Explainability in Dialogue Systems

Maunendra Sankar Desarkar
Department of CSE, and
Department of AI
IIT Hyderabad

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad

Work with PhD student Suvodip Dey

Natural Language and Information Processing Lab
IIT Hyderabad

https://nlip-lab.github.io/

graph neural networks
naturallanguageprocessing
content-based filtering
estimating traffic flow
search-engines
social network analysis
information-retreival
information retreivals
word2vec
empathetic
learning
image captioning
privacy
text classification
transformers
code-mix
knowledge
conversational-ai
e-commerce
indexing
multilingual
nlp
hostility
research
nlu
toxicity
query-processing
air quality index
negative sentiment
llm
dialog-systems
gnn
language
data protection
topic-modelling
conversationalai
bert
code-mix nlp
digital libraries
nlg
intrepretation
ranking
hostility detection
sentiment analysis
interpretability
recommendation-system
robustness
virtual assistants
reponsible nlp
pos tagging
machine translation
controllable nlg
large-language-models
recommendation system
representation learning
knowledge augmentation
low-resource language

Multilinguality

Dialog Systems

Hate Analysis

AutoSuggest

Non-Toxic Gen

Low-Resource NLP

Publications in: ECML-PKDD 23, ACL Findings-21,22,23; SIGKDD-23; SIGDIAL-21,23; INLG-23; ACL-22; NAACL-22; WWW-22, WebSci-20,22; ECIR-21; CIKM-20; SIGSPATIAL-20; RecSys-18; TCSS-18,22; …

Dialogue Systems

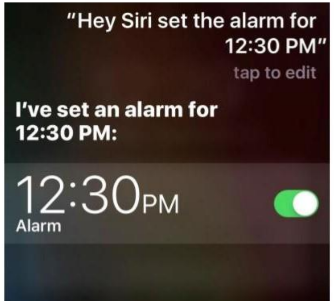
- What are dialog systems?
  - Conversations with automated agents
  - Takes *natural language input* and produces *natural language output*
  - System provides responses to user utterances
  - A sequence of (u_t, s_t) tuples
  - Also being commonly termed as conversational AI

- Useful in performing tasks, engaging with users
- Traditional systems were rule-based/flow-chart driven
- Recent advancements in NLP has made Dialog Systems smarter

Dialogue Systems

Goal-oriented

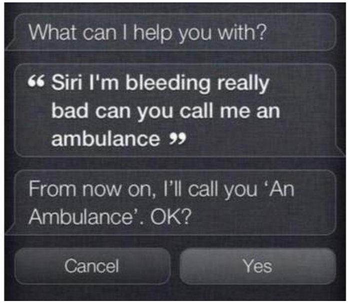
- Help users complete task through conversation
- Used for getting directions, searching restaurants/movies, booking tickets etc.
- Example: Siri, Alexa, Cortana, Google Now/Home
- Dataset: MultiWOZ, Schema-guided dialogue (SGD)



"Hey Siri set the alarm for 12:30 PM"

I've set an alarm for 12:30 PM:

12:30 PM



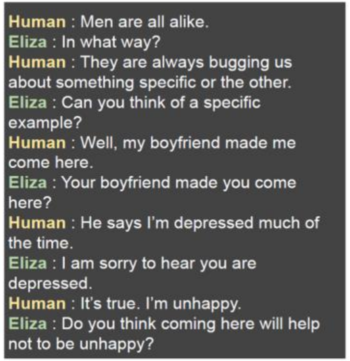
What can I help you with?

" Siri I'm bleeding really bad can you call me an ambulance "

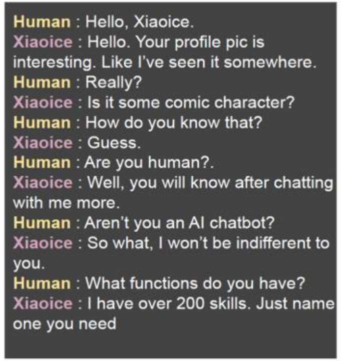
From now on, I'll call you 'An Ambulance'. OK?

