Multi-domain Multi-modal Task-Oriented Dialogue system
A survey over different related data sets

M. Ardestani

University of AmirKabir

The 2^{nd} sprint of the project

December 12, 2021

Agenda

1. Project overview
   - Terminology
   - Problem Definition
   - Project Pipe-line
   - Why MM? (multi-domain multi-modal)
   - Why TOD? (Task-Oriented dialogue)
2. Data Set
   - MMconv
   - MultiWOZ
3. Model
4. Evaluation
5. Demo

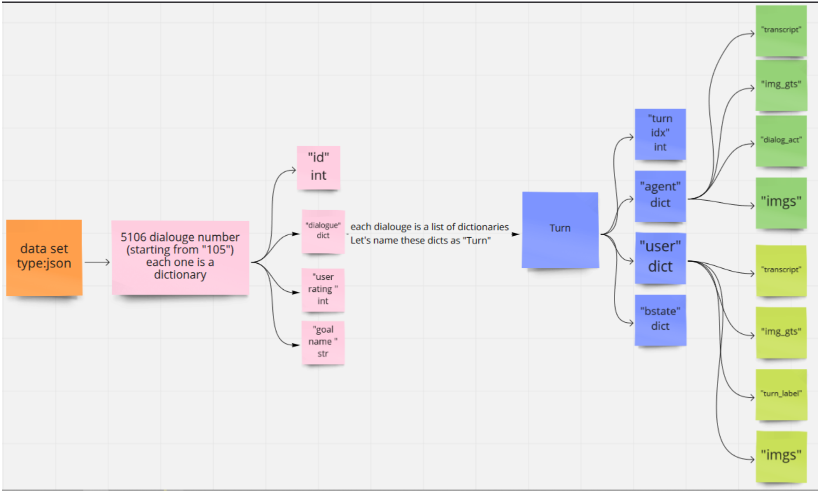


Figure: Json file structure of MMconv data set

Annotation means all the information that are extracted from **transcripts** and are stored in auxiliary structures like "belief state", "dialogue action"

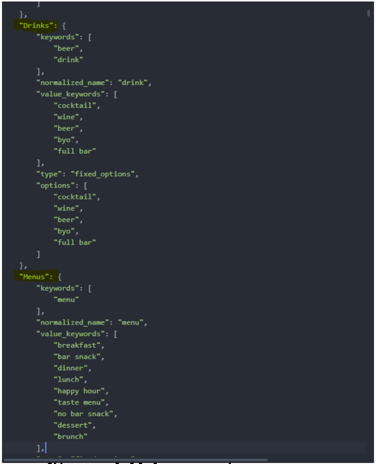


Figure: Dialogues and Ontology files - MMconv dataset

Terminology Cont.

Ontology means **concepts** that our system knows. **Slots** are **Keywords** in our Ontology.
we use **open span: X** when we don't have named entity X in as a **slot**.
Actions are predefined and we don't (need to) change them.

Table 4: Full ontology of all domains in our corpus. The upper script indicates which domain it belongs to. ⋆: universal, 1: food, 2: hotel, 3:nightlife, 4:mall, 5:sightseeing.

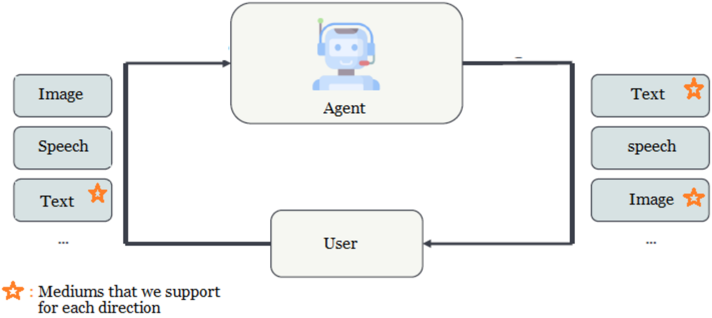
Action	inform / request/ recommend / negate / do not care/ confirm / show image/ greet / bye / others
Slots	drinks^{1,3} / music^{1,3} / reservations^{1,2,3,5} / dining options^{1,3} / stores^{4} / wifi* / menus^{1,2,3} / outdoor seating^{1,3} / venue domain* venue neighborhood* / wheelchair accessible ^{1,3}/ smoking ^{1}/ parking ^{1,3}/ restroom ^{1,2,3}/ credit cards * / pricerange ^{1,3}/ venuename* / venue score* / tips* / telephone* / venue address*

Figure: From MMconv paper

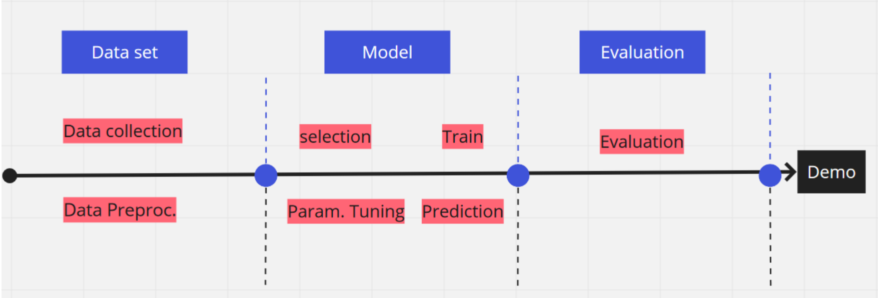
Problem Definition

In a few words, we are going to:

- Show an image when we make "recommendation".
- Show the most relevant image to our dialogue.



