MMTOD

Mohammadreza Ardestani

April 2022

Data Science Lab
Amirkabir University of Technology

Disclaimer

Hey, Sup?

Please Grab A Cup Of Coffee Since Q&A Might Take Long

Agenda

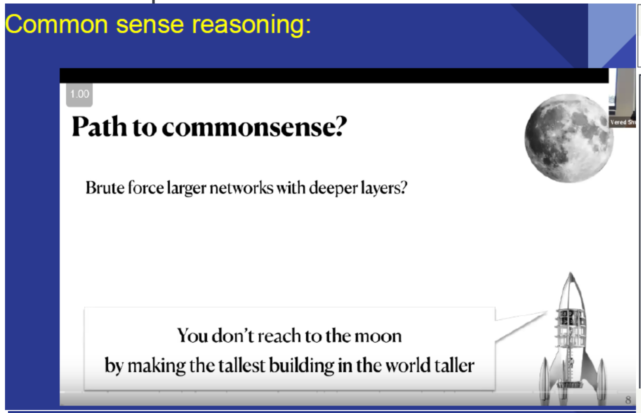
Retro

Sprint #0 | Sprint #1 | Sprint #2 | Sprint #3

slides

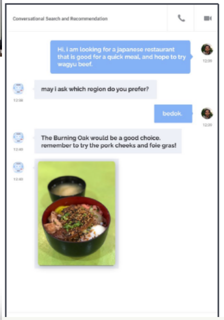
Sprint #0

Common sense reasoning:



Sprint #1

غنی سازی گفتگو



دو بخش اصلی برای غنی سازی گفتگو:

❑ تعیین زمان مناسب در بین گفتگو برای نمایش تصویر

❑ پیدا کردن تصویر مرتبط به متن گفتگو (برای مثال از پایگاه داده)

چالش‌ها – تصاویر

برخی از موارد استفاده تصاویر در مجموعه داده:

❑ نمایش غذا و محتویات آن

❑ پیدا کردن مکان‌ها و غذاهای مرتبط به تصویر کاربر

❑ توصیف فضا و اتمسفر مکان‌ها

در مجموعه‌داده‌هایی مانند MultiWOZ یا DSTC2 اطلاعات و حاشیه‌نویسی گفتگو وجود دارد اما فقط متنی و تک حالته هستند.

در مجموعه‌داده‌هایی مانند VisDial یا MMD تصاویر و ویژگی‌های آنها وجود دارد اما اطلاعات و حاشیه‌نویسی برای گفتگو وجود ندارد.

روش‌های موجود برای Response generation یا Dialogue state tracking برای مدیریت تصاویر در گفتگو با مشکل مواجه هستند.

Sprint #2

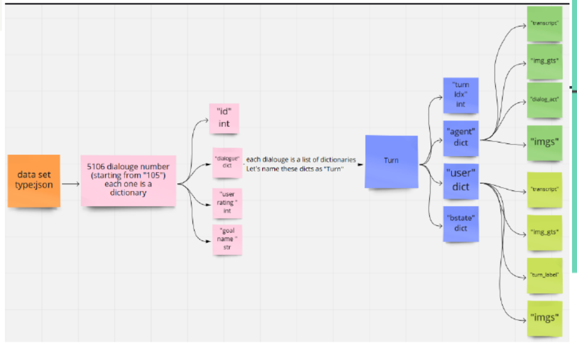
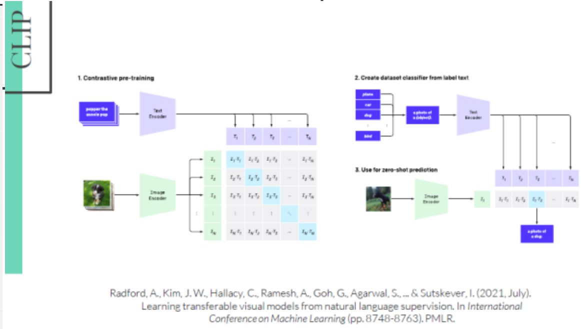