Non-goal oriented

- Try to mimic unstructured human-human conversation/chats
- Can act as a psychotherapist, assistant, virtual-friend etc.
- Example: ELIZA, Microsoft XiaoIce, BenderBot, ChatGPT
- Dataset: DailyDialog, PersonaChat, Wizard-of-Wikipedia, TopicalChat



Human : Men are all alike.
Eliza : In what way?
Human : They are always bugging us about something specific or the other.
Eliza : Can you think of a specific example?
Human : Well, my boyfriend made me come here.
Eliza : Your boyfriend made you come here?
Human : He says I'm depressed much of the time.
Eliza : I am sorry to hear you are depressed.
Human : It's true. I'm unhappy.
Eliza : Do you think coming here will help not to be unhappy?

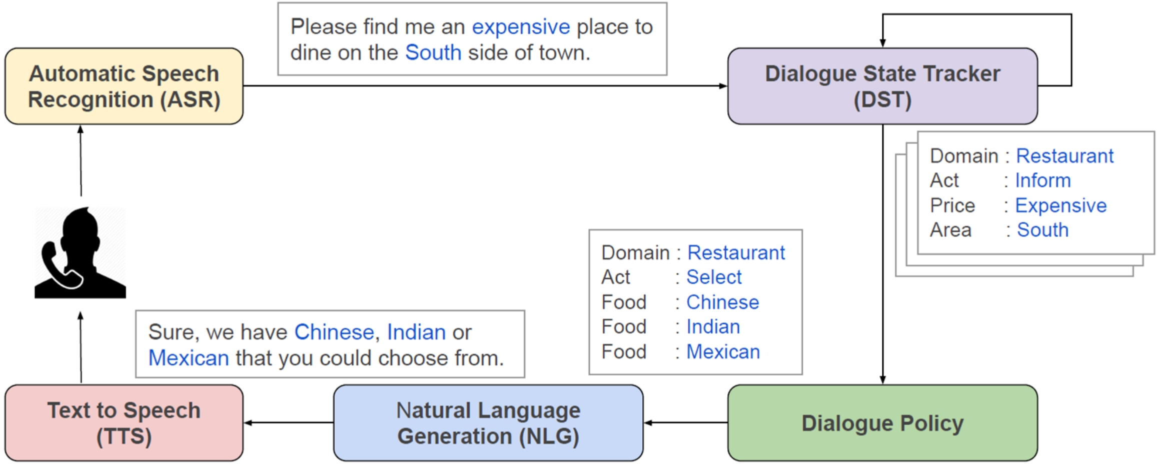


Human : Hello, Xiaoice.
Xiaoice : Hello. Your profile pic is interesting. Like I've seen it somewhere.
Human : Really?
Xiaoice : Is it some comic character?
Human : How do you know that?
Xiaoice : Guess.
Human : Are you human?.
Xiaoice : Well, you will know after chatting with me more.
Human : Aren't you an AI chatbot?
Xiaoice : So what, I won't be indifferent to you.
Human : What functions do you have?
Xiaoice : I have over 200 skills. Just name one you need

Goal-Oriented Dialogue Systems

Dialogue-State Architecture

- **Dialogue State Tracker (DST):** Keeps track of the overall dialogue act
- **Dialogue Policy:** Decides what dialogue act to generate
- **Natural Language Generation (NLG):** Generates response



Please find me an expensive place to dine on the South side of town.

Automatic Speech Recognition (ASR)

Dialogue State Tracker (DST)

Domain : Restaurant
Act : Inform
Price : Expensive
Area : South

Domain : Restaurant
Act : Select
Food : Chinese
Food : Indian
Food : Mexican

Sure, we have Chinese, Indian or Mexican that you could choose from.

Text to Speech (TTS)

