

Stanford CS224v Course

Conversational Virtual Assistants with Deep Learning

Lecture 6

The Dialogue State Tracking (DST) Problem

Monica Lam

Buying a Ticket from GPT-3

User: I'd like to buy a plane ticket to New York.

Agent: Sure, I can help you with that. What date and time would you like to depart?

User: July 7th, 9am.

Agent: Alright, I can book a flight for you departing on July 7th at 9am.
What airport would you like to depart from?

User: SFO

Agent: Great, I can book a flight for you departing from San Francisco
on July 7th at 9am. Would you like to book this flight now?

User: What is the flight number?

Agent: The flight number is AA123.

- + Perfect flow without training
- Hallucinates: American AA123
DFW → HNL
- Needs to consult DB

Current Commercial State of the Art

Let's give a call to an airline! (10/7/2023)

Welcome back, Monica.

- Commercial chatbots are conservative, despite LLM success

Quiz: Why?

Commercial SOTA: Dialogue Trees

User Utterance
I'd like to transfer \$100.00



Intent Classification



Intent & Slots

Transfer Money	
From	
To	
Amount	<i>\$100.00</i>

- **Initiatives**

- User initiates the transaction
- Agent solicits slots

- **Natural Language Understanding (NLU):**

Intent classification

- Intent: a small fixed set of possible intents
- slot-value pairs: values to slots

A Restaurant Reservation Agent Dialogue Tree

A: Hello, how can I help you?

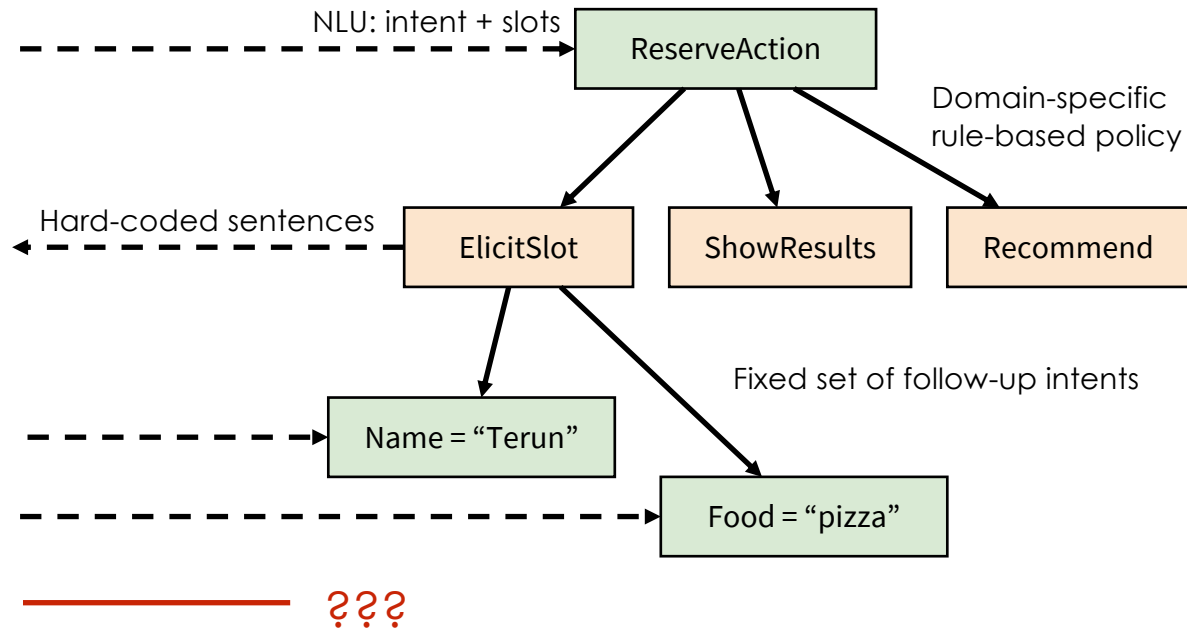
U: I'm looking to book a restaurant
for Valentine's Day

A: What kind of restaurant?

U: Terun on California Ave
-- or --

U: Something that has pizza
-- or --

U: I don't know, what do you
recommend?



Dialogue Trees: Pros & Cons

Pros: Control over conversation

Cons:

