Stanford CS224v Course

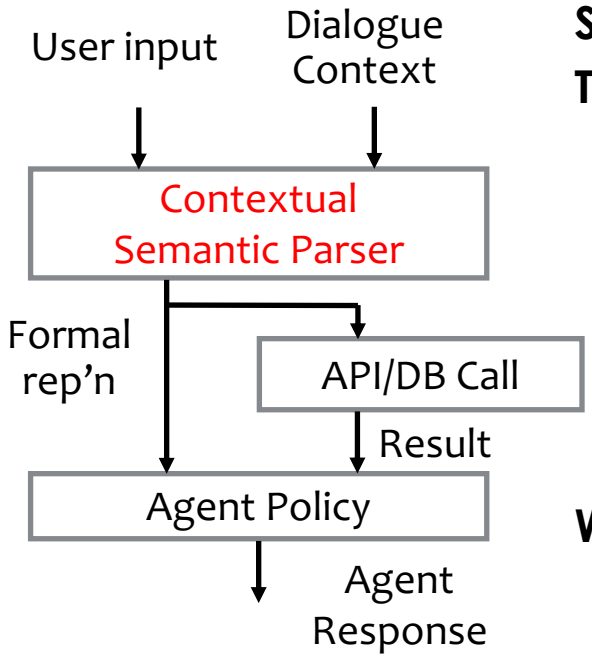
Conversational Virtual Assistants with Deep Learning

Lecture 7

Task-Oriented Dialogue Agents

Monica Lam

Lecture 6 Summary



Speech Act Theory

The dialogue state tracking (DST) problem

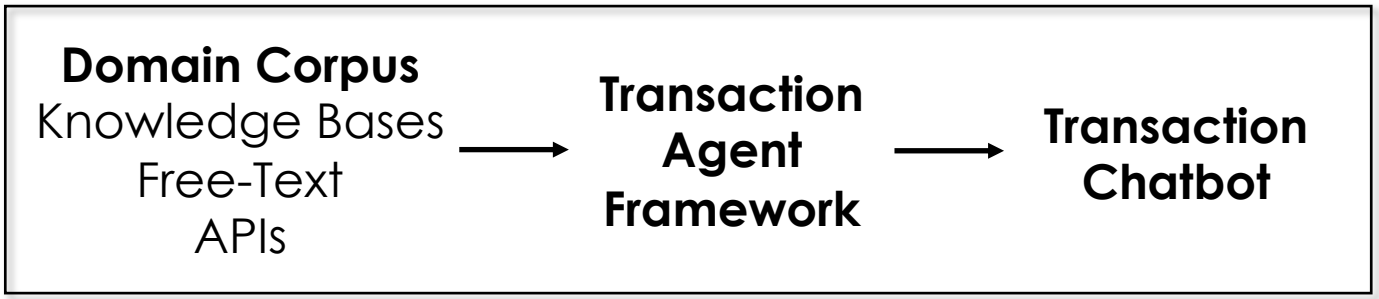
- Contextual semantic parsing (with dialogue history)

	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

Lecture Goals



- DST
  - How to track dialogue states for **long dialogues**?
  - How well do **LLMs** track dialogue states?
- Agent policy
  - How to implement a **neural agent policy?**
  - How to implement a **rule-based agent policy**?
    - Do real **H2H dialogues** have agent dialogue acts?

RiSAWOZ

MultiWOZ

RiSAWOZ: a Large Dataset in Chinese

Type	Single-domain					Multi-domain			
Dataset	**DSTC2**	**WOZ 2.0**	**FRAMES**	**KVRET**	**M2M**	**MultiWOZ**	**SGD**	**CrossWOZ**	**RiSAWOZ (ours)**
Language	EN	EN	EN	EN	EN	EN	EN	ZH	**ZH**
Speakers	H2M	H2H	H2H	H2H	M2M	H2H	M2M	H2H	**H2H**
Domains	1	1	1	3	2	7	16	5	**12**
Dialogues	1,612	600	1,369	2,425	1,500	8,438	16,142	5,012	**10,000**
Turns	23,354	4,472	19,986	12,732	14,796	115,424	329,964	84,692	**134,580**
Avg. turns	14.5	7.5	14.6	5.3	9.9	13.7	20.4	16.9	**13.5**
Slots	8	4	61	13	14	25	214	72	**159**
Values	212	99	3,871	1,363	138	4,510	14,139	7,871	**4,061**
Linguistic annotation	No	No	No	No	No	No	No	No	**Yes**

Table 1: Comparison of our dataset to other task-oriented dialogue datasets (training set). H2H, H2M, M2M represent human-to-human, human-to-machine, machine-to-machine respectively.

Jun Quan, et al.. 2020. RiSAWOZ: A large-scale multi- domain Wizard-of-Oz dataset with rich semantic an- notations for task-oriented dialogue modeling. In *EMNLP*, pages 930–940, Online. ACL.

RiSAWOZ Dataset

Domain	Attraction	Restaurant	Hotel	Flight	Train	Weather	Movie	TV	Computer	Car	Hospital	Courses
Entities	40	50	70	665	2,016	70	100	105	50	52	17	90
Slots	12	11	11	9	10	6	10	11	20	23	16	20

Domain	Informable Slots	Requestable Slots
Attraction	name, area, type, consumption, metro station	metro station, ticket price, phone number, address, score, opening hours, features
Restaurant	name, area, cuisine, pricerange, metro station	metro station, per capita consumption, address, phone, score, business hours, dishes
Hotel	name, area, star, pricerange, hotel type, room type, parking	parking, room charge, address, phone, score
Flight	departure, destination, date, class cabin	flight information, departure time, arrival time, ticket price, punctuality rate
Train	departure, destination, date, train type, seat type	train number, duration, departure time, arrival time, ticket price
Weather	city, date	weather condition, temperature, wind, UV intensity
Movie	production country or area, type, decade, movie star	director, movie title, name list, release date, film length, Douban score
TV	production country or area, type, decade, tv star	director, TV title, name list, premiere time, episodes, episode length, Douban score
Computer	brand, computer type, usage, memory capacity, screen size, CPU category, pricerange, series	product name, operating system, game performance, CPU model, GPU category, GPU model, features, colour, standby time, hard disk capacity, weight
Car	series, classification, size, number of seats, brand, hybrid, power level, 4WD, fuel consumption, price range	parking assist system, cruise control system, heated seats, ventilated seats
Hospital	name, level, type, public or private, area, general or specialized, key departments, metro station	address, phone, registration time, service time, bus routes, CT, 3.0T MRI, DSA
Class	grade, subject, level, Day, Time, area	campus, start date, end date, start time, end time, times, hours, classroom, teacher, price

Table 3: All informable and requestable slots in each domain.

Data Statistics

	Train	Dev	Test	Single-domain	Multi-domain	All
Dialogues	10,000	600	600	7,261	3,939	11,200
Turns	134,580	8,116	9,286	88,618	63,364	151,982
Tokens	1,462,727	87,886	108,032	954,298	704,347	1,658,645
Vocab.	11,486	4,514	4,927	10,065	8,031	11,971
Avg. turns	13.46	13.53	15.48	12.20	16.09	13.57
Avg. tokens	10.87	10.83	11.63	10.77	11.12	10.91

RiSAWOZ Data Collection

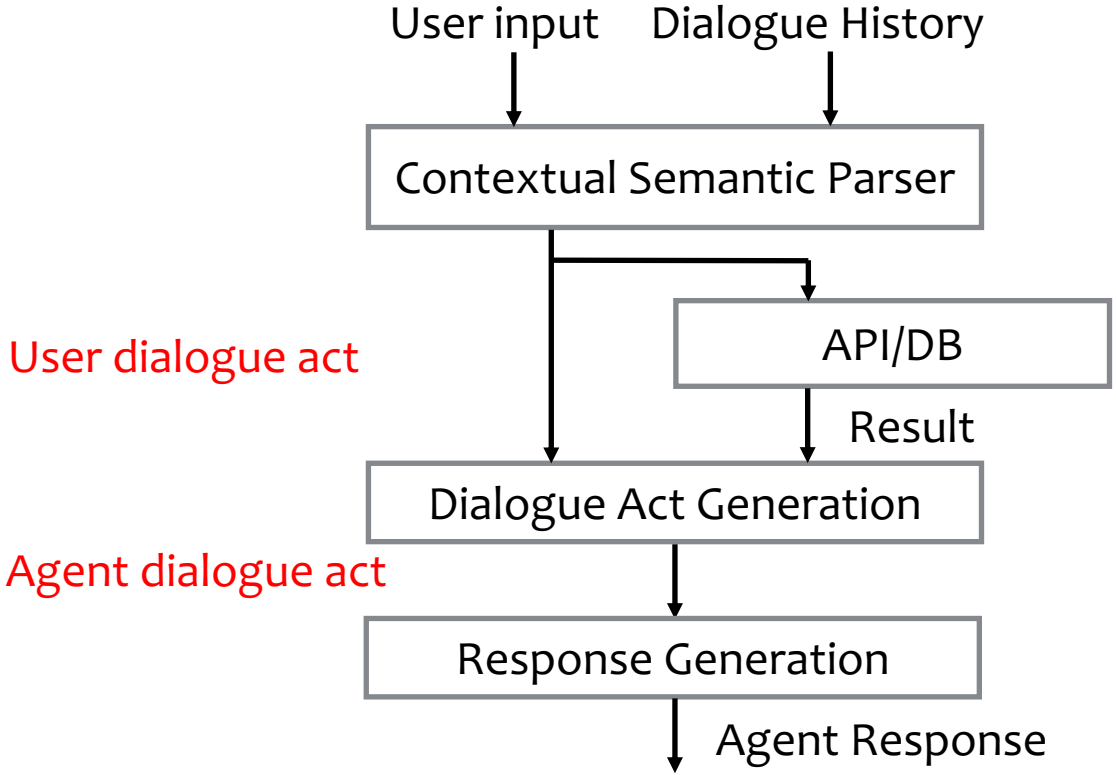
Predefined dialogue acts

User action type	Inform, request, greeting, bye, general
System action type	Inform, recommend, request, no-offer, greeting, bye, general

- Workers were trained and qualified
- User: given a set of constraints of product to search
- Agent worker:
  - Annotates the user dialogue acts
  - Provides the response and marks the agent dialogue state
- Result: a much more consistent annotation

Quiz: agent responses are more restrictive. Is this a good dataset?