Project Pipe-line



multi-domain multi-modal?

Multi-modal?: Various real world applications

Multi-domain?: 73.8% times topics jumps from one domain to another

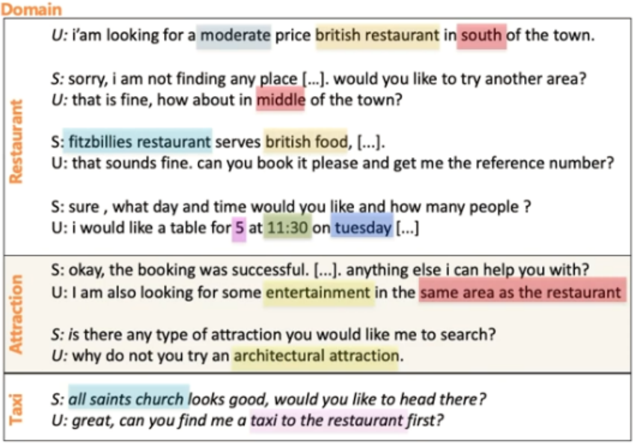
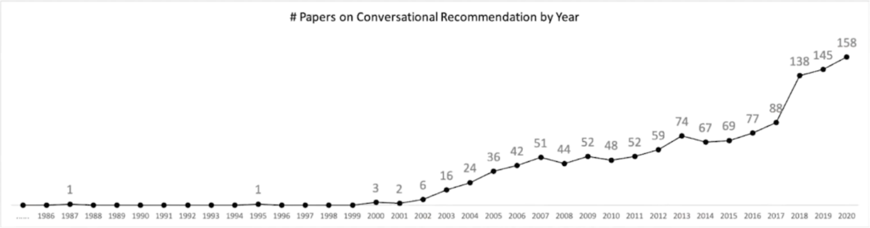


Figure: From Multi-domain Dialogue State Tracking with Recursive Inference paper

Why Task Oriented Dialogue system?



Papers in Google Scholar using query ("conversational recommendation" OR "conversational recommender").
May not represent all papers in this direction since many papers on the related topic may not include these exact words.

Figure: from ACM RecSys 2020 Conversational Recommender Systems

"Conversational-search based recommendation" and "Task oriented dialogue system" are relatively the same concepts.
For more information on how our paradigms on Conversational Search have been evolved, visit the mentioned paper.

Investigating data collection and properties of two related data set

Table 1: Comparison of our dataset MMConv to existing task-oriented dialogue datasets across domain, modality and tasks. 'Conv.' and ' Rec.' stand for 'conversational' and 'recommendation' respectively.

Datasets	# Dialogues	# Utters	Types	Domains	User Data	Modality	State Label
Facebook Rec [8]	1M	6M	Conv. Rec.	Movie	×	Text	×
REDIAL [17]	10K	163K	Conv. Rec.	Movie	×	Text	×
TG-ReDial [44]	10K	129K	Conv. Rec.	Movie	√	Text	×
OpenDialKG [23]	15K	143K	Conv. Rec.	Movie, book	×	Text	×
DuRecDial [21]	10K	156K	Conv. Rec.	Movie, music, news etc.	√	Text	×
MGConvRex [40]	7K	73K	Conv. Rec.	Restaurant	√	Text	√
WOZ 2.0[25]	1.2K	12K	Conv. Search	Restaurant	×	Text	√
DSTC2 [38]	1.6K	23K	Conv. Search	Restaurant	×	Text	√
FRAMES [9]	1.3K	20K	Conv. Search	Flight, hotel, budget	×	Text	√
KVRET [10]	3K	15K	Conv. Search	In-car assistant	×	Text	×
MultiWOZ [3]	8K	115K	Conv. Search	Hotel, restaurant etc.	×	Text	√
VisDial [5]	123K	2.4M	Image-based QAs	Concepts in image	×	Multi.	×
GuessWhat [6]	155K	1.6M	Image-based QAs	Concepts in image	×	Multi.	×
IGC [24]	4K	25K	Image-based QAs	Concepts in image	×	Multi.	×
MMD [29]	150K	6M	Fashion Search	Fashion	×	Multi.	×
MMConv	5.1K	39.7K	Conv. Search	5 domains in travel	√	Multi.	√

Figure: from MMconv paper

Data sets cont.

Three ways of collecting dialogue data:

1. machine synthesized
2. human-to-machine
3. human-to-human

Both MMconv and MultiWoZ are collected by the third way. With different setups, however.
While MMconv uses **open** ontology, MultiWoZ has a small **fixed** ontology.