1. Expressiveness: Intent & slots are limited

What are the top 3 restaurants either near the airport or the Fisherman's Wharf

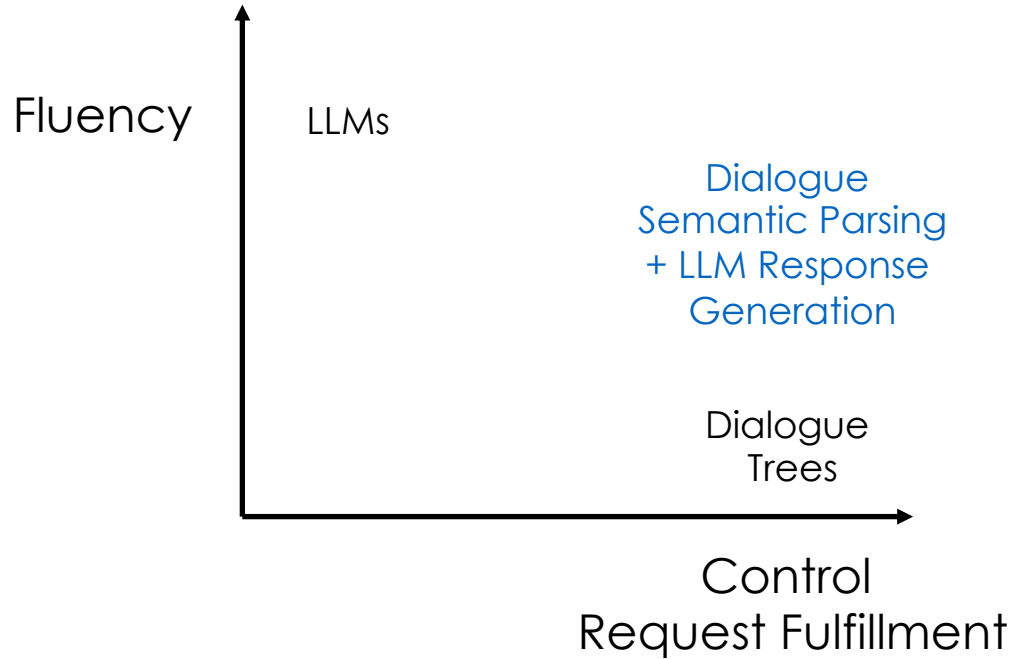
2. Tedious: exponentially many paths

- Product navigation where searches are modified in many possible ways
- Handcrafted for every problem

3. Fragile: Developer cannot anticipate all possible utterances

- Need to handle mixed initiatives for unexpected answers

How to Do Better?



Note: The dialogue history is needed to interpret user sentences

Knowledge Assistants (Example)

My dad is visiting me in Mountain View. What do you think about dinner choices?

Query:
*SELECT * FROM restaurants
WHERE location = "Mountain View"*

I found Cascal, a moderate Spanish restaurant with great tapas platters, appetizers, and drinks. It has a great ambience with live music on the weekends and both indoor and outdoor seating with heat lamps to keep guests warm.

How about Italian instead? I really like the food there when I visited Florence last year.

Query:
*SELECT * FROM restaurants
WHERE location = "Mountain View"
AND cuisines = ANY("Italian")*

I found Vaso Azzurro Ristorante, a moderate Italian restaurant near Mountain View. It has a lively atmosphere and offers excellent customer service. Reviews say their pasta is authentic and perfectly seasoned.

Oh, what is the ham called in Italian?

GPT question

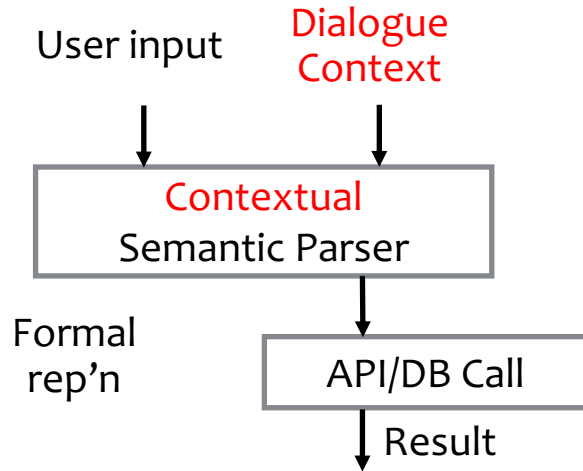
Prosciutto is the Italian word for ham.

Thanks. We'll be visiting Stanford. How about in Palo Alto?

Query:
*SELECT * FROM restaurants
WHERE location = "Palo Alto"
AND cuisines = ANY("Italian")*

I found Terún, a 4 star Italian and Pizza restaurant near Stanford in Palo Alto. Reviews say the pizza is great and they can seat customers outside without a reservation.

Single Sentence → Dialogues



- Semantic parser needs to be **contextual**
- Input: User utterance + the conversation so far

Transaction Assistants (Example)

USER	GENIE
I like a recommendation for French food.	
	I'd recommend Chez TJ in Mountain View. It has one Michelin star!
How far is it from here?	
	The distance to Chez TJ is 7.5 miles.
Can you recommend something nearby?	
	The distance to Left Bank is 2.8 miles. Would you like to book it?
Sure.	
	Which day would you like to book Left Bank?
November 2, 7pm, for 2 people.	
	Could you confirm that you want a table for 2 people at Left Bank on November 2 at 7pm?
Yes.	
	I booked Left Bank for 2 on November 2 at 7pm.