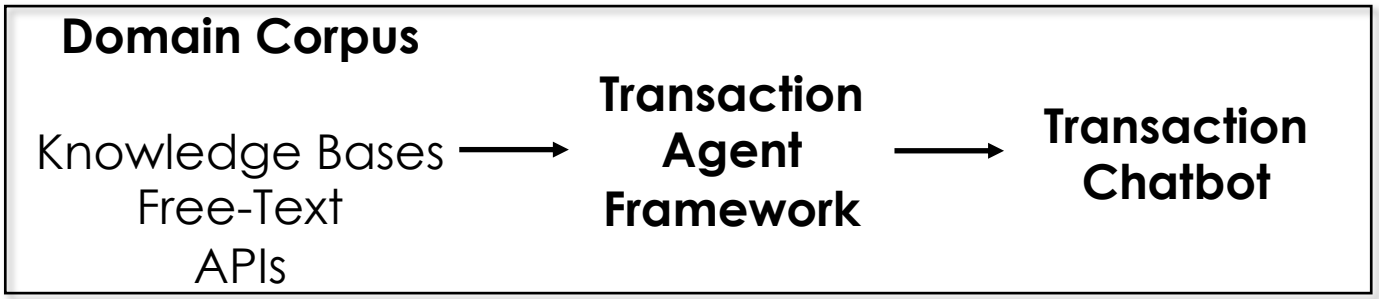
RiSAWOZ: Dataset for the Entire Agent



4 Problems:

- DST: Dialogue State Tracking
  - Semantic parsing
- API call detection
- DAG: Dialogue Act Generation (agent policy)
- RG: Response Generation

Lecture Goals



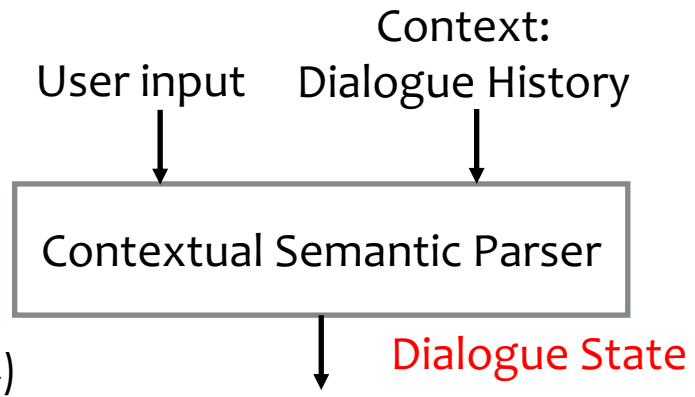
- DST
  - How to track dialogue states for **long dialogues**?
  - How well do **LLMs** track dialogue states?
- Agent policy
  - How to implement a **neural agent policy?**
  - How to implement a **rule-based agent policy**?
    - Do real **H2H dialogues** have agent dialogue acts?

RiSAWOZ

MultiWOZ

Original Dialogue State Tracking (DST) Problem

Dialogue State (DS): Meaning of the user utterance



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

For each turn i

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

Training data:

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

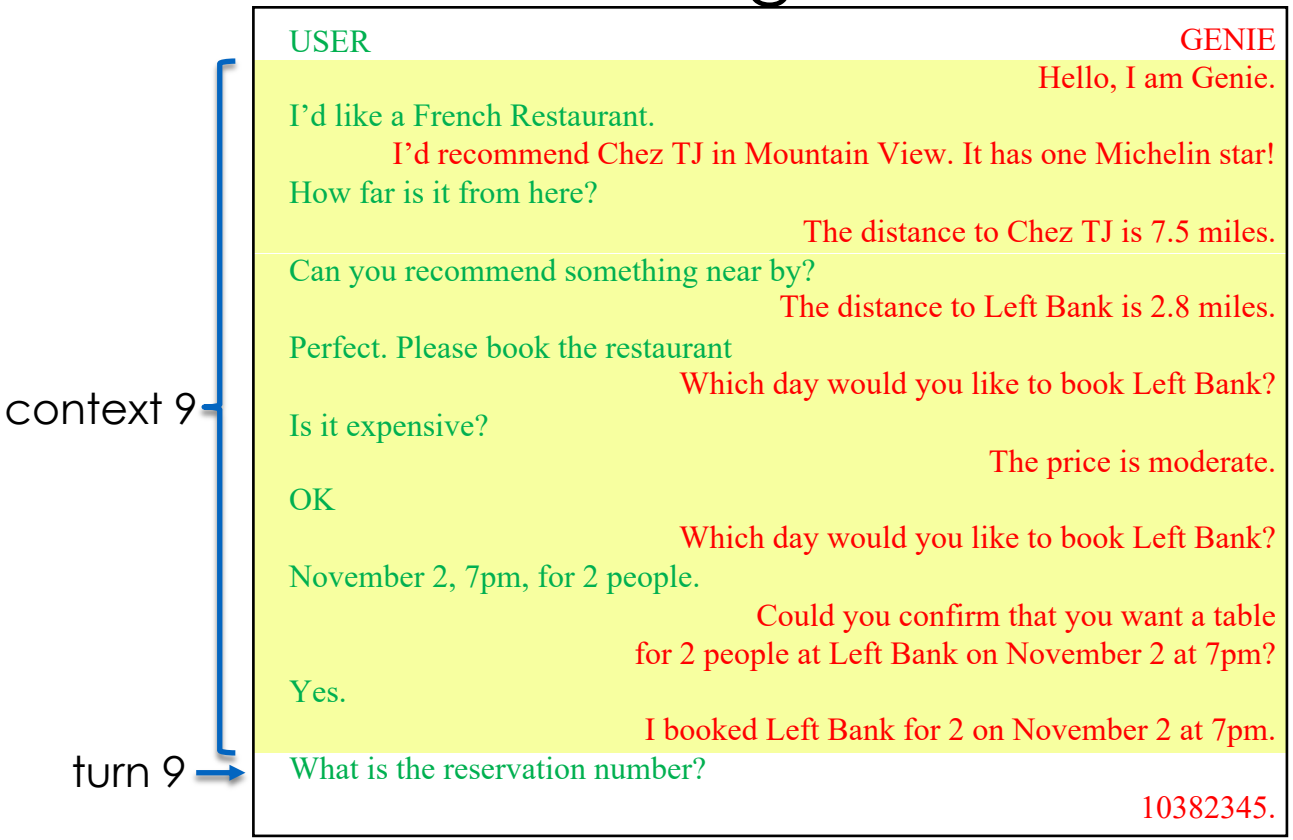
A dialogue with n turns → n training samples

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.

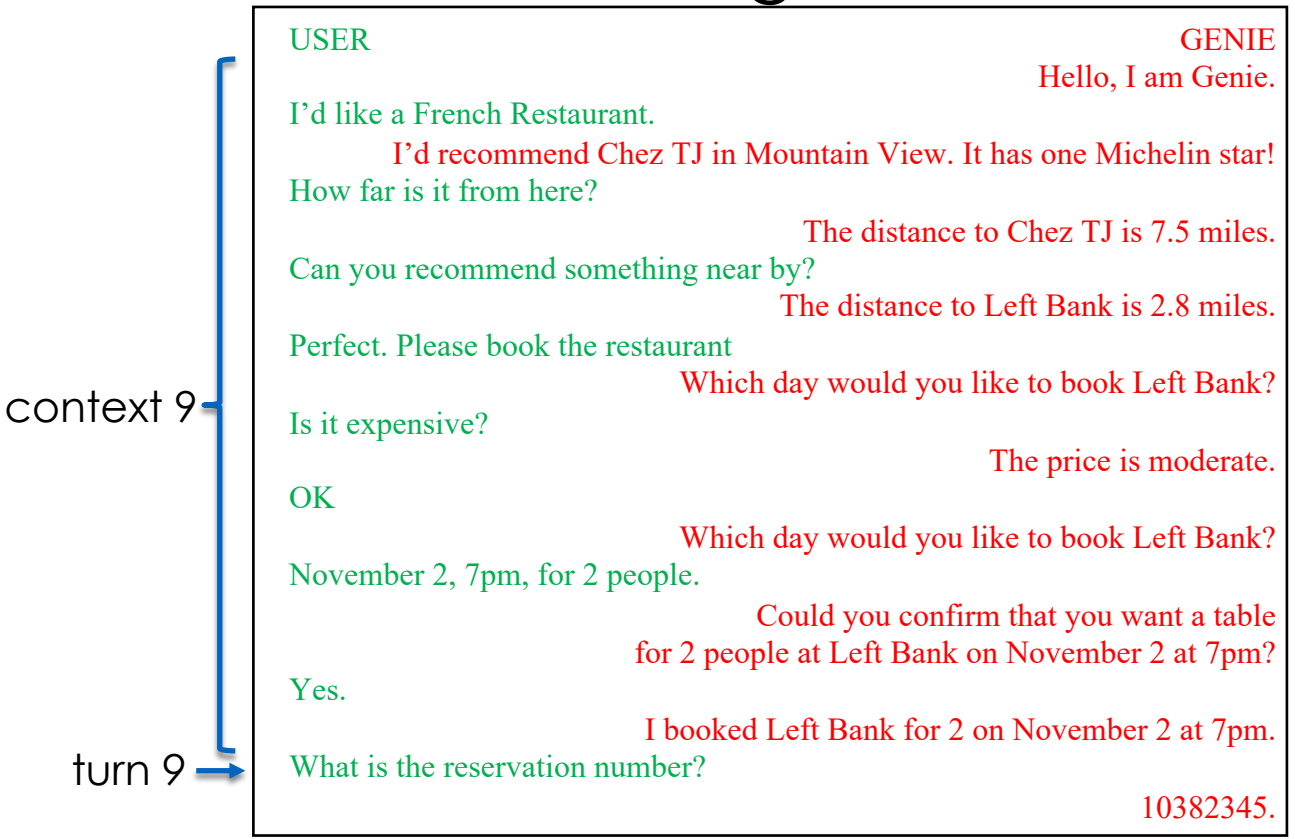
Context of Dialogue State Tracking



Quiz:

Can we make parsing conversation easier?