MMConv dataset

They have recruited about 87 people and have trained them to generate dialogues based on a special setting and the dataset has been reevaluated 5 times

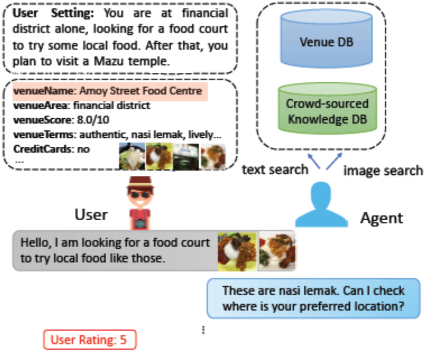


Figure 1: The multimodal conversation collection setting.

Table 5: The general statistics of the MMConv corpus.

Entry	Number
# dialogues	5,106
# turns	39,759
# single domain v.s. multi-domain	808 v.s. 4,298
# single modality v.s. multi-modality	751 v.s. 4,355
# goals	386
# total venues in DB	1,771
# total images	113,953
# total reviews	42,850
# average user ratings	4.67

The general statistics of the MMConv corpus are listed in Table

Figure: from MMconv paper

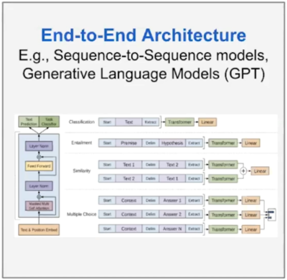
MultiWOZ

From MultiWoZ multiple versions have been released (2.0 up to 2.4)

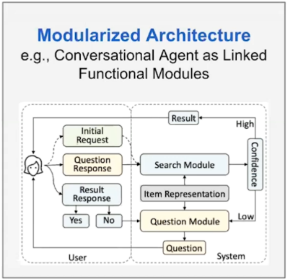
MultiWOZ 2.1 have had following issues(from SimpleTOD paper):

1. User provided multiple options, but context does not provide sufficient information to determine the true belief state.
2. Belief state is not labeled, but context provides sufficient information.
3. Belief state is labeled, but context lacks necessary information.
4. Belief state value is misspelled according to the context information.

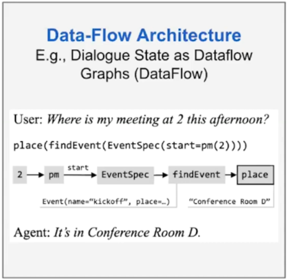
Model



Radford, Alec et. al. Improving Language Understanding by Generative Pre-Training. arXiv 2018.



Zhang, Yongfeng et. al. Towards Conversational Search and Recommendation: System Ask, User Respond. CIKM 2018.



Andreas, Jacob et. al. Task-Oriented Dialogue as Dataflow Synthesis. TACL 2020.

Figure: from ACM RecSys 2020 Conversational Recommender Systems

Evaluation of the chat-bot

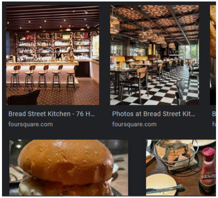
- Turn-level Metrics
  - Recommendation accuracy per turn (e.g., Precision, Recall, NDCG)
  - Frequencies and distributions of recommendation acts
  - Limitation: cannot measure the overall recommendation performance of the whole dialog
- Dialogue-level Metrics
  - Recommendation accuracy at round k (e.g., Precision@k, Recall@k, NDCG@k)
  - Dialogue success rate (e.g., SuccessRate@ k)
- Business-level Metrics
  - Conversion rate per dialog
  - Sales revenue
  - User satisfaction rating, user retention, customer loyalty

Figure: from ACM RecSys 2020 Conversational Recommender Systems

Evaluation of the image handler module

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


```
"1375": {
    "id": "1375",
    "dialogue": [
        {
            "turn idx": 0,
            "agent": {
                "transcript": "",
                "img_gts": [],
                "dialog_act": {},
                "imgs": []
            },
            "user": {
                "transcript": "hi i am looking for an authentic english restaurant that serves breakfast buffet.
                 it should be good for groups with scenic views. it should have banana pudding and caesar salad.",
                "img_gts": [],
                "turn_label": {
                    "open span: good for group": "inform",
                    "open span: authentic": "inform",
                    "open span: scenic view": "inform",
                    "menus: breakfast": "inform",
                    "placeholder": "greet"
                },
                "imgs": []
            },
            "bstate": {
                "open span: good for group": "inform",
                "open span: authentic": "inform",
                "open span: scenic view": "inform",
                "menus: breakfast": "inform"
            }
        },
        {
            "turn idx": 1,
            "agent": {
                "transcript": "you can go to bread street kitchen located in the financial district area.",
```

Demo

We are considering two following ways:

- Deploying on an online web-page
- Deploying as a Telegram chat-bot

Thanks
Do you have any questions?

Data Science lab
Amirkabir University of Technology

http://dslab.aut.ac.ir
https://github.com/MateAnderson/MMTOD