Query

Query

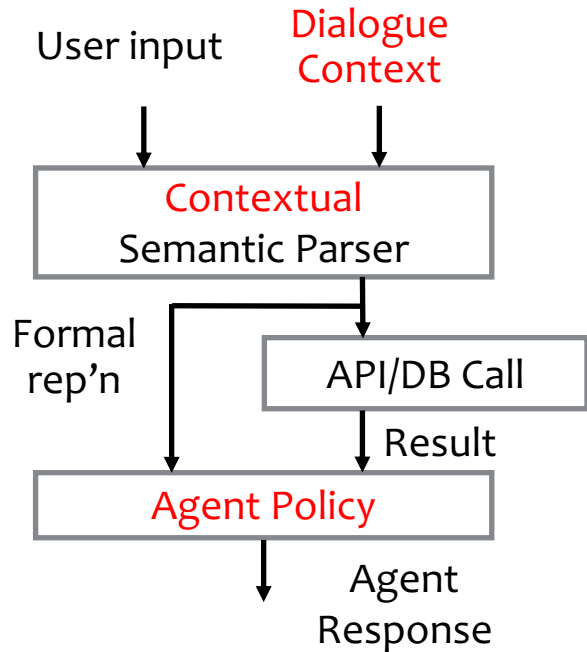
Offer to book

Slot fill

Ask for confirmation

Result

Knowledge vs. Task-Oriented Agent (ToD)



- A knowledge assistant is passive
 - Only reports the result from the API call
- The **agent policy** of task-oriented dialogue (ToD): [Next lecture]
 - Respond to users' questions
 - Perform actions
 - Making API / DB calls
 - Ask user questions
 - Offer unrequested information

How to represent the user's meaning?

Problem: Dialogue State Tracking

2 Variants

- **Tracking multiple speakers:**
 - Given a dialogue between two or more humans
 - Understand what is being said by each person turn by turn
 - Applications:
 - Create a meeting summary
 - Learn how to automate one of the attendees
e.g. a call agent, the lawyer doing the deposition
- **Tracking 1 speaker: (This lecture)**
 - The user talking to a virtual assistant
 - Understand what the user says in each turn of the dialogue

Lecture Goals

- Agent architectures
 - (1) LLMs, (2) dialogue trees, (3) contextual semantic parser
- The DST (Dialogue State Tracking) Problem
 - **Speech Act Theory**
 - Data acquisition methods
 - Wizard-of-Oz dialogues
 - Synthesized dialogues

The Meaning (Semantics) of Dialogues

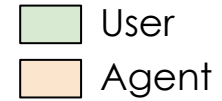
- **Speech Act Theory** [John L. Austin (1962)]
 - We can *classify* every utterance as one of a finite number of high-level purposes:
 - Requests, warnings, invitations, promises, apologies, predictions, ...
- In NLP, speech acts are also known as “dialogue acts”
 - We annotate each utterance by its dialogue act
 - Dialogue act state machine:
 - user dialogue act → agent dialogue act
 - agent dialogue act → user dialogue act

<https://plato.stanford.edu/entries/speech-acts/>

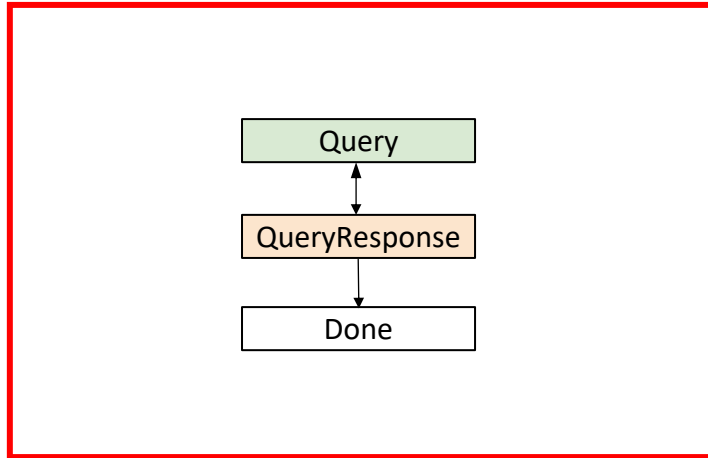
Dialogue Acts of *Basic* Transaction Agents

- User can initiate queries (Lecture 2-5)
- User can initiate actions:
 - Turn on the lights, play a song, make a tweet
 - Book a restaurant, transfer money, ...
- User can answer agent's questions to accomplish queries/actions

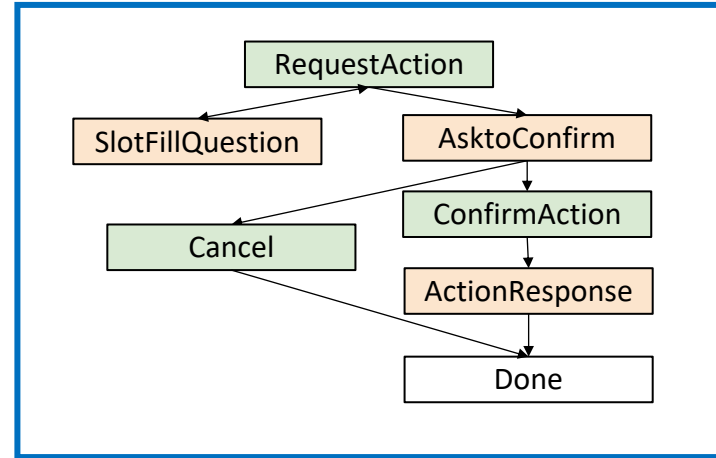
Example of a Dialogue State Machine



KB Navigation



Action



Dialogue act names: Independent of the conversation domain
Actual content is domain-specific