Context of Dialogue State Tracking



Parses more & more user and agent utterances!

- Do we need the whole conversation?
- Why do we have to parse the agent state?
- Why do we reparse the user utterances over and over again?

History: Replace NL with Formal States

4 Tasks: DST, ACD, DAG, RG

- Fine-tune 1 multi-task neural network (MBART)

Formal state captures all useful info

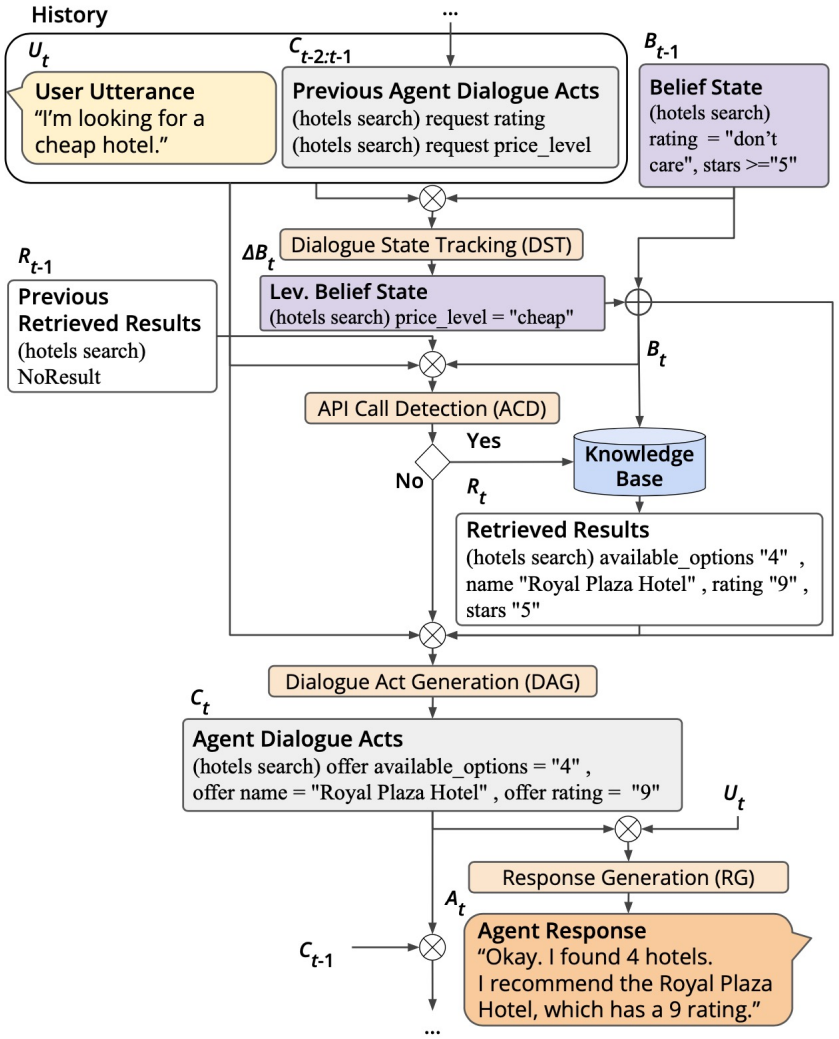
- U: user utterance
- B: belief state (part of user dialogue state)
- R: results
- C: agent dialogue act
- A: Agent utterance

Context for semantic parser:

- what is needed to understand the user

Levenstein belief state:

- What is new in the user utterance
- Added to the previous state to get the full query



Mehrad Moradshahi, et al. 2023. X-RiSAWOZ: High-quality end-to-end multilingual dialogue datasets and few-shot agents. In *Findings of ACL 2023*, pages 2773–2794, Toronto, Canada.

SOTA Result

Joint goal accuracy	API call detection	Agent dialogue act accuracy	Response BLEU score
78.2	72.7	73.7	34.7

- Much better results than previous MultiWOZ semantic parsers
- Quiz: why?

Mehrad Moradshahi, et al. 2023. X-RiSAWOZ: High-quality end-to-end multilingual dialogue datasets and few- shot agents. In *Findings of ACL 2023*, pages 2773–2794, Toronto, Canada.

X-RiSAWOZ MultiLingual Dataset

- A few-shot training data set (100 train, 600 dev, 600 test)
- Chinese: Original language
- English: translated manually
- Rest:
  - Auto-translated from English
  - Manually fixed with many tools

Languages	#People
English	1.5B
Chinese (written)	1.3B
Hindi	0.6B
Hindi+English (mixed code)	
French	0.3B
Korean	0.08B

X-RiSAWOZ: High-Quality End-to-End Multilingual Dialogue Datasets and Few-shot Agents, Moradshahi, Mehrad et al. ACL Findings 2023.

X-RiSAWOZ Dataset Stats

	Dataset		
	Few-shot	**Validation**	**Test**
# Domains	12	12	12
# Dialogues	100	600	600
# Utterances	1,318	8,116	9,286
# Slots	140	148	148
# Values	658	2,358	3,571

Full Training: 10,000 dialogues!

DST on Dev

These results are hard to get
-- need to fix many translation errors
and agree on the conventions

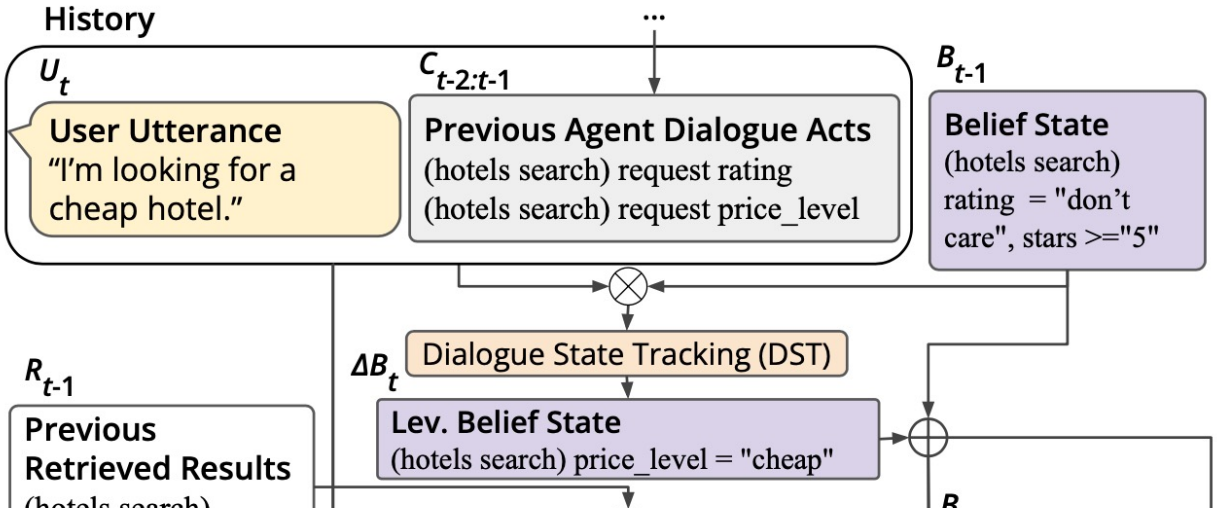
Language	Full-shot Training	Few-shot Training*	Zero-shot Training*
Chinese	96.4	82.8	
English		84.6	84.2
Korean		71.2	34.6

*For English: Includes machine-translated full-shot data from Chinese
*For Korean: machine-translated full-shot data from English

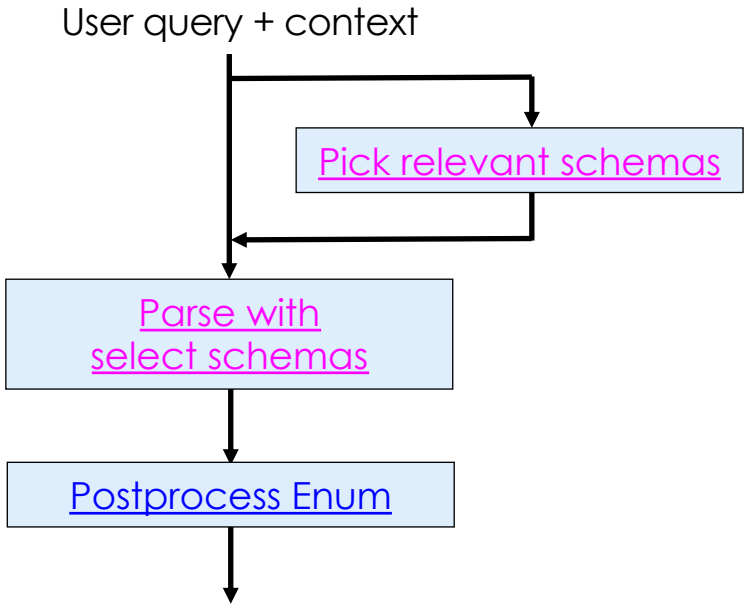
- Training is fine-tuning MBart
- Optimizations: dictionary and neural alignment

What if We Use GPT-4?