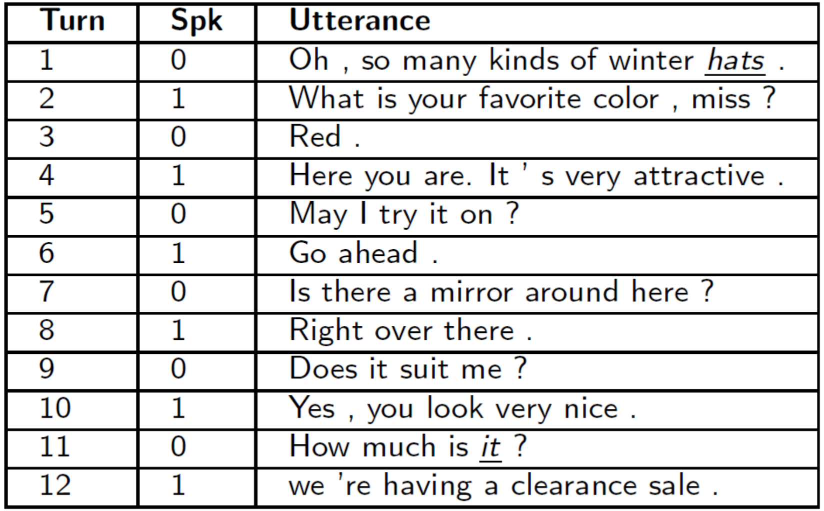
Natural Language Generation (NLG)

Dialogue Policy

Non-Goal Oriented or Open-domain Dialogue Systems

Pure Chat

- Generate response given dialogue history

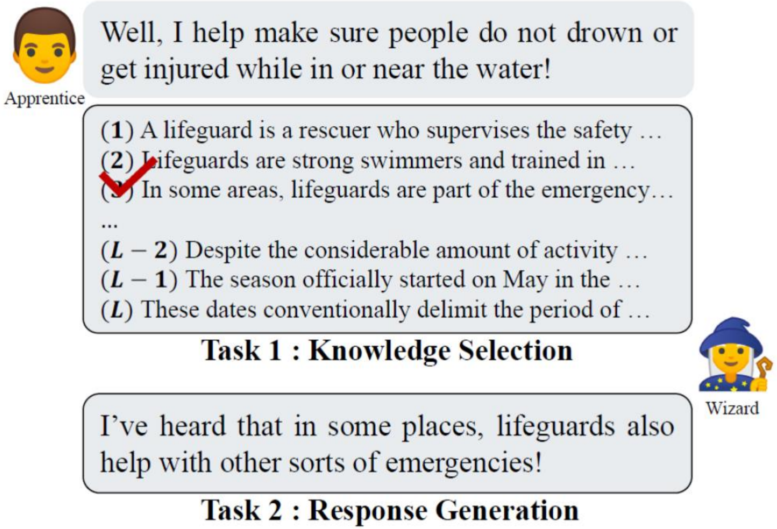


Turn	Spk	Utterance
1	0	Oh , so many kinds of winter *hats* .
2	1	What is your favorite color , miss ?
3	0	Red .
4	1	Here you are. It ' s very attractive .
5	0	May I try it on ?
6	1	Go ahead .
7	0	Is there a mirror around here ?
8	1	Right over there .
9	0	Does it suit me ?
10	1	Yes , you look very nice .
11	0	How much is *it* ?
12	1	we 're having a clearance sale .

Sample from DailyDialog dataset

Knowledge-grounded

- Generate response given dialogue history and other relevant conditions



Well, I help make sure people do not drown or get injured while in or near the water!

(**1**) A lifeguard is a rescuer who supervises the safety …
(**2**) Lifeguards are strong swimmers and trained in …
(**3**) In some areas, lifeguards are part of the emergency…
…
(***L*** − **2**) Despite the considerable amount of activity …
(***L*** − **1**) The season officially started on May in the …
(***L***) These dates conventionally delimit the period of …

Task 1 : Knowledge Selection

I've heard that in some places, lifeguards also help with other sorts of emergencies!

Task 2 : Response Generation

Sample from Wizard-of- Wikipedia dataset

Explainability: The Need

- Trust
- Causality
- Informativeness
- Fair and Ethical Decision Making
- Accountability
- Making Adjustments

Reference: Nadia Burkart and Marco F. Huber. A Survey on the Explainability of Supervised Machine Learning. Journal of Artificial Intelligence Research (JAIR), 2020.