Example of a Restaurant Dialogue

USER
DIALOGUE ACTS

AGENT
DIALOGUE ACTS

	USER	GENIE	
query	I like a recommendation for French food.		
query	How far is it from here?	I'd recommend Chez TJ in Mountain View. It has one Michelin star!	query-response
query	Can you recommend something nearby?	The distance to Chez TJ is 7.5 miles.	query-response
request-action	Great. Could you please book it?	The distance to Left Bank is 2.8 miles.	query-response
request-action	November 2, 7pm, for 2 people.	Which day would you like to book Left Bank?	slot-fill
confirm-action	Yes.	Could you confirm that you want a table for 2 people at Left Bank on November 2 at 7pm?	ask-confirmation
		I booked Left Bank for 2 on November 2 at 7pm.	action-response

Discussion

- Consider other kinds of dialogues
 - What kind of dialogue acts do you expect?

Context of Dialogue State Tracking

context 1 {
turn 1 →

USER	GENIE
	Hello, I am Genie.
I'd like a French Restaurant.	I'd recommend Chez TJ in Mountain View. It has one Michelin star!
How far is it from here?	The distance to Chez TJ is 7.5 miles.
Can you recommend something near by?	The distance to Left Bank is 2.8 miles.
Perfect. Please book the restaurant	Which day would you like to book Left Bank?
Is it expensive?	The price is moderate.
OK	Which day would you like to book Left Bank?
November 2, 7pm, for 2 people.	Could you confirm that you want a table for 2 people at Left Bank on November 2 at 7pm?
Yes.	I booked Left Bank for 2 on November 2 at 7pm.
What is the reservation number?	10382345.

Quiz: What is the context of turn 2?

Context of Dialogue State Tracking

USER	GENIE
	Hello, I am Genie.
I'd like a French Restaurant.	I'd recommend Chez TJ in Mountain View. It has one Michelin star!
How far is it from here?	The distance to Chez TJ is 7.5 miles.
Can you recommend something near by?	The distance to Left Bank is 2.8 miles.
Perfect. Please book the restaurant	Which day would you like to book Left Bank?
Is it expensive?	The price is moderate.
OK	Which day would you like to book Left Bank?
November 2, 7pm, for 2 people.	Could you confirm that you want a table for 2 people at Left Bank on November 2 at 7pm?
Yes.	I booked Left Bank for 2 on November 2 at 7pm.
What is the reservation number?	10382345.

context 9

turn 9

Dialogue State Tracking (DST) Problem Definition

Dialogue State (DS): Meaning of the user utterance

Problem: Given a dialogue with alternating user-agent turns in NL, $u_1, a_1, u_2, a_2, \dots$

For each turn i

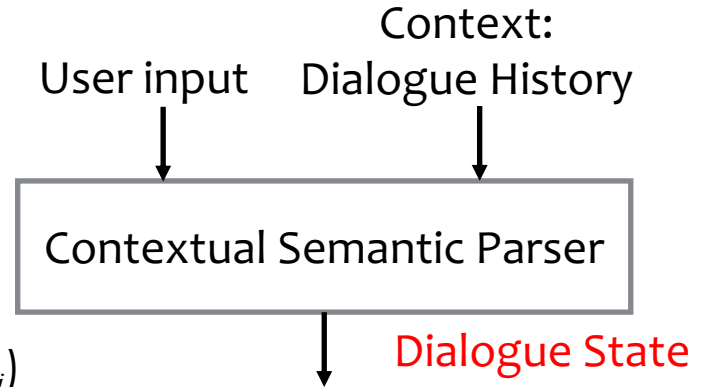
Predict the dialogue state: DS (u_1, a_1, \dots, u_i)

Training data:

$((u_1, a_1, \dots, a_{i-1}, u_i), \text{DS}(u_1, a_1, \dots, u_i))$

A dialogue with n turns $\rightarrow n$ training samples

Metric: *Joint Accuracy*: Accurate only if *all slots* are correct.



Quiz

HOW TO ACQUIRE DATA?

Lecture Goals

- Agent architectures:
 - (1) LLMs, (2) dialogue trees, (3) contextual semantic parser
- Dialogue State Tracking Problem
 - Speech Act Theory
 - Task-oriented agents
 - **Data acquisition methods**
 - Wizard-of-Oz dialogues
 - Synthesized dialogues

Where do we get training data?

