed for Grounded Goal-driven Communication

[shen et al, 2021] ViDA-MAN: Visual Dialog with Digital Humans

[chen et al, 2021] Multi-Modal Chatbot in Intelligent Manufacturing

[liao et al, 2021] An Environment for Multimodal Conversational Search across Multiple Domains

[yang et al, 2021] UniMF: A Unified Framework to Incorporate Multimodal Knowledge Bases intoEnd-to-End Task-Oriented Dialogue Systems

[saha et al, 2018] Towards Building Large Scale Multimodal Domain-Aware Conversation Systems

[moon et al, 2020] Situated and Interactive Multimodal Conversations

[das et al, 2016] Visual dialog. In CVPR

[dodge et al, 2016] Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems.. In ICLR.

[cui et al, 2019] User Attention-guided Multimodal Dialog Systems [radford et al, 2018] Language Models are Unsupervised Multitask Learners

[hosseini-asl et al, 2021] A Simple Language Model for Task-Oriented Dialogue

[redford et al, 2021] Learning Transferable Visual Models From Natural Language Supervision

[Wikipedia: PCA] Principal component analysis. last edited on 6 February 2022. https://en.wikipedia.org/wiki/Principal_component_analysis

[bloom et al, 2016] Demography of Global Aging