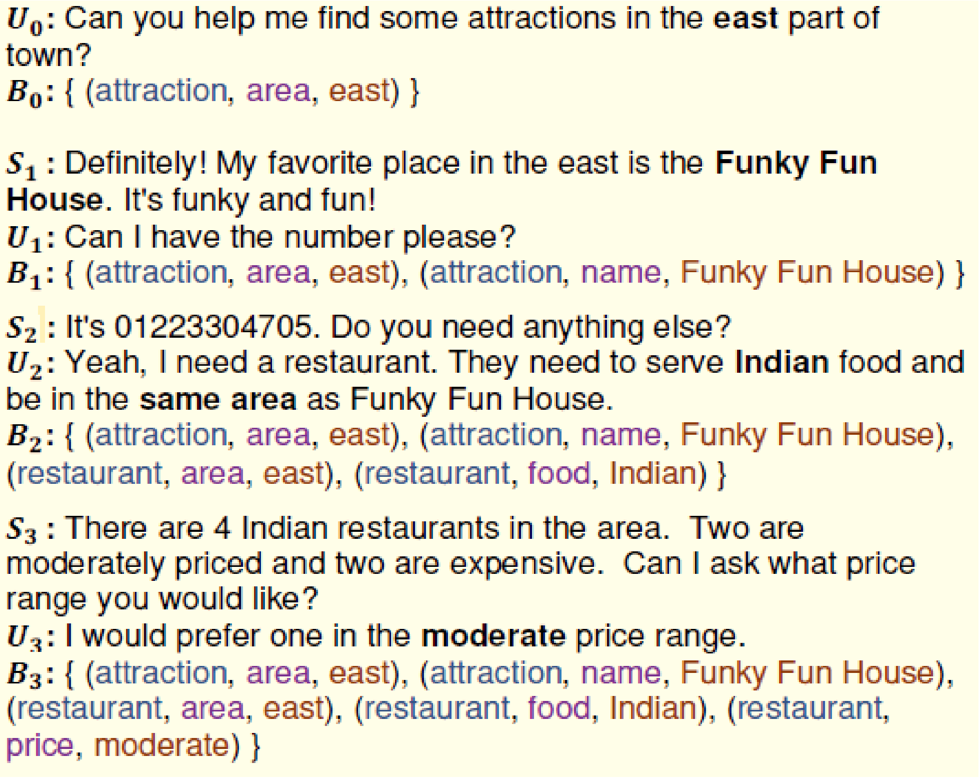
Our efforts towards Explainable Dialogue Systems

- Hi-DST: A Hierarchical Approach for Scalable and Extensible Dialogue State Tracking [1]
- DialoGen: Generalized Long-Range Context Representation for Dialogue Systems [2]
- Dial-M: A Masking-based Framework for Dialogue Evaluation [3]

1. Suvodip Dey, Maunendra Sankar Desarkar: Hi-DST: A Hierarchical Approach for Scalable and Extensible Dialogue State Tracking. SIGDIAL 2021.
2. Suvodip Dey, Maunendra Sankar Desarkar, and P. K. Srijith. "Towards Generalized and Explainable Long-Range Context Representation for Dialogue Systems." *arXiv preprint arXiv:2210.06282* (2022).
3. Suvodip Dey, Maunendra Sankar Desarkar: Dial-M: A Masking-based Framework for Dialogue Evaluation. SIGDIAL 2023.

Scalable and Extensible Dialogue State Tracking

- Dialog State/Belief State:
  - As a goal-oriented conversation flows, different slots/aspects and values get mentioned
  - Dialog state is a collection of those slot-value pairs
- How does it help?
  - The system frames DB queries from this Dialog States, and DB responses are used to generate the response
  - One way to validate response generation



U_0: Can you help me find some attractions in the **east** part of town?
B_0: { (attraction, area, east) }

S_1 : Definitely! My favorite place in the east is the **Funky Fun House**. It's funky and fun!
U_1: Can I have the number please?
B_1: { (attraction, area, east), (attraction, name, Funky Fun House) }