1. Real-Life Recordings of Human Agents

- Human user, human agent
- From real conversations
(phone marketing, customer support, ...)
 - “This call may be monitored for quality and training purposes”
 - Confidentiality:
Outsource providers often cannot see a client's data
 - Not available to academia
- Cannot see the mistakes of an automated agent

2. Real-Life Recordings of an Agent

- Human user, computer agent
- Chicken and egg (aka *bootstrapping*):
we don't have a real agent until we train the model
- Same issues
 - Confidentiality
 - Outsource providers often cannot see a client's data
 - Not available to academia

3. Wizard-of-Oz (WOZ)

- Human user, human agent
- Paid workers (crowdworkers) on both sides
 - One worker pretends to be an automated agent
 - Another worker pretends to be a user, following a prompt
- The worker is given specific tasks
 - Does not reflect real life
 - changing their minds, exploring options
 - errors made by the agent

How to Annotate Human-to-Human Traces?

- Human agent uses GUI to access the database
- Record interaction with that UI (clicking, typing, ...) each turn
- Convert UI interaction trace to slots

- Simple, cheap (no extra effort on top of call center worker)
 - Are we done?

Case Study: MultiWOZ [Budzianowski et al 2018]

1140 paper citations

- **A large dataset of dialogues**
 - Annotated with domain + slots at every turn
 - Application: concierge service in Cambridge, UK
 - 7 domains → 5 domains
 - restaurants, hotels, attractions, train, taxi
 - police, hospital dropped due to bad quality
 - Number of slots: 30
 - **Closed ontology:** 1,829 values
 - **Representation:** slot-value pairs

MultiWOZ Ontology

Table 2: Full ontology for all domains in our data-set. The upper script indicates which domains it belongs to. *: universal, 1: restaurant, 2: hotel, 3: attraction, 4: taxi, 5: train, 6: hospital, 7: police.

act type	inform* / request* / select ¹²³ / recommend/ ¹²³ / not found ¹²³ request booking info ¹²³ / offer booking ¹²³⁵ / inform booked ¹²³⁵ / decline booking ¹²³⁵ welcome* / greet* / bye* / reqmore*
slots	address* / postcode* / phone* / name ¹²³⁴ / no of choices ¹²³⁵ / area ¹²³ / pricerange ¹²³ / type ¹²³ / internet ² / parking ² / stars ² / open hours ³ / departure ⁴⁵ destination ⁴⁵ / leave after ⁴⁵ / arrive by ⁴⁵ / no of people ¹²³⁵ / reference no. ¹²³⁵ / trainID ⁵ / ticket price ⁵ / travel time ⁵ / department ⁷ / day ¹²³⁵ / no of days ¹²³

Example of a MultiWOZ dialogue

U: I am looking for a restaurant. The restaurant should serve world food and should be in the centre.

A: Unfortunately, I could not find a restaurant that serves that kind of food in that area.

U: What expensive restaurants are there?

A: The Cambridge Chop House is expensive and it's British food. Is that okay?

U: No. Try looking for a restaurant that serves Korean food.

A: I found the Little Seoul restaurant. Would you like me to make a reservation?

U: Yes, a table for 2.

A: I would be happy to reserve for you, what day and time are you looking for?

U: Thank you, I need a reservation for 2 people at 13:15 on Monday.

A: gg3dcufd is your confirmation code, may I help you with anything else?

U: No that was everything. Thanks

Search

**Result; Action
suggestion call**

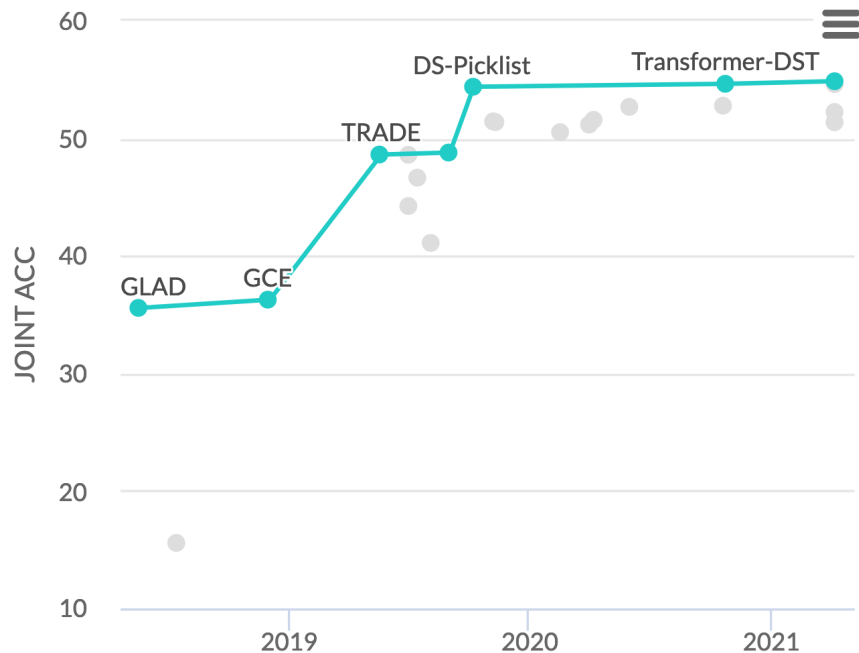
Action Slot Fill

API call

MultiWOZ Dataset

- **Users are asked to book**
 - One of various places (restaurant, hotel, train, taxi) and ask about attractions
 - Many dialogues cross multiple domains
- **Collected using Wizard-of-OZ conversations by crowdworkers**
 - ~56k turns train
 - ~7k turns test and ~7k turns dev
 - ~7 turns/dialogue on average