Figure: Json file structure of MMconv data set

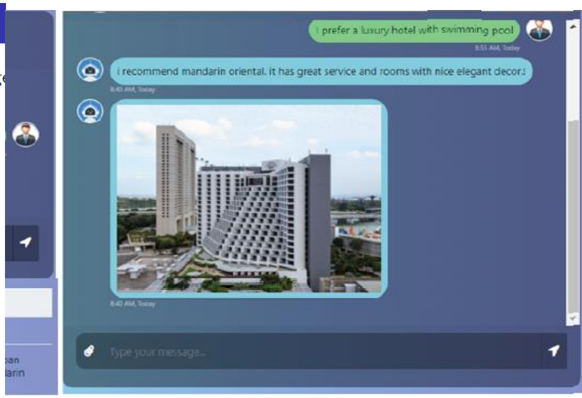
Evaluation of the image handler module

We are currently searching for a suitable evaluation meter for our image handler module.

Sprint #3



Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning* (pp. 8748-8763). PMLR.



Feedbacks

Sprint #0: Fine, Better to switch to travel domain

Sprint #1: Why Bother?

Sprint #2: Not clear and elaborated examples and use-case

Sprint #3:

+ How evaluate?

+Outdoor birthday?

+Only Image? That's it?

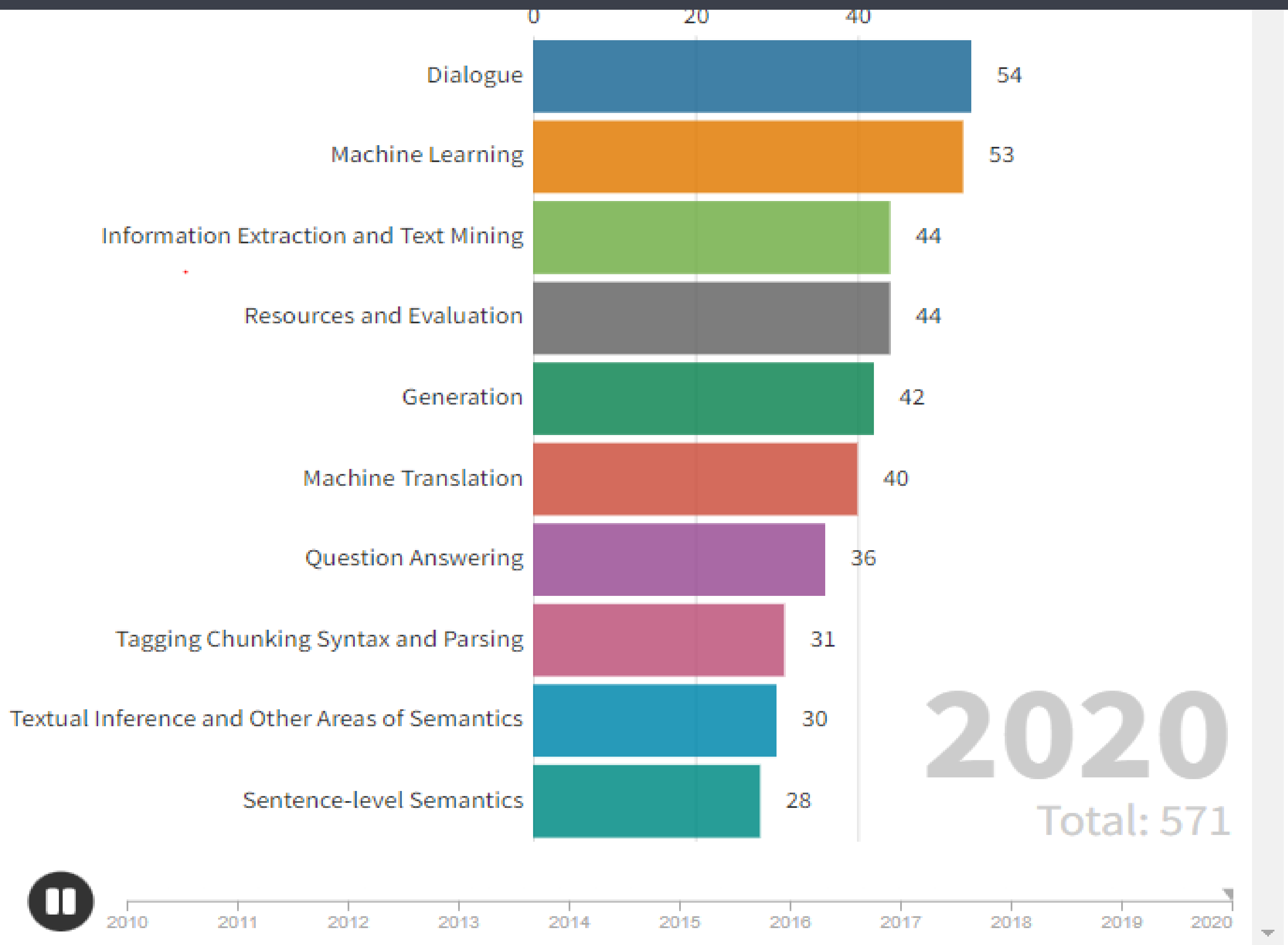
DataScience Lab

introduction