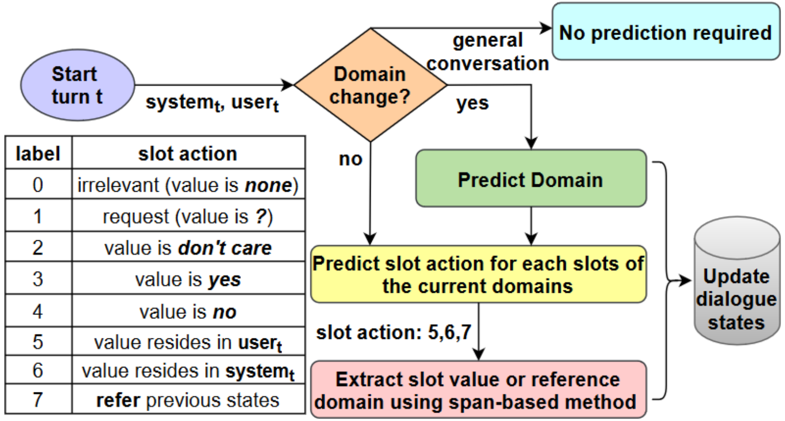
S_2 : It's 01223304705. Do you need anything else?
U_2: Yeah, I need a restaurant. They need to serve **Indian** food and be in the **same area** as Funky Fun House.
B_2: { (attraction, area, east), (attraction, name, Funky Fun House), (restaurant, area, east), (restaurant, food, Indian) }

S_3 : There are 4 Indian restaurants in the area. Two are moderately priced and two are expensive. Can I ask what price range you would like?
U_3: I would prefer one in the **moderate** price range.
B_3: { (attraction, area, east), (attraction, name, Funky Fun House), (restaurant, area, east), (restaurant, food, Indian), (restaurant, price, moderate) }

Suvodip Dey, Maunendra Sankar Desarkar: **Hi-DST: A Hierarchical Approach for Scalable and Extensible Dialogue State Tracking**. SIGDIAL 2021.

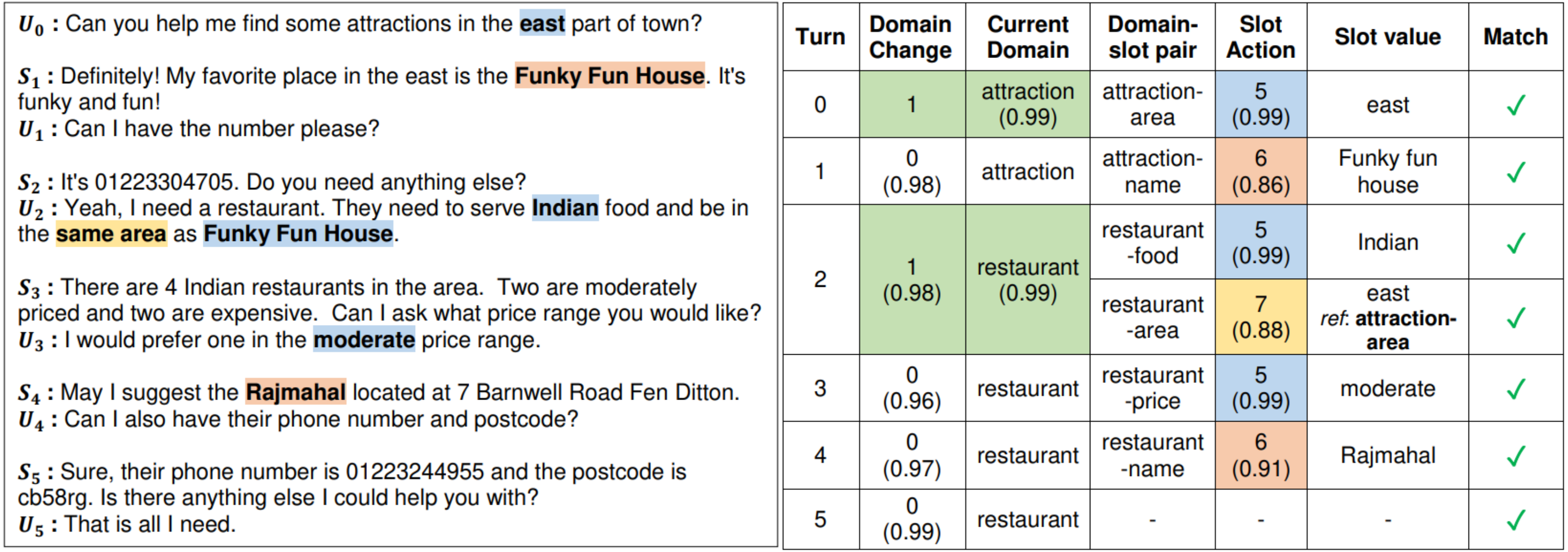
Scalable and Extensible Dialogue State Tracking