MultiWOZ 2.0



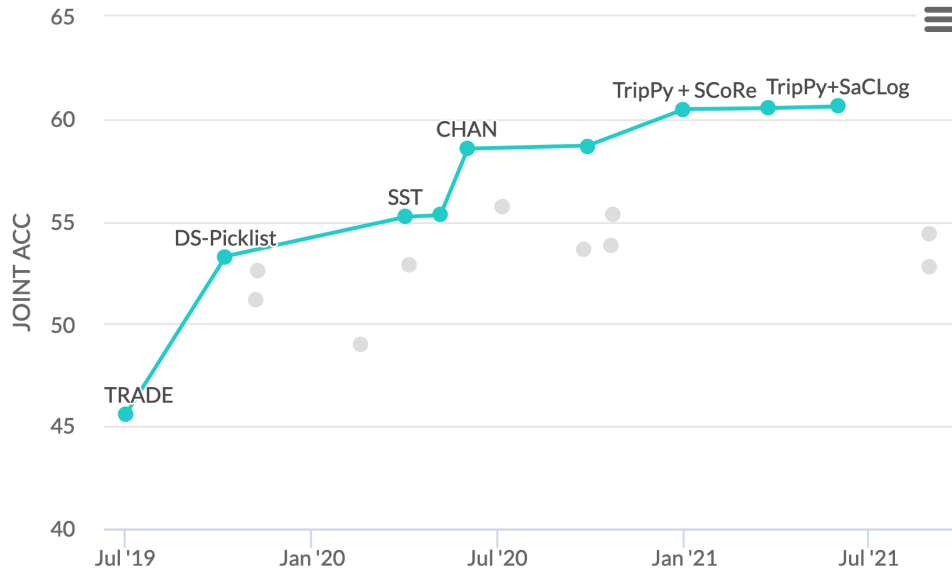
Annotated by
recording the UI of the
agent worker
Joint accuracy: 54.9

<https://paperswithcode.com/sota/multi-domain-dialogue-state-tracking-on>

Errors in Annotation

- UI traces do not capture exact DST at every turn
 - Agents delay entering search criteria by one-two turns
 - “What cuisine?” “Italian” “What price?” “cheap”
→ type cheap & Italian at once
- Agents resolve the answer in their mind only
 - “What cuisine?” “Italian” “What price?” “cheap”
→ type Italian, top result is cheap → reply to user

MultiWOZ 2.1 [Eric et al]



<https://paperswithcode.com/sota/multi-domain-dialogue-state-tracking-on-1>

- 2.1: complete reannotation
 - Move slots to the right turn, add slots that the agent ignored
 - fixed 32% of dialogue state annotations across 40% of the dialogue turns
- Joint accuracy
 - TripPyTripPy+ScoRe: **60.5%**
 - TripPy+CoCoAug: **60.5%**
 - TripPy+SaCLog: **60.6%**
- TripPy:
 - Constructs a label map to handle value variants.
 - Three-way loss to be robust to errors
 - Copy from sentence, context, or ontology
- ScoRe:
 - Schema-aware curriculum learning

Re-Annotations of MultiWOZ

- Reannotated **4 times** manually to fix annotation errors
 - 2.1: Move slots to the right turn, add slots that the agent ignored
 - 2.2: *Normalize types*: numbers, times, enums
 - 2.3: Fix **errors introduced** in 2.1
 - 2.4: Enforced *convention* on when to include a slot or not