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].

ACM Research Tracks



Source: https://public.flourish.studio/visualisation/2431551/

ACM Research Tracks

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].

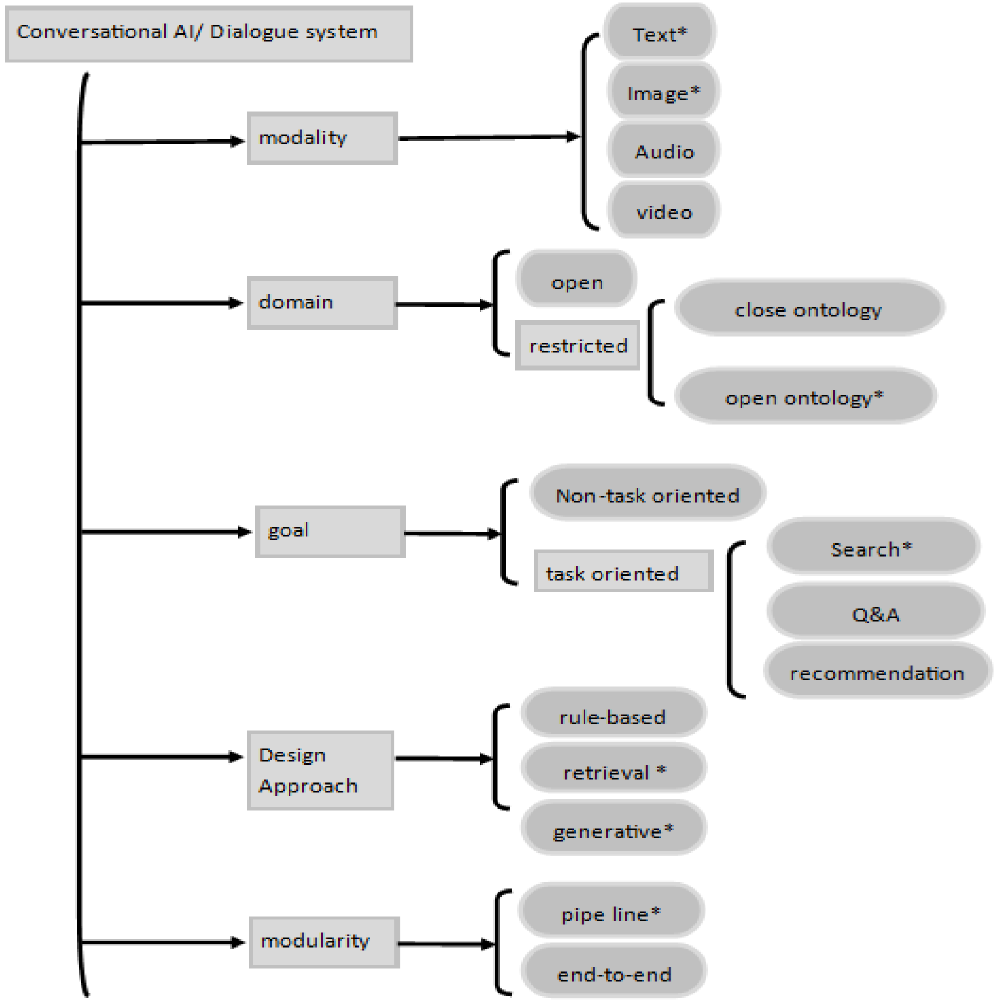


Fig1 Broad classification of dialogue systems: Our system type is marked with *. There are hybrid methods and rarely more modalities are used. Usually "Chabot" refers to open domain systems aiming to engage and inform users while "TOD system" refers to restricted-domain systems aimed at doing a specific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