- Expect more slot values in the future
  - Flat multi-class classification becomes unreliable
  - Break it down into smaller pieces
- A series of predictions:
  - Domain change / Domain / Slot-action / Slot-value
- Need to be extensible
  - More slots, values can be added in future
  - Can we avoid (detailed) retraining?
  - Use representation of the slot and value names wherever possible
  - Convert domain-slot values into questions
    - hotel-name: What is the name of the hotel?
  - Context: User/system utterance based on slot action


Start turn t
system_t, user_t
Domain change?
general conversation
No prediction required
yes
no
Predict Domain
Predict slot action for each slots of the current domains
slot action: 5,6,7
Extract slot value or reference domain using span-based method
Update dialogue states

label	slot action
0	irrelevant (value is ***none***)
1	request (value is ***?***)
2	value is ***don't care***
3	value is ***yes***
4	value is ***no***
5	value resides in **user_t**
6	value resides in **system_t**
7	**refer** previous states

Scalable and Extensible Dialogue State Tracking



U_0 : Can you help me find some attractions in the **east** part of town?

S_1 : Definitely! My favorite place in the east is the **Funky Fun House**. It's funky and fun!
U_1 : Can I have the number please?

S_2 : It's 01223304705. Do you need anything else?
U_2 : Yeah, I need a restaurant. They need to serve **Indian** food and be in the **same area** as **Funky Fun House**.

S_3 : There are 4 Indian restaurants in the area. Two are moderately priced and two are expensive. Can I ask what price range you would like?
U_3 : I would prefer one in the **moderate** price range.

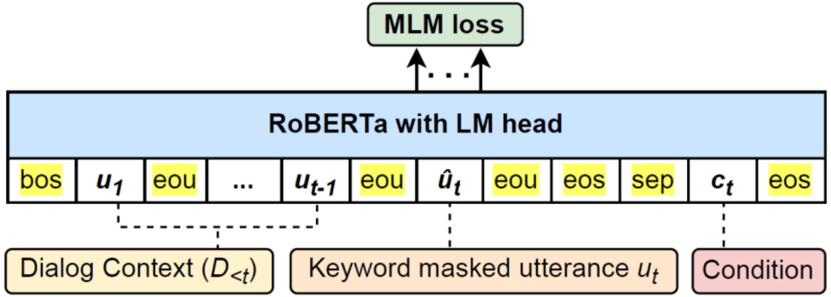
S_4 : May I suggest the **Rajmahal** located at 7 Barnwell Road Fen Ditton.
U_4 : Can I also have their phone number and postcode?

S_5 : Sure, their phone number is 01223244955 and the postcode is cb58rg. Is there anything else I could help you with?
U_5 : That is all I need.

Turn	Domain Change	Current Domain	Domain-slot pair	Slot Action	Slot value	Match
0	1	attraction (0.99)	attraction-area	5 (0.99)	east	✓
1	0 (0.98)	attraction	attraction-name	6 (0.86)	Funky fun house	✓
2	1 (0.98)	restaurant (0.99)	restaurant -food	5 (0.99)	Indian	✓
			restaurant -area	7 (0.88)	east *ref*: **attraction-area**	✓
3	0 (0.96)	restaurant	restaurant -price	5 (0.99)	moderate	✓
4	0 (0.97)	restaurant	restaurant -name	6 (0.91)	Rajmahal	✓
5	0 (0.99)	restaurant	-	-	-	✓

Dial-M

- Good quality response often contains cues related to the ongoing dialog context
- If not, then the response may become incoherent/random
- Proposed approach to check response quality:
  - ○ Mask important words from the response
  - ○ From the context, we should be able to predict those words
- Take a response
- Mask *important* keywords
- Predict the masked keywords
  - ○ Use standard high-quality denoising models
- No re-learning for separate datasets

