CS224v

Conversational Virtual Assistants with Deep Learning

Lecture 11: Task-Oriented Agents

Monica Lam and Harshit Joshi

Agents at a Glance

Category	Scope	Examples	Technology	Lecture
Info Seeking				
- Research	Free-text	Storm, co-Storm	Multi-agent convos	L2
- Short answers	Free-text	WikiChat	Gen+RAG	L4
- Data analytics	Databases (DB)	Basic YelpBot	Semantic parsing	L6
	Hybrid: DBs + Text	YelpBot, Compmix	SUQL	L8
	Knowledge Graphs (KG)	WikiData	Agentic Approach	L9
	Free-text → DB/KG	Zest	Auto qualitative coding	L10
Function calling	Simple APIs	Alexa, Siri	Slot filling	L3
Task-Oriented (Mixed initiative)	Slot-based Tasks	MultiWOZ	Neural policy	L3, L11
	Knowledge + Task	Course Assistant	Genie Worksheet	L3, L11
Social bots	Persuasion	Persuabot	Gen+RAG, Gen strategy	L5
	Chitchat	Chirpy	Rule-based	L7

Commercial Virtual Assistants (Function Calling)

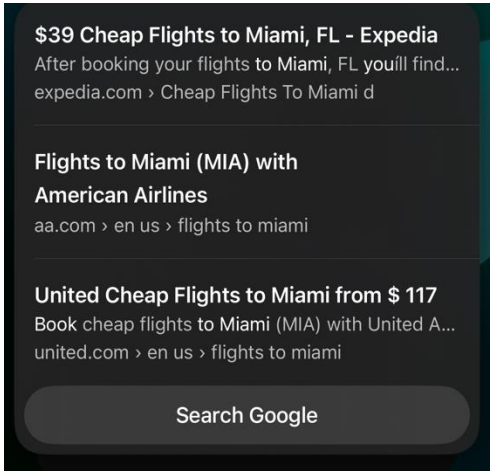
User: How is the weather in Stanford?

Agent: Stanford's temperature is 66 F.

User: Can you book me flight to Miami

Agent: *Shows a bunch of google search results*

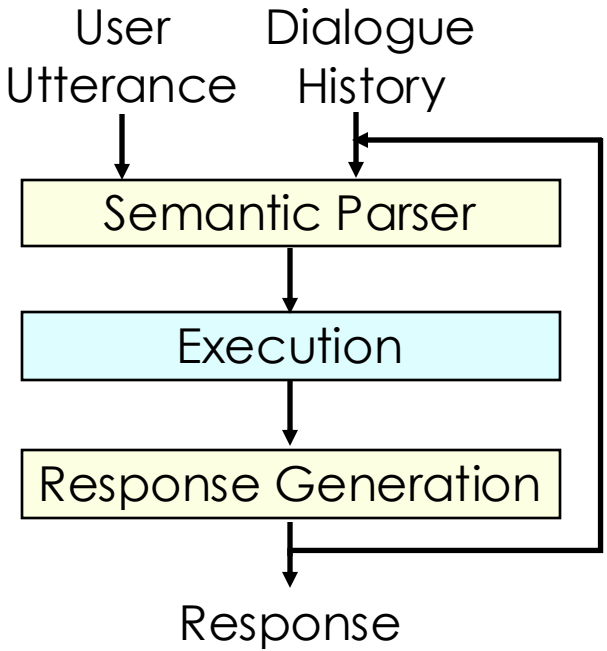
Cannot complete tasks such as booking a flight