ACM Research Tracks

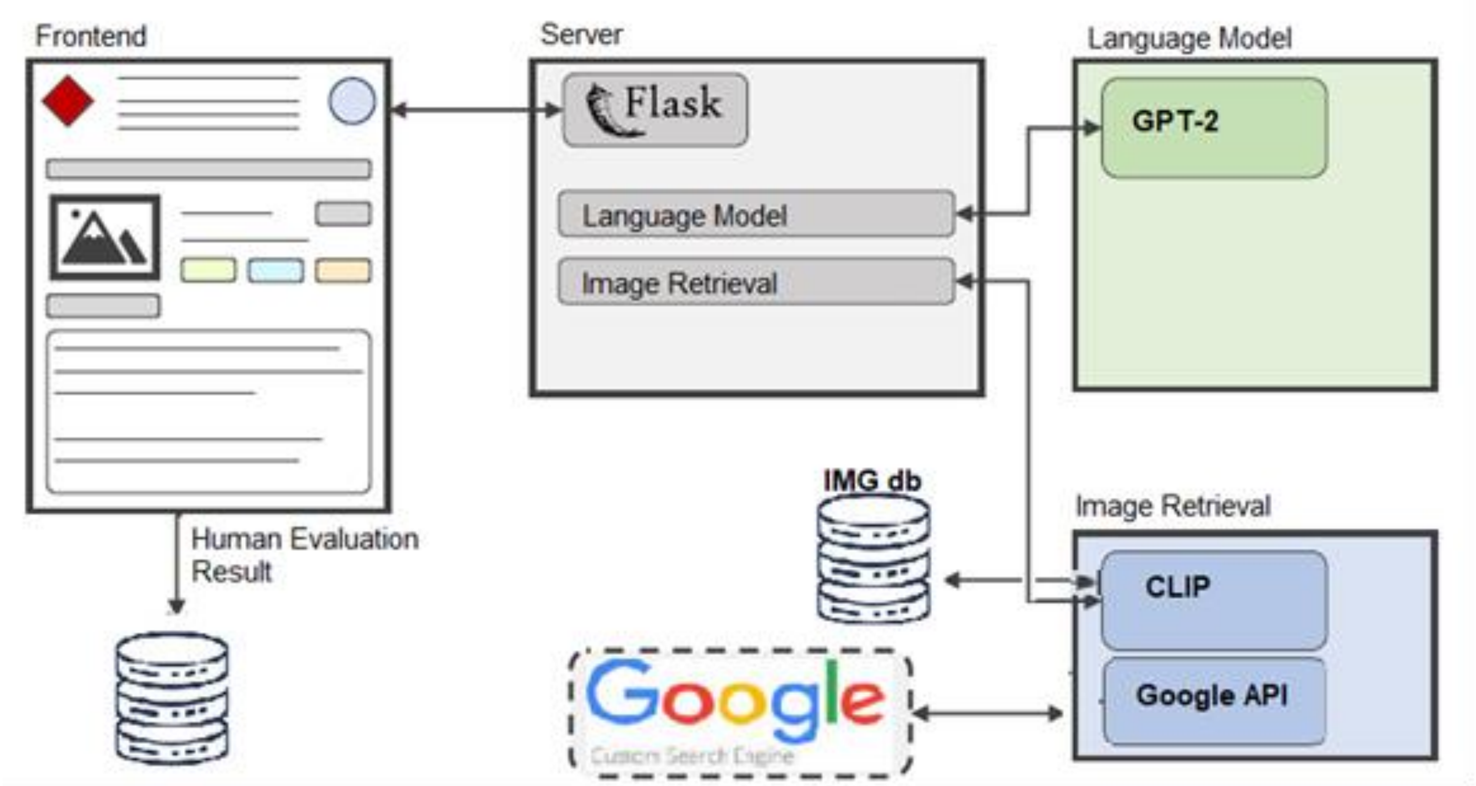


Fig2: Design of the proposed system

The highlights of our work

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.

Related works

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.