- Same formal state as context



Semantic Parsing with GPT-4



Courtesy of Andrew Lee

Click the links to see the prompts (written in jinja syntax)

- Pick relevant schemas
  - Choose 1 or more of the 12 domains
  - 1 example for each domain
- Parse with select schemas
  - 7 examples for the given domain covering all the slots
- Postprocess enum
  - LLMs do not know the enumerated values
  - Shows the enumerated values for slots used

Results with GPT-4

Language	Full-shot Training	Few-shot Training*	Zero-shot Training*	GPT-4
Chinese	96.4	82.8		
English		84.6	84.2	74.9
Korean		71.2	34.6	63.7

*For English: Includes machine-translated full-shot data from Chinese
*For Korean: machine-translated full-shot data from English

Results with GPT-4

Language	Full-shot Training	Few-shot Training*	Zero-shot Training*	GPT-4	GPT-4 + postprocess
Chinese	96.4	82.8			
English		84.6	84.2	74.9	79.3
Korean		71.2	34.6	63.7	68.6

*For English: Includes machine-translated full-shot data from Chinese
*For Korean: machine-translated full-shot data from English

Quiz: Is this the End of the Story?

Quiz: What should we do now?

Error Analysis on GPT-4 (English)

Reference: Few-shot + auto-translated training has an accuracy of 84.6

Factors	Cause of errors	%	Overall Accuracy
Measured			79.3
Poor gold label		9.3	88.6
Convention Mismatch		8.3	96.8
Errors		3.2	
	Slots	1.7	
	Post-processing	0.8	
	Slot value	0.4	
	Domains	0.3	

Error Analysis on GPT-4 (Korean)

Reference: Few-shot + auto-translated training has an accuracy of 71.2

Factors	Cause of errors	%	Overall Accuracy
Measured			68.6
Poor gold label		15.2	83.8
Convention Mismatch		9.8	93.6
Errors		6.4	
	Slots	2.2	
	Post-processing	1.8	
	Slot value	2.0	
	Domains	0.3	

Discussion

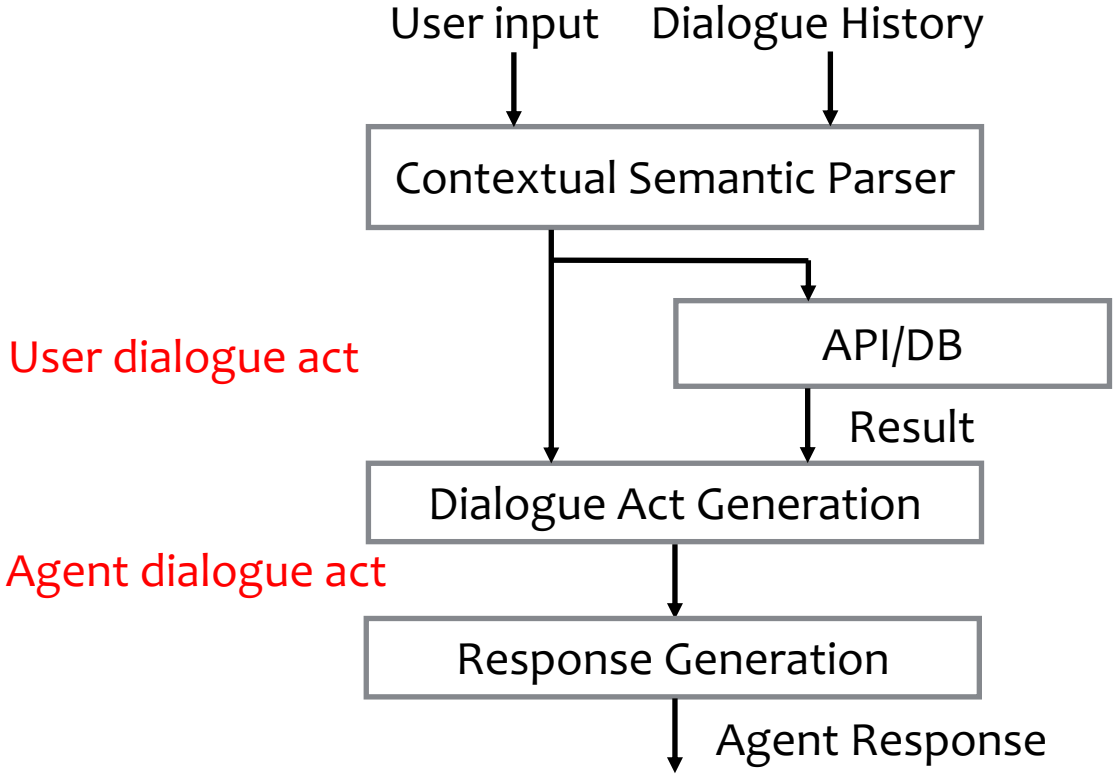
Language	Full-shot Training	Few-shot Training*	Zero-shot Training*	GPT-4	GPT-4 + postprocess	Manual Eval
Chinese	96.4	82.8				
English		84.6	84.2	74.9	79.3	96.8
Korean		71.2	34.6	63.7	68.6	93.6

- Training results better aligned with annotations → Not important
- Be careful with automatic evaluation!
  - Benchmark is useful: only need to examine 20% of the data
- This is not a difficult dataset

LLM's Usefulness and Limitations

Method	Applicability
In-Context Learning	Hospitality domain: • 12 domains, 159 total slots (RiSAWOZ) • < 10 attributes for SQL queries (Yelp)
Fine-tuning LLaMA with few-shot data	• WikiData for WikiWebQuestions
Hypothesis: Fine-tuning LLaMA with few-shot + synthesized data	• Large databases • WikiData for less popular topics • Complex SPARQL queries

RiSAWOZ: Dataset for the Entire Agent



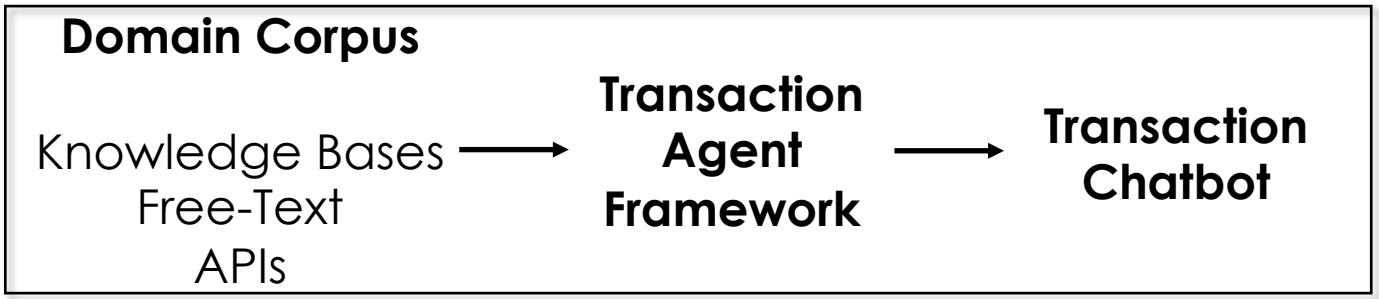
4 Problems:

- DST: Dialogue State Tracking
  - Semantic parsing
- API call detection
- DAG: Dialogue Act Generation (agent policy)
- RG: Response Generation

Neural Agent Policies

- **Relies on training data – expensive**
  - User task and agent policy for RiSAWOZ were pre-determined before data collection
  - How do we know those are the right agent dialogue acts?
  - Does the training data cover what users may say?
- **Cannot control precisely what the agent says**
  - Mimic human agents; but why should it be statistical?
  - There is a difference between good and bad agents
- **Hard to update a policy!**