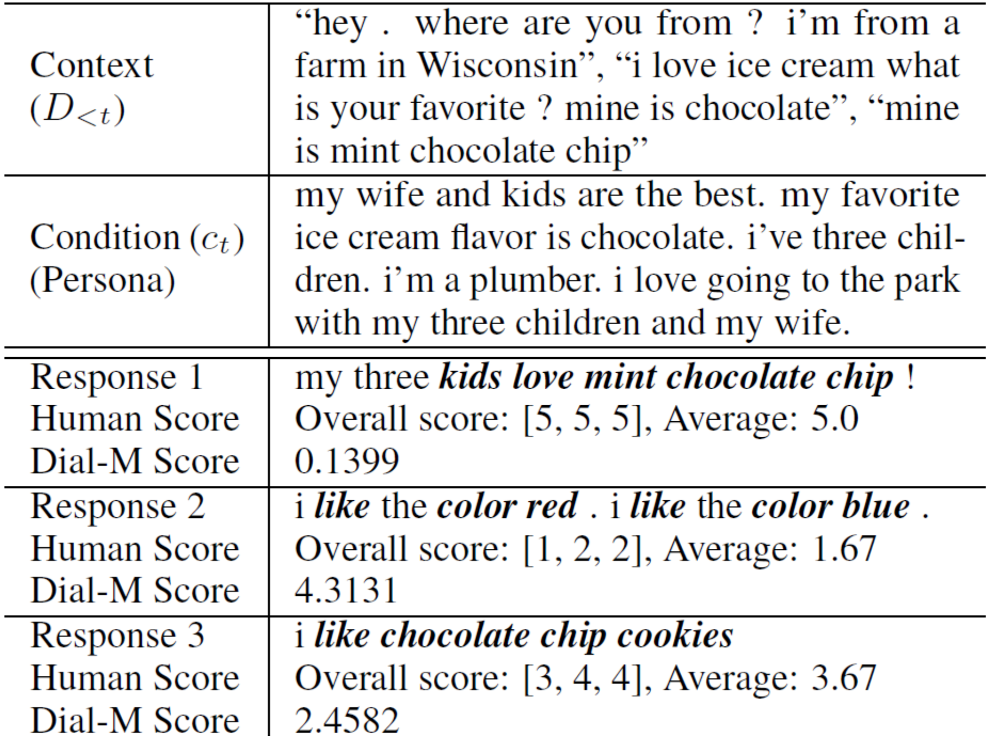


$$s = \frac{1}{|K_t|} \sum_{j=1}^{|K_t|} \left(\frac{1}{|T_{t,j}|} \sum_{y \in T_{t,j}} -\log p(y|D_{<t}, \hat{u}_{t,j}, c_t) \right)$$

Lower scores are better

Suvodip Dey, Maunendra Sankar Desarkar: **Dial-M: A Masking-based Framework for Dialogue Evaluation**. SIGDIAL 2023.

Dial-M



Context ($D_{<t}$)	“hey . where are you from ? i’m from a farm in Wisconsin”, “i love ice cream what is your favorite ? mine is chocolate”, “mine is mint chocolate chip”
Condition (c_t) (Persona)	my wife and kids are the best. my favorite ice cream flavor is chocolate. i’ve three children. i’m a plumber. i love going to the park with my three children and my wife.
Response 1	my three ***kids love mint chocolate chip*** !
Human Score	Overall score: [5, 5, 5], Average: 5.0
Dial-M Score	0.1399
Response 2	i ***like*** the ***color red*** . i ***like*** the ***color blue*** .
Human Score	Overall score: [1, 2, 2], Average: 1.67
Dial-M Score	4.3131
Response 3	i ***like chocolate chip cookies***
Human Score	Overall score: [3, 4, 4], Average: 3.67
Dial-M Score	2.4582

An example. Lower Dial-M score indicates better response quality

- Lower Dial-M scores indicate better responses
- Responses with higher human scores get less score for the Dial-M metric
- Correlates with human scores

Conclusion

- Explainability builds trust
    - Helps in debugging, and has other advantages
- Quality without explainability, for human-facing tasks is not adequate
- Generic and explainable methods are needed
    - For policy making
    - Response generation
    - Evaluation
- Some initial approaches are proposed
- A lot more need to be done

Thank you.
Questions?

..!!.....XX……???

……….

Maunendra Sankar Desarkar

Email: maunendra@cse.iith.ac.in
Website: https://people.iith.ac.in/maunendra/
Lab: https://nlip-lab.github.io/

భారతీయ సాంకేతిక విజ్ఞాన సంస్థ హైదరాబాద్
भारतीय प्रौद्योगिकी संस्थान हैदराबाद
Indian Institute of Technology Hyderabad