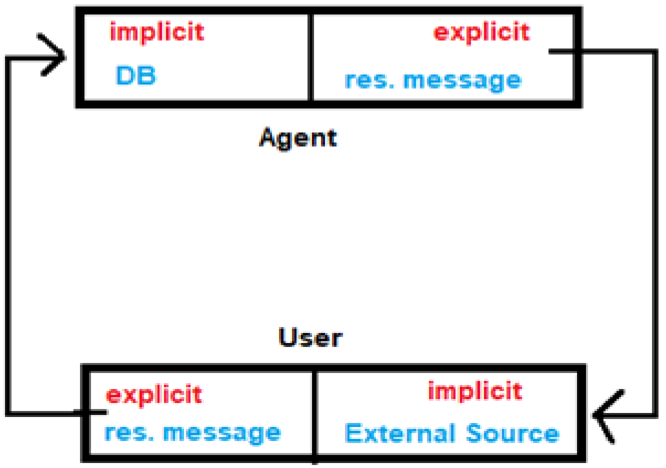


Fig3: Modality formulation. Users' implicit modalities do not need to enter into our equations since it does not affect the system architecture and it is up to users to handle it. CoDraw users [kim et al,2019], for example, describe a picture to the agent verbally; hence, we only need to consider handling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

Related works

			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.

Related works

[cui et al, 2019]

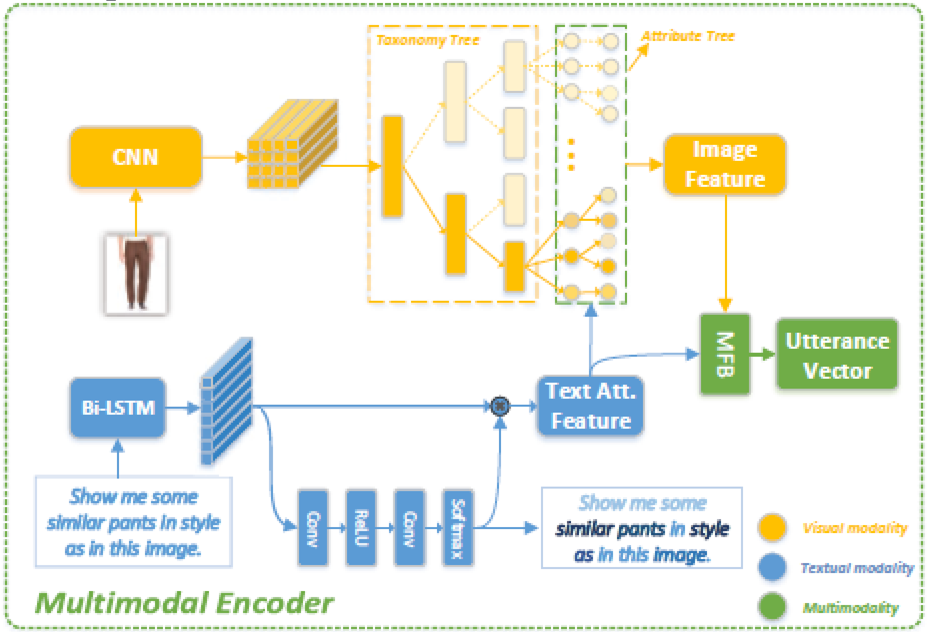


Figure 3: Schematic illustration of the multimodal encoder. A taxonomy-attribute combined tree is applied to learn the visual representation. The attention-augmented RNN encoder is incorporated to output attentive textual features and then the visual features are weighted by textual ones in the attribute level. They are ultimately fed into a multimodal fusion layer (MFB module) to generate the utterance vector.