Lecture Goals

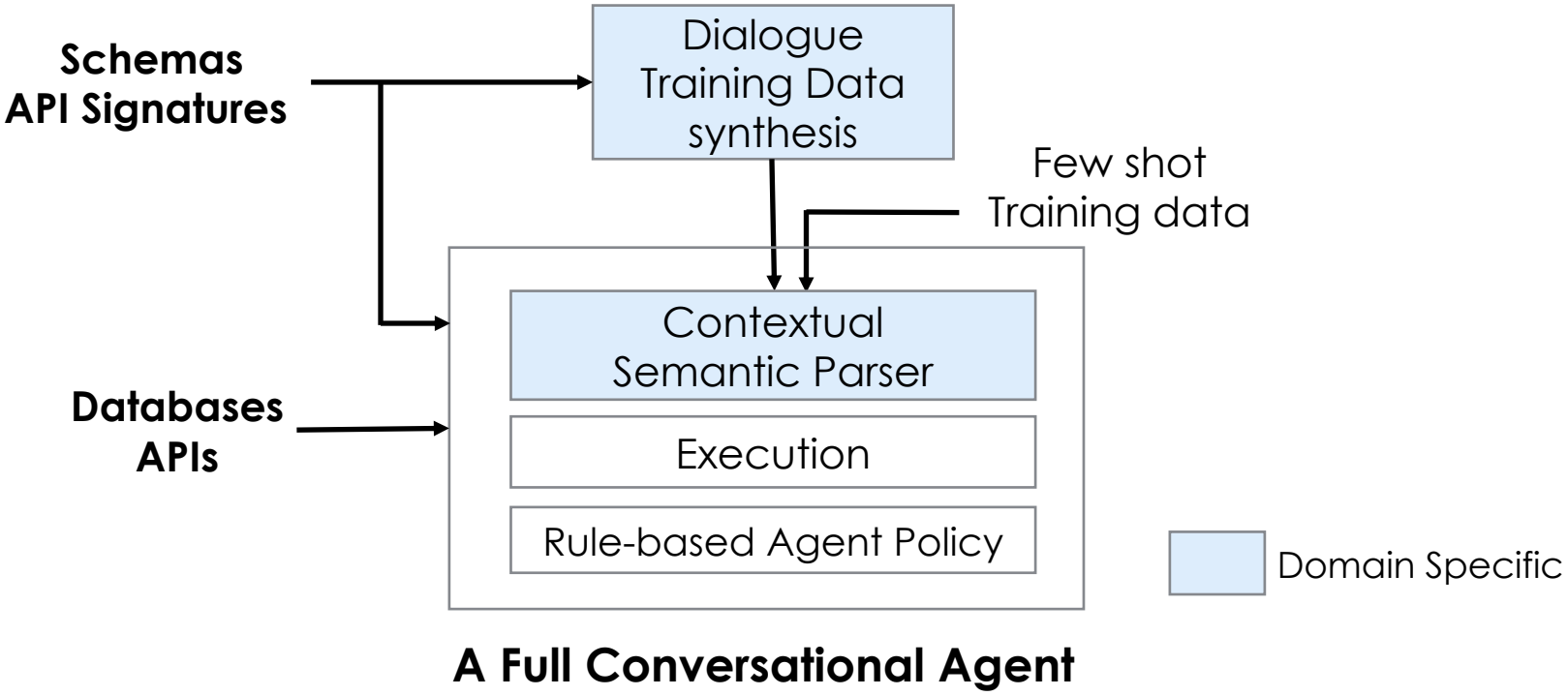


- DST
  - How to track dialogue states for **long dialogues**?
  - How well do **LLMs** track dialogue states?
- Agent policy
  - How to implement a **neural agent policy?**
  - How to implement a **rule-based agent policy**?
    - Do real **H2H dialogues** have agent dialogue acts?

RiSAWOZ

MultiWOZ

Generating Transaction Agents Automatically (before LLM)



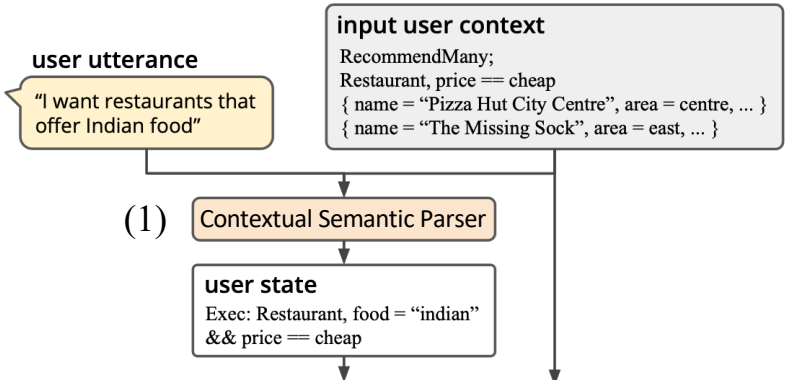
MultiWOZ 3.0

- MultiWOZ 2.1
  - Has too many errors
  - The slot-value pairs are not precise enough
- MultiWOZ 3.0
  - Change the slot-value pairs to ThingTalk
    - Similar to SQL for queries, but support API calls
    - With dialogue acts
  - **Reannotated 2% training data, dev set, test set**

Giovanni Campagna, et al. A few-shot semantic parser for Wizard-of-Oz dialogues with the precise ThingTalk representation. In *Findings of the ACL* pages 4021–4034, Dublin, Ireland.

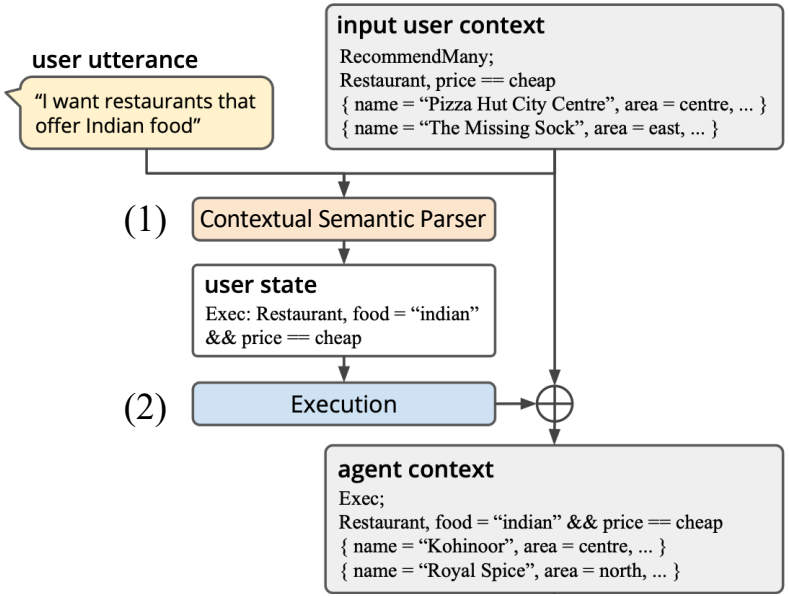
Agent Anatomy

1. **Contextual Semantic Parser**:
   (Dialogue State Tracking DST)
   User utterance + context → user state



Agent Anatomy

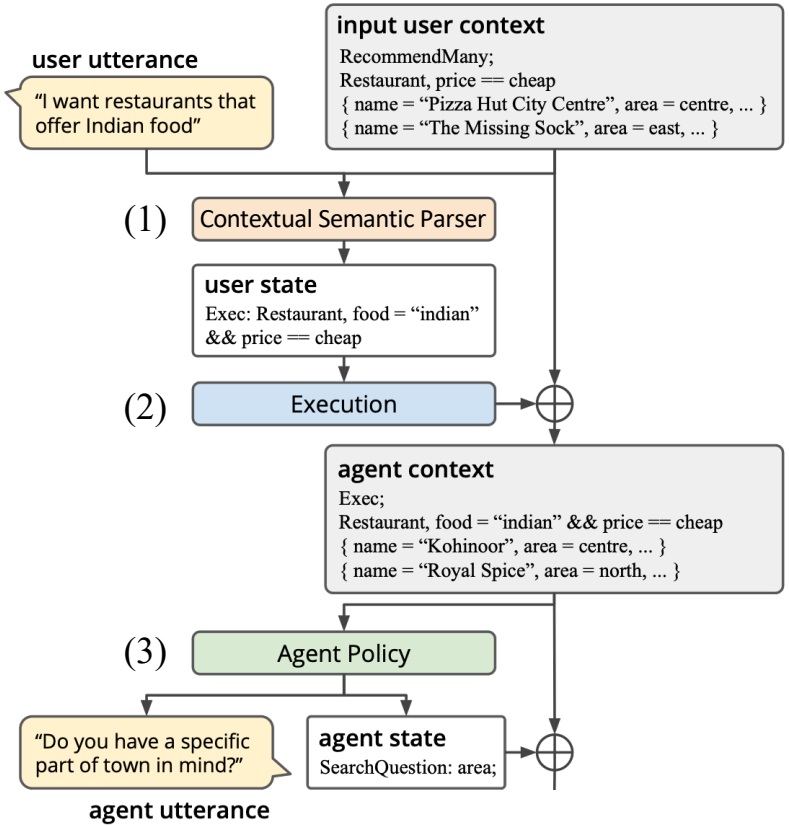
1. **Contextual Semantic Parser**:
   (Dialogue State Tracking DST)
   User utterance + context → user state
2. **ThingTalk Execution:**
   ThingTalk → result state



Agent Anatomy

1. **Contextual Semantic Parser**:
   (Dialogue State Tracking DST)
   User utterance + context → user state
2. **ThingTalk Execution:**
   ThingTalk → result state
3. **Agent policy:** code to decide the response
   User context + result → agent state

   Natural Language Generator (RG):
   agent state → agent utterance



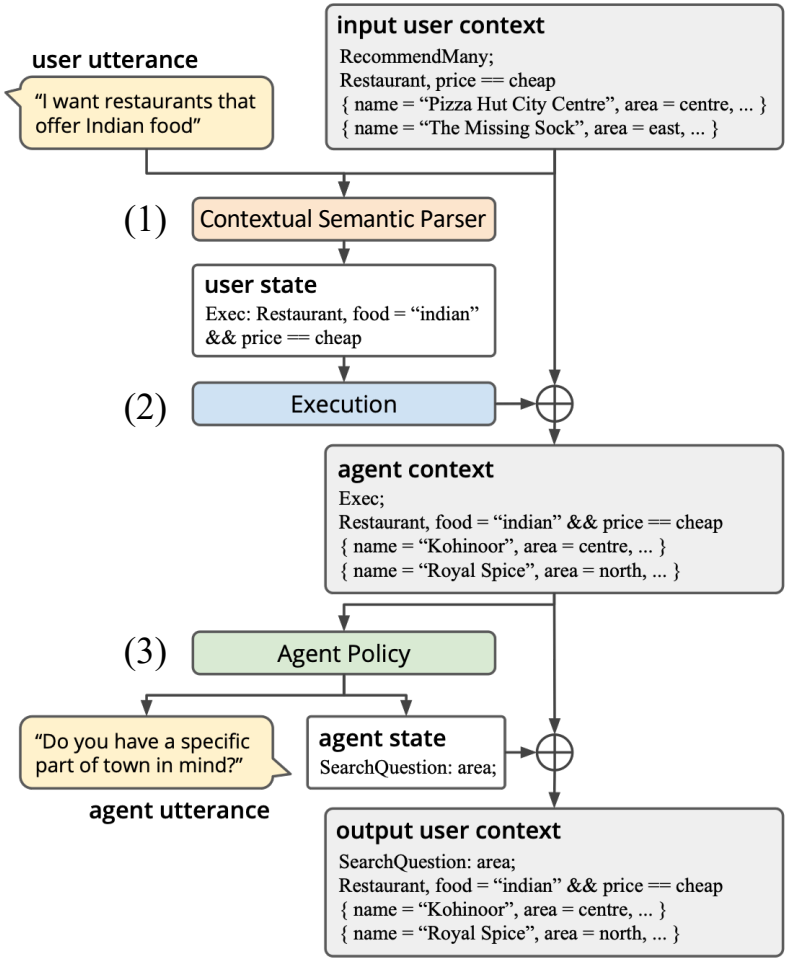
Agent Policy is Rule-Based, Not Neural

Agent Anatomy

1. **Contextual Semantic Parser**:
   (Dialogue State Tracking DST)
   User utterance + context → user state
2. **ThingTalk Execution:**
   ThingTalk → result state
3. **Agent policy:** code to decide the response
   User context + result → agent state