Error Type	Conversation	MultiWOZ 2.1	MultiWOZ 2.4
Context Mismatch	Usr: Hello, I would like to book a taxi from restaurant 2 two to the museum of classical archaeology.	taxi-destination=museum of archaeology and anthropology	taxi-destination=museum of classical archaeology
Mis-Annotation	Usr: I need a place to dine in the centre of town.	rest.-area=none	rest.-area=centre
Not Mentioned	Usr: I am planning a trip in Cambridge.	hotel-internet=dontcare	hotel-internet=none
Multiple Values	Usr: Something classy nearby for dinner, preferably Italian or Indian cuisine?	rest.-food=Indian	rest.-food=Indian Italian
Typo	Usr: I am looking for a restaurant that serves Portuguese food.	rest.-food=Portugese	rest.-food=Portuguese
Implicit Time Processing	Usr: I need a train leaving after 10:00.	train-leaveat=10:15	train-leaveat=10:00
Slot Mismatch	Usr: Can you please help me find a place to go in town in the same area as the hotel? Preferably a college.	attraction-name=college attraction-type=none	attraction-name=none attraction-type=college
Incomplete Value	Sys: I recommend Charlie Chan. Would you like a table? Usr: Yes. Monday, 8 people, 10:30.	rest.-name=Charlie	rest.-name=Charlie Chan
Delayed Annotation	Usr: Please recommend one and book it for 6 people.	hotel-book people=none	hotel-book people=6
	Sys: I would recommend express by holiday inn Cambridge. From what day should I book? Usr: Starting Saturday. I need 5 nights for 6 people.	hotel-book people=6	hotel-book people=6
Unnecessary Annotation	Usr: I am looking for a museum. Sys: The Broughton house gallery is a museum in the centre. Usr: That sounds good. Could I get their phone number?	attraction-area=centre	attraction-area=none

Sentences That Cannot Be Represented As Slots

“I was hoping you could **recommend** something”.

“Are there any churches **or** museums on the east side?”

“I would like the **latest** train leaving that will arrive by 9:15 please”.

The agent cannot possibly return the result needed!

SOTA Architecture on MultiWOZ 2.4

- State of the art on version 2.4 [Ye et al 2021]:
 - Joint accuracy: **73.6%**
 - STAR achieves **56%** on 2.1
 - Annotations make a big difference!
 - Very specialized architectures, closed terminology, not general

STAR Architecture (Closed Ontology)

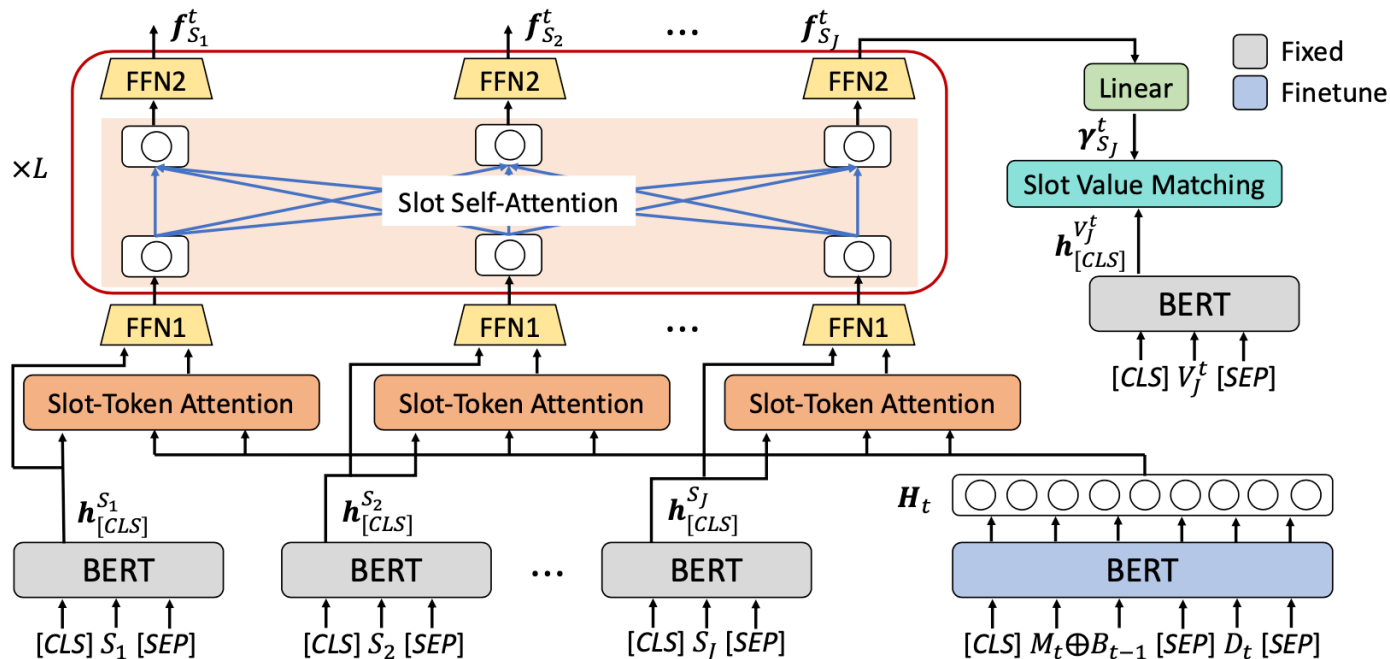


Figure 2: The architecture of our approach STAR. A fine-tuning BERT is used to encode dialogue contexts, another fixed BERT is utilized to generate aggregated vector representations for slots and values. For simplicity, layer normalization and residual connection are omitted, and only the value matching of slot S_j is included. Both FFN1 and FFN2 are feed-forward networks.