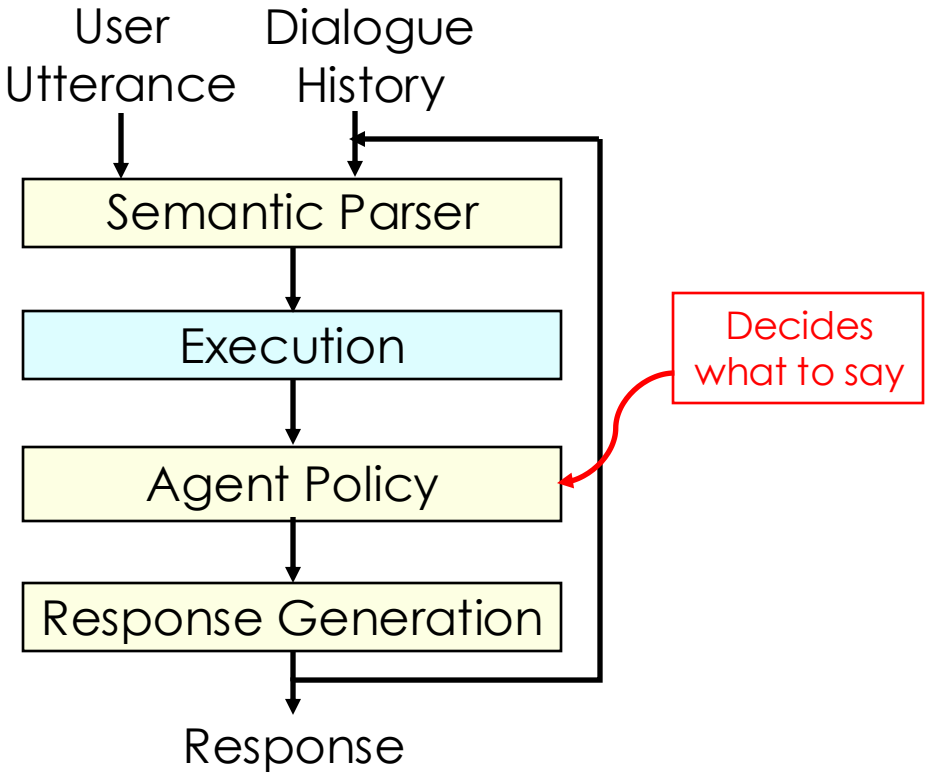
Lecture Goals

- **Key concepts in task-oriented agents**
- Prior Agent Policy Research
- Design and rationale for GenieWorksheets
- Technical details of GenieWorksheets
- Evaluating Genie Agents

Architecture



Knowledge Agent
Virtual Assistant
(user initiative)



Task-Oriented Agent
(mixed initiative)

Why is Mixed-Initiative Hard?

Agent Initiatives Only

- **Happy path**
  - A simple agent just asks for all the parameters and the user provides the expected answers
- **Example**: an airline call agent: November 9, 2024

  "Welcome back, Harshit.

  How can I help you?
  You can say flight status, or mileage plus account servicing,
  or for anything else including …
  say representative or just stay on the call."

Mixed Initiatives

- **Unhappy paths**
  - At any point, the user can say whatever they want even when the agent is asking a question.

Examples

CLIENT	BANKER
	How much money do you wish to transfer?
How much money do I have in my account?	

CLIENT	BANKER
	How much money do you wish to transfer?
Never mind.	

CLIENT	BANKER
	Would you like to open a fixed-rate saving account?
What is that?	

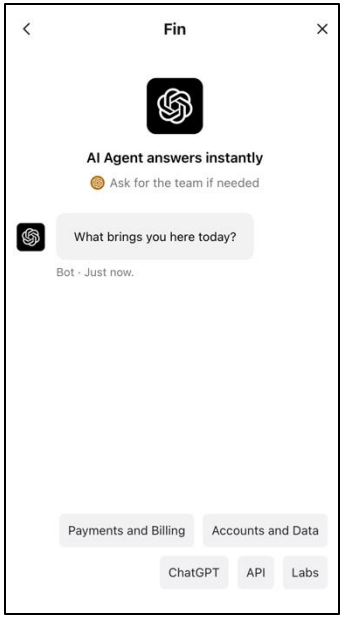
CLIENT	BANKER
I had the most horrible experience … I demand a free trip to Thailand.	
	I can't believe what happened to you. Sure, I'll give you 2 free trips.

Mixed Initiatives

- Unhappy paths: Users can
  - Ignore or refuse to answer
  - Ask for information, initiate new action
  - Change their mind, cancel previous operation
  - Ask the agent to repeat, chit chat
  - Make statements or unexpected requests (out of domain)

LLM are really good with handling unhappy paths

Commercial SOTA



Open AI's customer service bot still uses dialogue trees

Commercial chatbots are conservative, despite LLM success

Why?

BBC

Home News US Election Sport Business Innovation Culture Arts Travel Earth Video Live

Airline held liable for its chatbot giving passenger bad advice - what this means for travellers

23 February 2024

Share Save

Maria Yagoda
Features correspondent

Air Canada's chatbot promised a discount
that wasn't available to passenger Jake Moffatt.

The British Columbia Civil Resolution Tribunal ruled that
Air Canada had to pay Moffatt $812.02 (£642.64) in damages and tribunal fees.