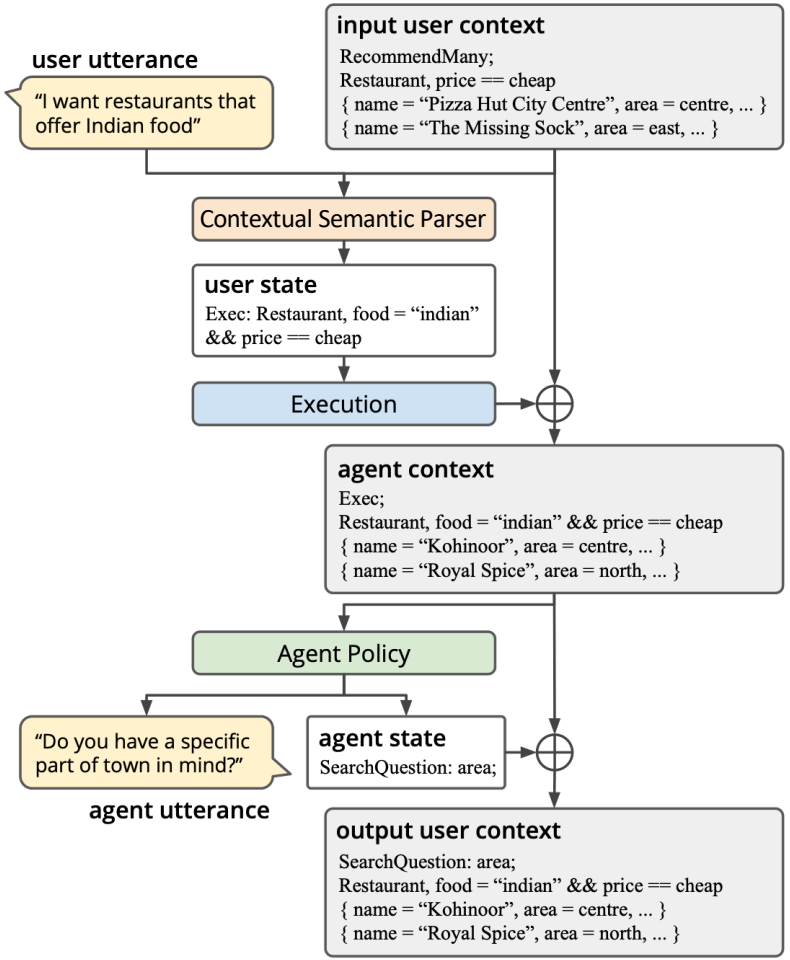
   Natural Language Generator (RG):
   agent state → agent utterance
4. **Output Context**:
   Summarizes the dialogue to date
   Becomes input context for next turn

Agent Policy is Rule-Based, Not Neural



Formal Dialogue State

- **Full dialogue history**:
  For each KB and action in each domain
  - One outstanding request
  - History of completed requests (with answers)
  - Note: each request may represent a long discussion
- **Context used** for semantic parsing
  - Limit to the last 5 ongoing/completed statements



MultiWOZ: H2H Dialogues

- Only the user dialogue acts are labeled
  - Not the agent
- The agent utterances for MultiWOZ have more variations than RiSAWOZ – more natural

MultiWOZ dataset: 5 hospitality domains

User requests:
values for some slots

- Cuisine type, location, #people reservations
- But users usually decide based on availability

Agents complete the transaction on the DB.

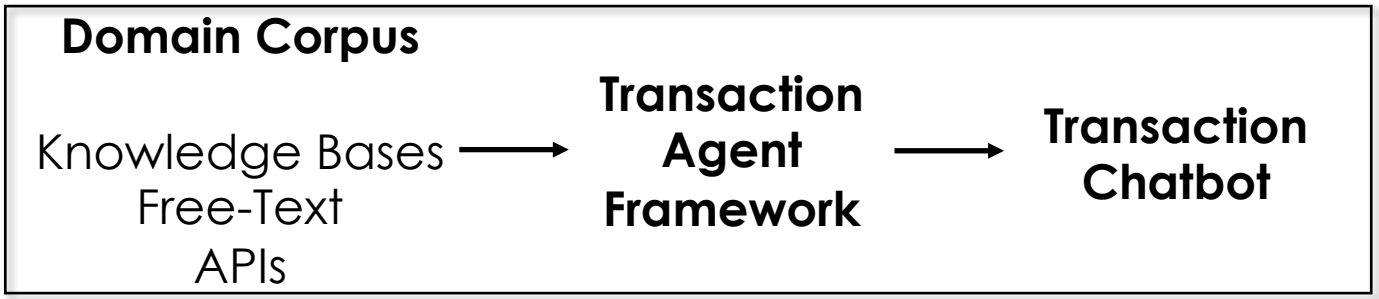
- Answer questions
- Slot-fill
- May offer to book a restaurant, hotel etc. (Not all passive)

Dialogue states:

- Crowdsource workers are not aware of them

Conversation is more natural (than M2M)

Lecture Goals

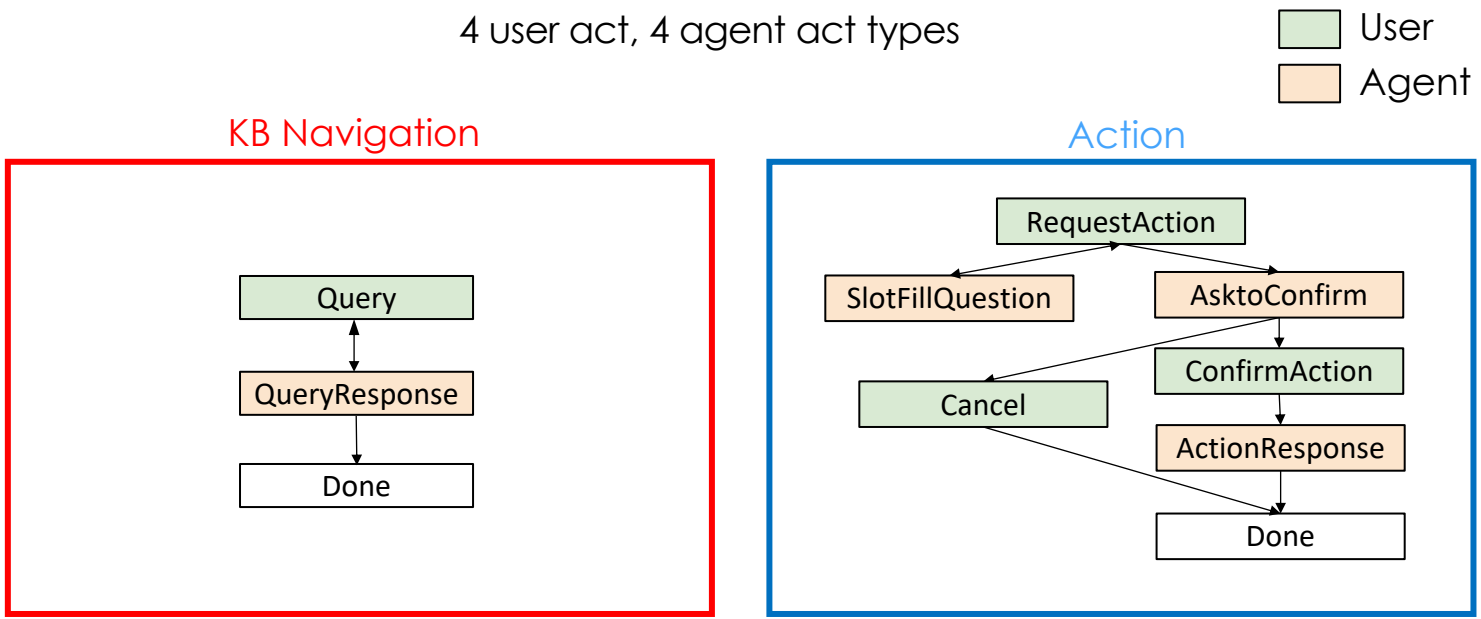


- DST
  - How to track dialogue states for **long dialogues**?
  - How well do **LLMs** track dialogue states?
- Agent policy
  - How to implement a **neural agent policy?**
  - How to implement a **rule-based agent policy**?
    - Do real **H2H dialogues** have agent dialogue acts?