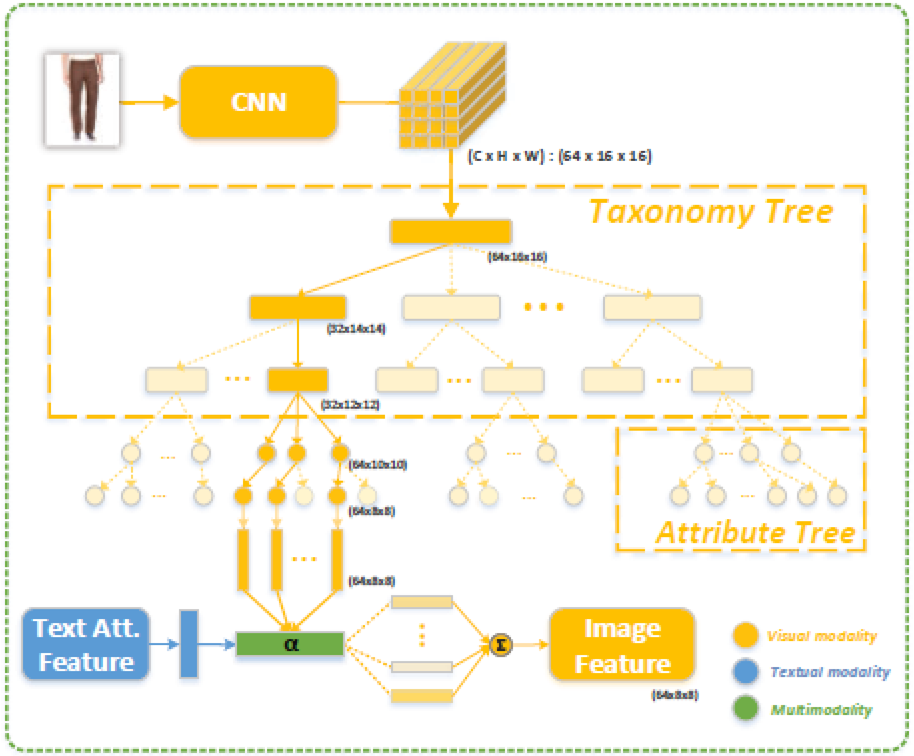
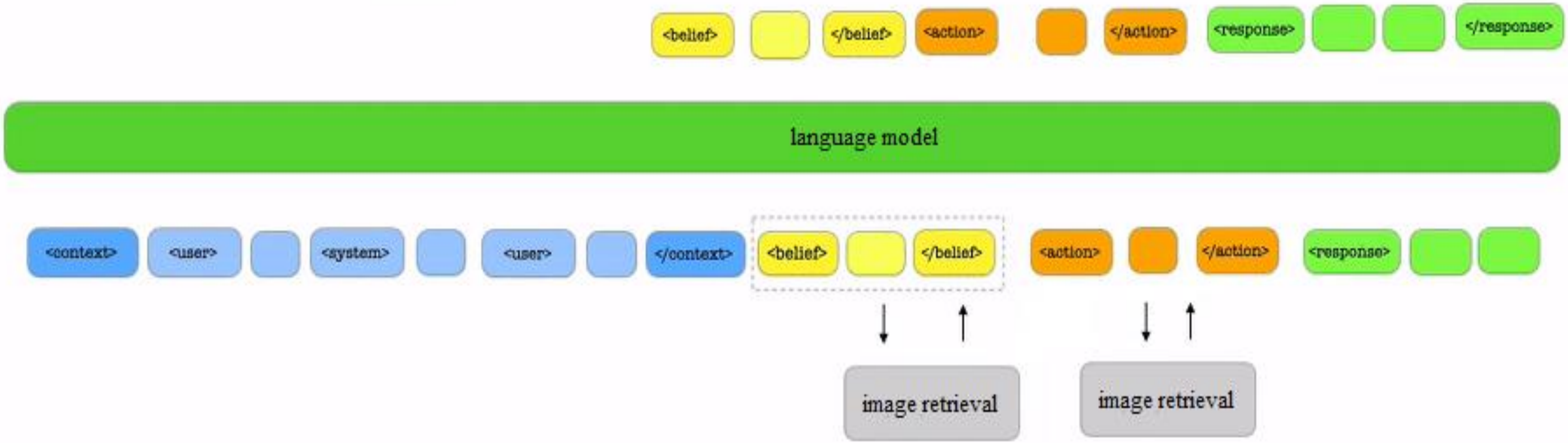


Figure 4: The proposed taxonomy-attribute combined tree. The solid lines connect the nodes that the image will pass through from top to bottom; whereas the dash lines denotes the irrelevant categories. Notably, all products share N common attribute nodes in the attribute tree.