Questions from the STAR Result

- Are slots good enough?
 - How do we return the answer without precise representation?
- Would a closed ontology work?
- Is 73.6% good enough?
- Who can afford reannotations?
- How to cover the world's conversations in all domains and in all languages?

The answer is nope!

Limitations of MultiWOZ Data Set

- Issues
 - Closed taxonomy of 1,829 values: not realistic
 - Dialogue state representation not expressive enough
 - Need to change from slots to queries
 - Estimated error rate for MultiWOZ 2.1 is at least 15%
 - Does not seem to be surmountable (4 re-annotations)
- Results: joint accuracy = 60.6% (22% in 3 turns)
 - Specialized architecture of MultiWOZ 2.4 = 73.6%
 - Not scalable, not scientifically worthwhile

Limitations of Wizard-of-Oz Dialogues

1. Missing real-world state transitions
(based on crowdworker instructions)
 - Predefined what product to search
(user never changes their mind)
 - No out-of-domain questions, no info beyond slot values
 - No interruptions of what the agent asks
2. Fundamentally
 - Human annotation of dialogue states is error-prone!

Lecture Goals

- Agent architectures:
 - (1) LLMs, (2) dialogue trees, (3) contextual semantic parser
- Dialogue State Tracking Problem
 - Speech Act Theory
 - Task-oriented agents
 - Data acquisition methods
 - Wizard-of-Oz dialogues
 - **Synthesized dialogues**

Machine-2-Machine Synthesized Dialogues

- SGD (schema-guided dialogue dataset)
- To synthesize data from a given schema
- Both user and agents have dialogue acts
 - 11 user dialogue acts, 10 system dialogue acts
 - (Probabilistic) domain independent rule-based systems to implement the state machine
- Meaning representation: slot-value

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan (2019).

“Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset”. In: arXiv:1909.05855

Data Acquisition (Synthesize → paraphrase)



- a. Generated scenario with (Probabilistic) domain independent rule-based systems
- b. Vary the expression of the value
- c. Generation with 1 template
- d. Crowdsourced paraphrase

Variety relies heavily
on manual paraphrases!

Train, validate, and test with paraphrase

SGD Dataset

Metric ↓ Dataset →	DSTC2	WOZ2.0	FRAMES	M2M	MultiWOZ	SGD
No. of domains	1	1	3	2	7	16
No. of dialogues	1,612	600	1,369	1,500	8,438	16,142
Total no. of turns	23,354	4,472	19,986	14,796	113,556	329,964
Avg. turns per dialogue	14.49	7.45	14.60	9.86	13.46	20.44
Avg. tokens per turn	8.54	11.24	12.60	8.24	13.13	9.75
Total unique tokens	986	2,142	12,043	1,008	23,689	30,352
No. of slots	8	4	61	13	24	214
No. of slot values	212	99	3,871	138	4,510	14,139

Synthesis/Paraphrase: many more dialogues than MultiWOZ

Cost of paraphrases (including verification) is still a limiting factor

Much lower annotation error

Evaluation

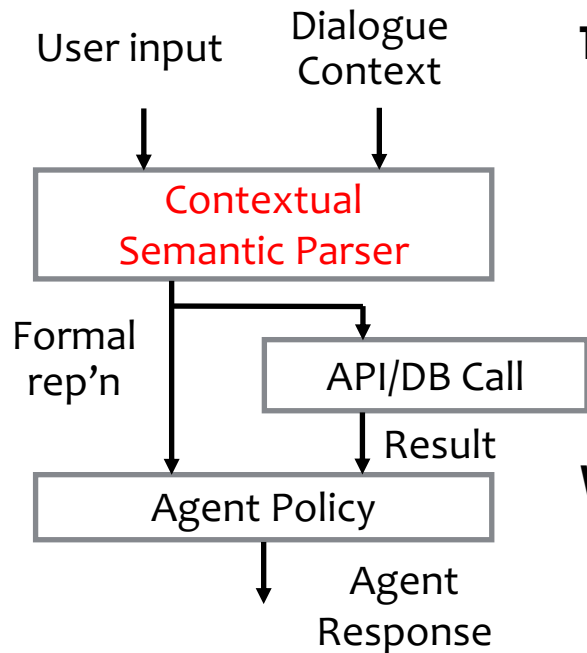
- Evaluation set: synthesized then paraphrased
 - >= **95%** accuracy intent classification
 - >= **95%** F1 score slot tagging

Quiz: Great results! Are we done?

Summary of M2M Dialogues

- M2M – not useful for as a real test
 - All dialogue acts, transitions, original utterances are derived from a small state machine
 - Not meaningful because it is easily solved
- But, M2M dialogues can be used to generate training data!

Summary



Speech Act Theory

The dialogue state tracking (DST) problem

- Contextual semantic parsing (with dialogue history)

	More Realistic	Correct Annotation
H2H: Wizard-of-Oz	✓	✗
M2M: Synthesized	✗	✓

What should we do?

- Train with: few-shot H2H + synthesized M2M data (correct annotation)
- Validate / Test on H2H