RiSAWOZ

MultiWOZ

Recall: a Simple Dialogue State Machine

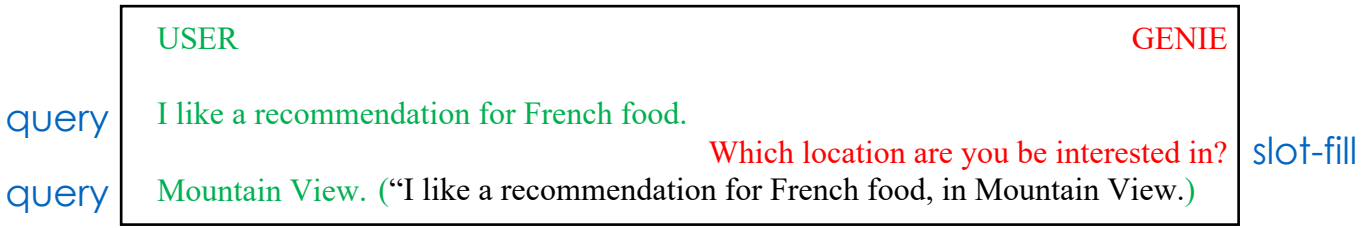


Do H2H Transaction Dialogues Have a Small # of Dialogue Acts?

Methodology

- Sample conversations
- Derive a domain-independent state-machine
  - Dialogue act types: 10 for users, 20 for agents

Example: collapses user state based on overall intent



User Dialogue Acts (10)

User initiative	execute; <query>	"Please "find me a restaurant", "Book a table"
User responses to agent utterances	ask_recommend	[Agent: "what food would you like"] User: "Can you give me a recommendation?"
	insist	"Can you check again?"
	learn more	"Can you tell me more about that?"
	action-query (<param>)	"What was the reservation number?"
Conversational flow	greet, reinit	"Hi", "I wanna ask about something else"
	cancel	"Thank you, that will be enough"
	end	"Good bye!"
	init	Start state

Agent Dialogue Acts (20)

Slot Fill	slot_fill (<slot>)	"How many people is the reservation for?"
Simple result	display_result	"It is cloudy today at Stanford" (pure text output)
Query flow (agent Initiatives)	search_question (param>)	"Do you have a preference of the kind of food?"
	recommend[1|2|3|4|*]: [<query>]	"I found X and Y. Both are restaurants that …" "I found X. Would you like to book it?"
	learn_more_what	"What would you like to know about?"
	empty_search [_question (<param>)]	"I couldn't find any restaurant like that." "Would you like a different area?"
Action flow	confirm_action	"OK I'll book Terun for 5 ppl. Is that correct?"
	action_success	"I booked Terun for you"
	action_error [_question(<param>)]	"Sorry, I could not book Terun: how about Jinsho?"
Conversational Flow	greet	"Hi, how can I help you?"
	anything_else	"Can I help with you anything else?"
	end	"Alright, bye!"

Relatively Few Transitions

- Transitions:
  - 20 transitions from agent to user acts
  - 43 transitions from user acts to agent acts
- Non-deterministic agent behavior
  - Same user act & result → different agent acts since different source workers may behave differently
  - Developer can just always use one, or choose probabilistically for variety
- Domain Agnostic!

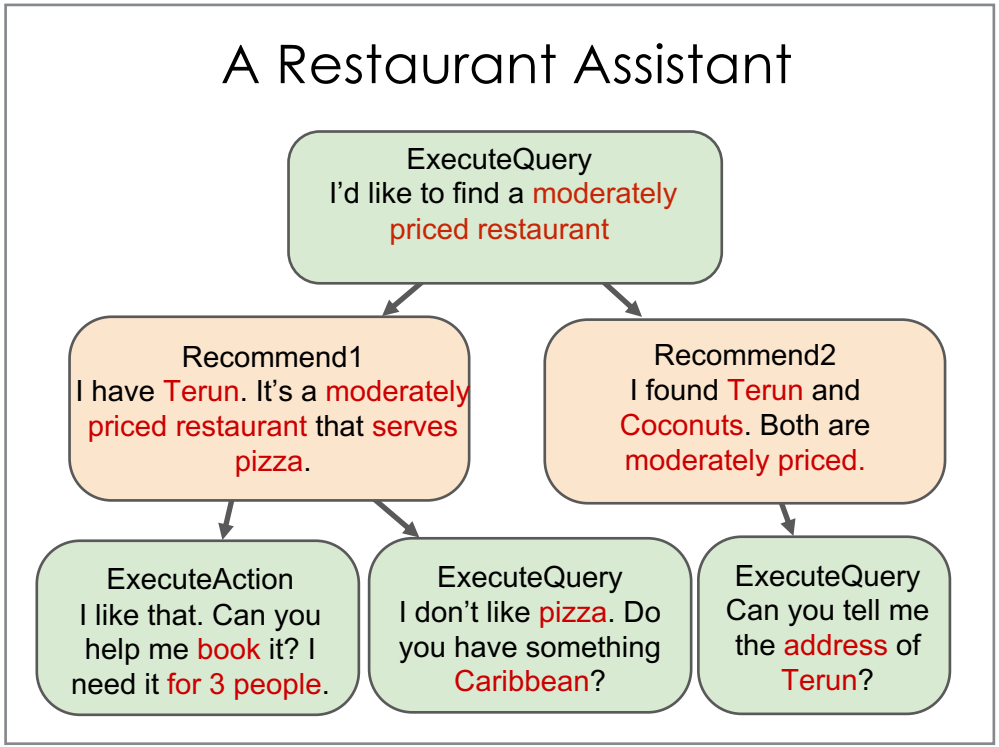
State-Machine+ Domain Info → Domain Chat

Restaurant

DB schema and API signature

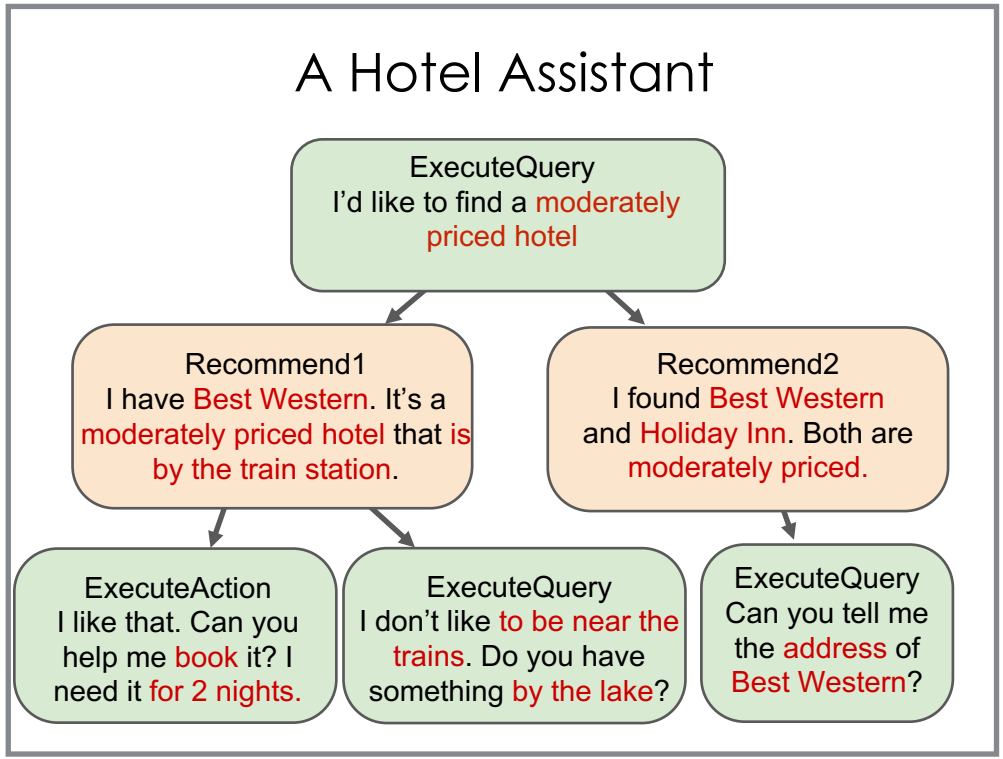
```
Restaurant
id:          Entity(Restaurant)
geo:         Location
price:       Enum(cheap, moderate, expensive)
cuisines:    Array(Entity(Cuisine))
```

```
MakeReservation
restaurant:    Entity(Restaurant)
book_people:   Number [min=1]
book_day:      Date
book_time:     Time
```



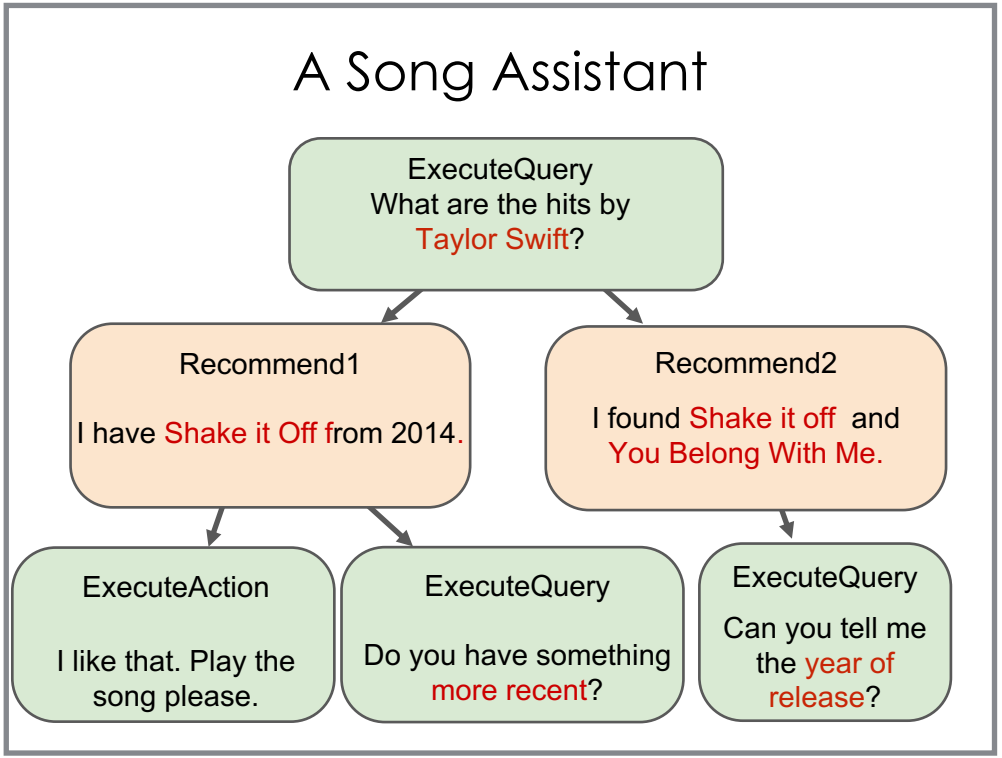
State-Machine+ Domain Info → Domain Chat