3 Proposed Method

Proposed Method



Proposed Method

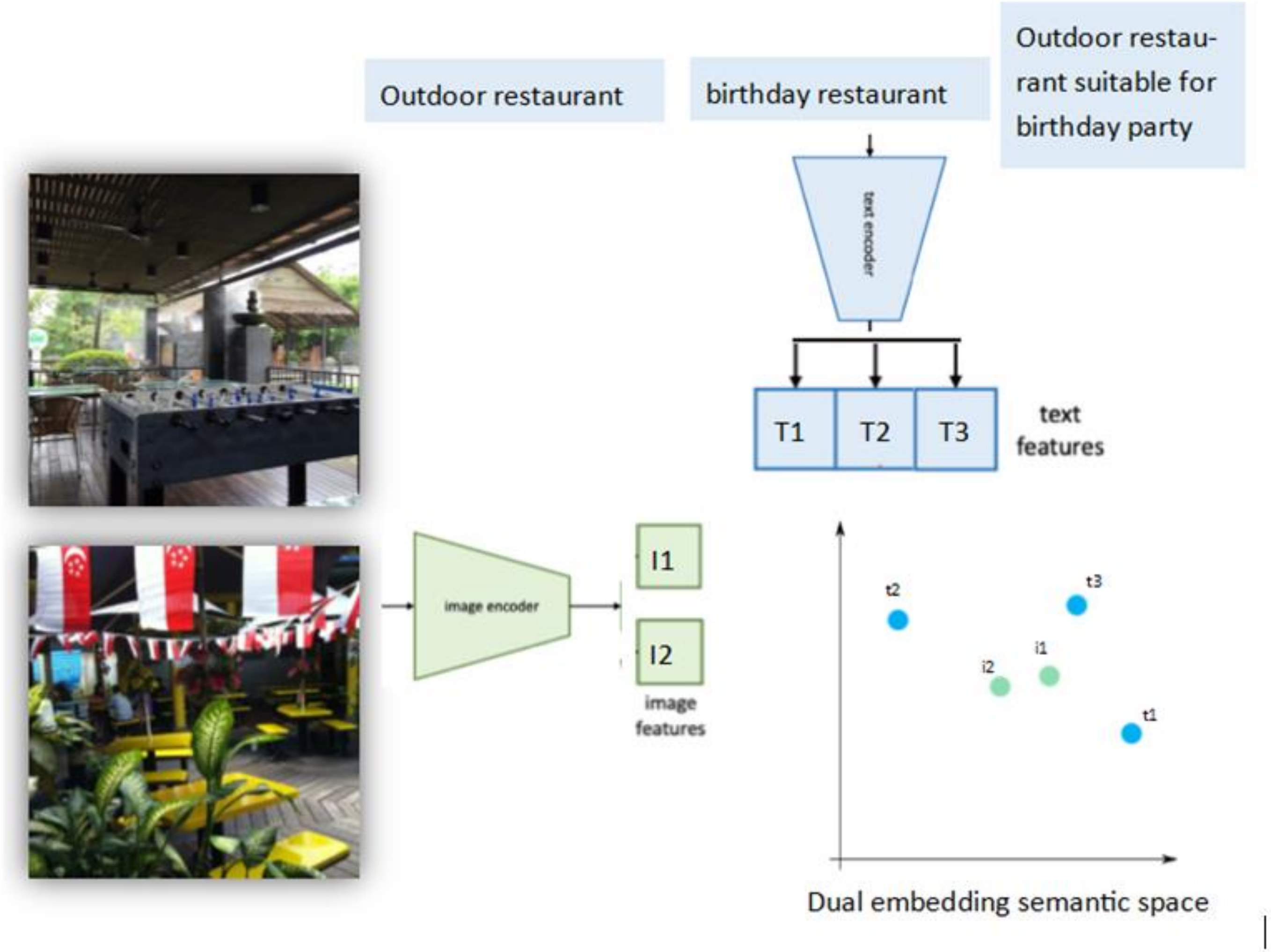


Fig 6 CLIP text and image encoders. (pictures from MMConv image dataset and the 2 D dual embedding semantic space is generated with initial 512 D feature vectors by PCA method [wikipedia: PCA] in sklearn library[1])

Evaluation

Evaluation

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)

Evaluation

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

Conclusion

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 due to the scarcity of young people working in the service sector of the economy which is caused by demographic aging [bloom et al, 2016].

REFERENCES

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

[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

References

[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

That’s that

Data Science lab
Amirkabir University of Technology

dslab.aut.ac.ir
+98 (21) 64545875
info@dslab.aut.ac.ir