Hotel Info



State-Machine+ Domain Info → Domain Chat

Song Info



Evaluation of our Dialogue Modeling

- For MultiWOZ H2H data set
  - **How good is the representation?**
    - 10 User Acts, 20 Agent acts
    - ThingTalk query language
  - **How good is the state machine?**
    - 20 transitions from agent to user acts
    - 43 transitions from user acts to agent acts

Expressiveness of Dialogue Act+ThingTalk

- **97.7% of the validation set is representable**
  - 99.8% of the user utterances
  - 97.6% of the agent utterances

- **What cannot be represented?**
  - User: (Treated as errors in DST experiment)
    - Questions that can't be answered using the given database
    - Out-of-domain questions
    - Note: random chitchat is fine (but ignored)
  - Agent: (Marked as 'invalid' in DST experiment)
    - Asking the user to wait

Speech Act Theory

- The crowdsource workers are not aware of dialogue acts
- Good representation
  - 99.8% of user utterances represented with only 10 acts
  - 97.6% of agent utterances represented with only 20 acts
- Confirms the Speech Act Theory"
  for WOZ transactional dialogues

Gap Between State-Machine & Real Life

U: Please book a table for 5 at 14:30 on Wednesday at Royal Spice.
I also need to find a place to stay

Two user dialogue acts in different domains

A: I was able to book your table.
Your reference number is kqmxil0z.
Now, what kind of hotels are you looking for?

Two agent dialogue acts in different domains

- Utterances are representable with 2 dialogue acts
- Not modeled in state machine
  - Multi-domain turns
  - Domain switches
  - Abandoned transactions
- State Machine will blow up in size to cover all the transitions
- Gap = 14%

What Can We Do with the State Machine?

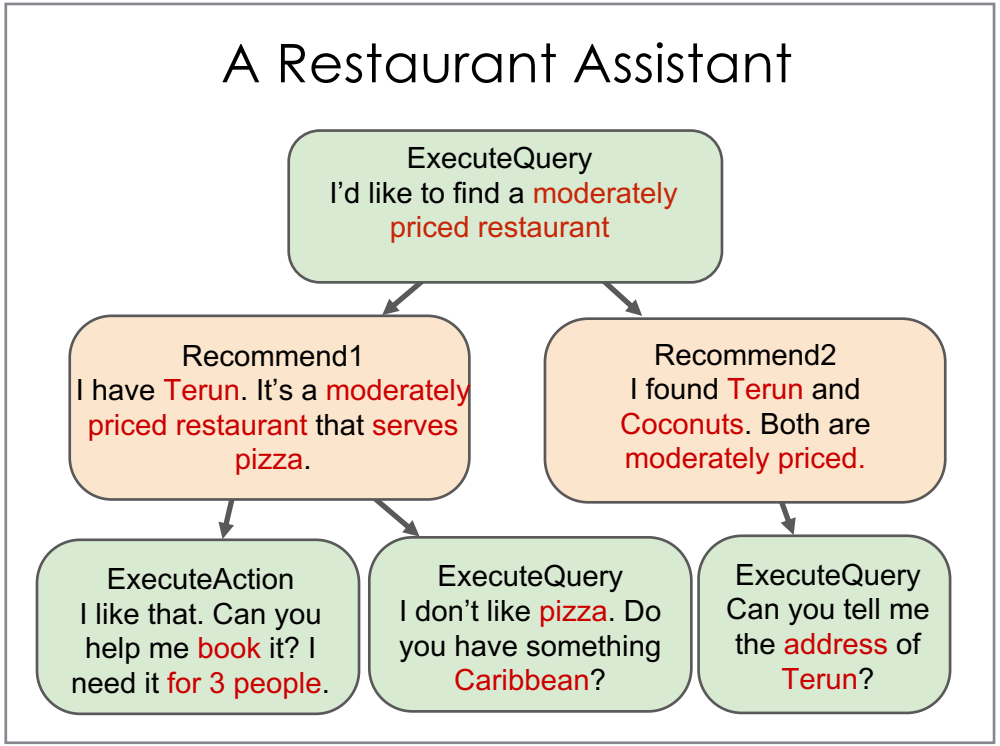
- Debugging the agent: Generate test data
- Augment training data with synthesized data
  - State machine: Similar to the SGD paper
  - Sentence-level: grammar based + language variety
- DO NOT use synthesized data for evaluation or test

Test Accuracy

Training Set	**Exact match**	**Slots only**
Few-shot only	73.7%	81.6%
Genie	**79.2%**	**87.5%**

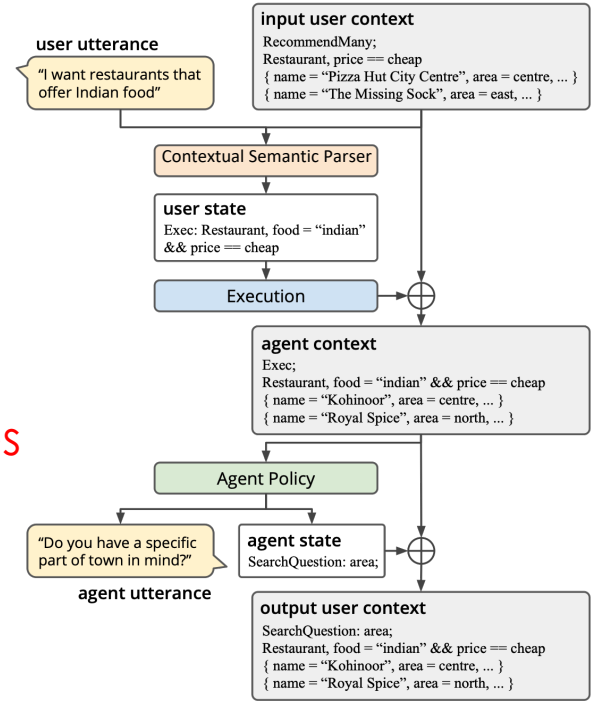
Use the State Machine as Rule-Based Agent Policy?

- Agent state
  = transition (User state, Result)
- Agent Policy:
  All possible transition rules
- Quiz: Is this a good idea?



Must Handle Everything Representable

- 14% gap between state machine and representable dialogues
- One agency policy for all domains
  - Must handle all representable dialogue states
  - E.g. multiple domain requests in 1 sentence, abandoned transitions etc …

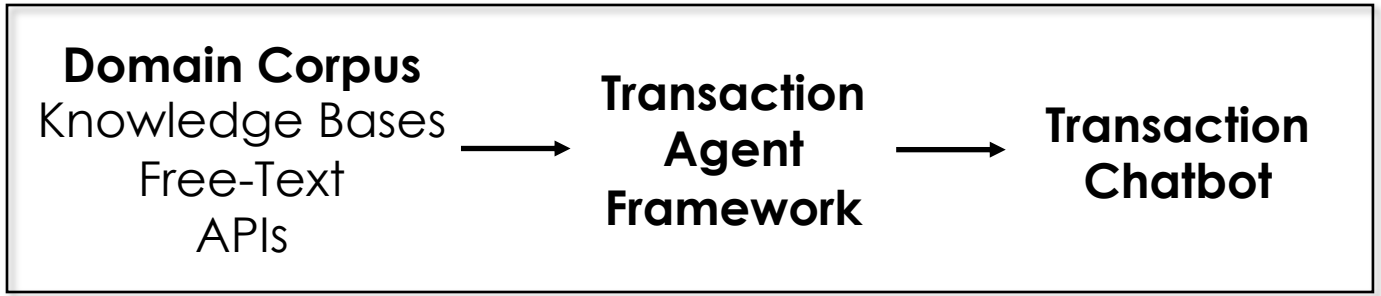


Neural vs. Rule-Based Agent Policy

	Neural Agent Policy	**Rule-Based Policy**
WOZ dataset acquisition	Full training (expensive)	Few-shot to derive a formal representation
Ease of coding	✓	×
Control	×	✓
Interpretability	×	✓
Ease of Updates	×	✓

How to Build a Transaction Agent

- Just supply the domain corpus



How to Build Other Conversational Agents

- **Design and refine the formal dialogue representation**
  - Get a few-shot data set, possibly synthesized by GPT
  - Label them to design the user and agent dialogue acts
  - Design the entire formal dialogue state
- **Implement the first agent**
  - Create a semantic parser with an LLM with in-context learning
  - Write the agent policy for the full formal state representation
- **Test and refine your agent**
  - With simulated users by GPT
  - With real users

Next Lecture: A proposal to simplify agent development

Conclusion

- Dialogue state tracking
  - Use the formal state for dialogue context
  - LLMs can handle contextual semantic parsing for simple schemas
  - Be wary of old manually annotated datasets
    - Even the best has errors: gold annotation, conventions
    - Useful for evaluation: manually check only 20%
    - Synthesis useful for larger knowledge bases
- Agent policy
  - Neural policies requires annotated data, less interpretable
  - Rule-based policies should be written for all possible formal states