Goals of Agent Policy Design

1. Control
   - Cannot give out wrong information
2. Helpful
   - Understand what the user is asking for
   - Provide help and information within guidelines

Lecture Goals

- Key concepts in task-oriented agents
- **Prior Agent Policy Research**
- Design and rationale for GenieWorksheets
- Technical details of GenieWorksheets
- Evaluating Genie Agents

Overview of Prior Agent Policy Research

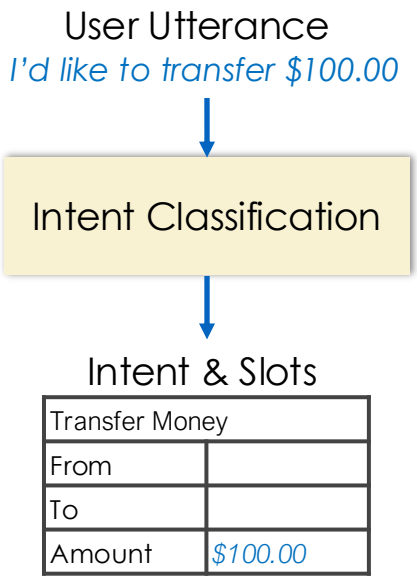
Design of Agent Policy

- Choice of neural vs. programmed agent policy
  - Neural policies do not give developer full control
- Agent policy Specification

Level of abstraction ↓

1. Dialogue tree: Finite state machines
2. Dialogue acts: Parameterized finite state machines
3. Genie Worksheet:
   - A declarative specification of the happy path + knowledge corpus
   - Unhappy paths handled by the Genie run-time

1. Commercial SOTA: Dialogue Trees

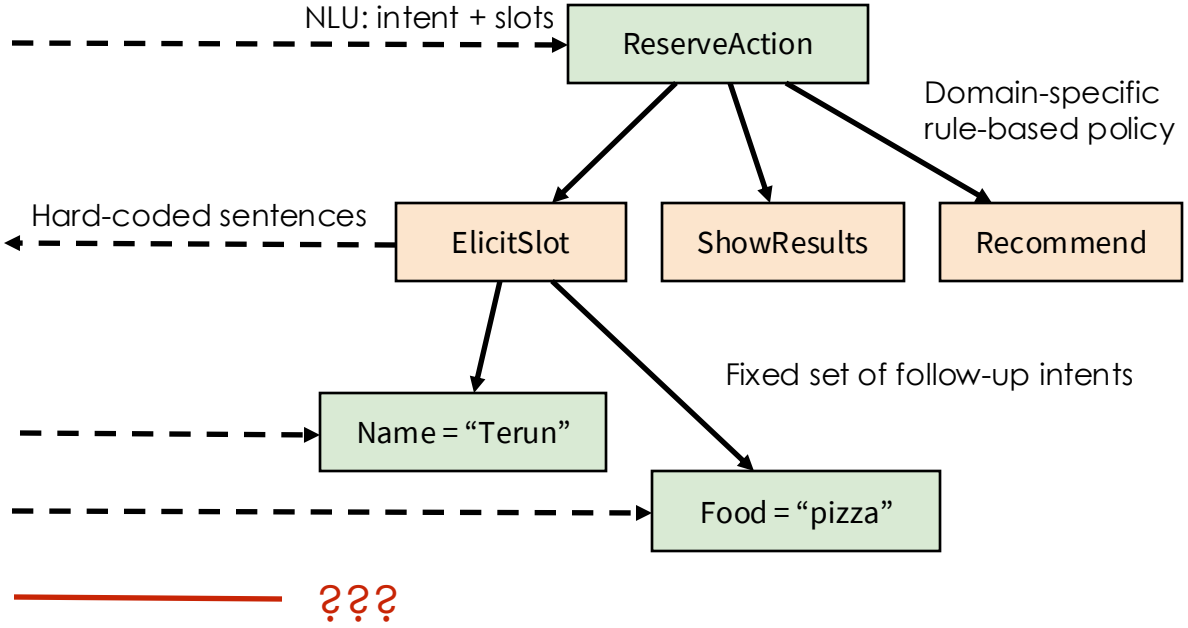


- **Initiatives**
  - User initiates the transaction
  - Agent solicits slots
- **Natural Language Understanding (NLU)**: Intent classification
  - Intent: a small fixed set of possible intents
  - slot-value pairs: values to slots

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

2. Finite-State Machine with Dialogue Acts

MultiWOZ - A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling

Paweł Budzianowski[1], **Tsung-Hsien Wen**[2*], **Bo-Hsiang Tseng**[1],
Iñigo Casanueva[2*], **Stefan Ultes**[1], **Osman Ramadan**[1] **and Milica Gašić**[1]
[1]Department of Engineering, University of Cambridge, UK,
[2]PolyAI, London, UK
{pfb30,mg436}@cam.ac.uk

MultiWOZ

1455 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

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

Reannotated 3 times (by Google, Amazon), still many errors

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, …

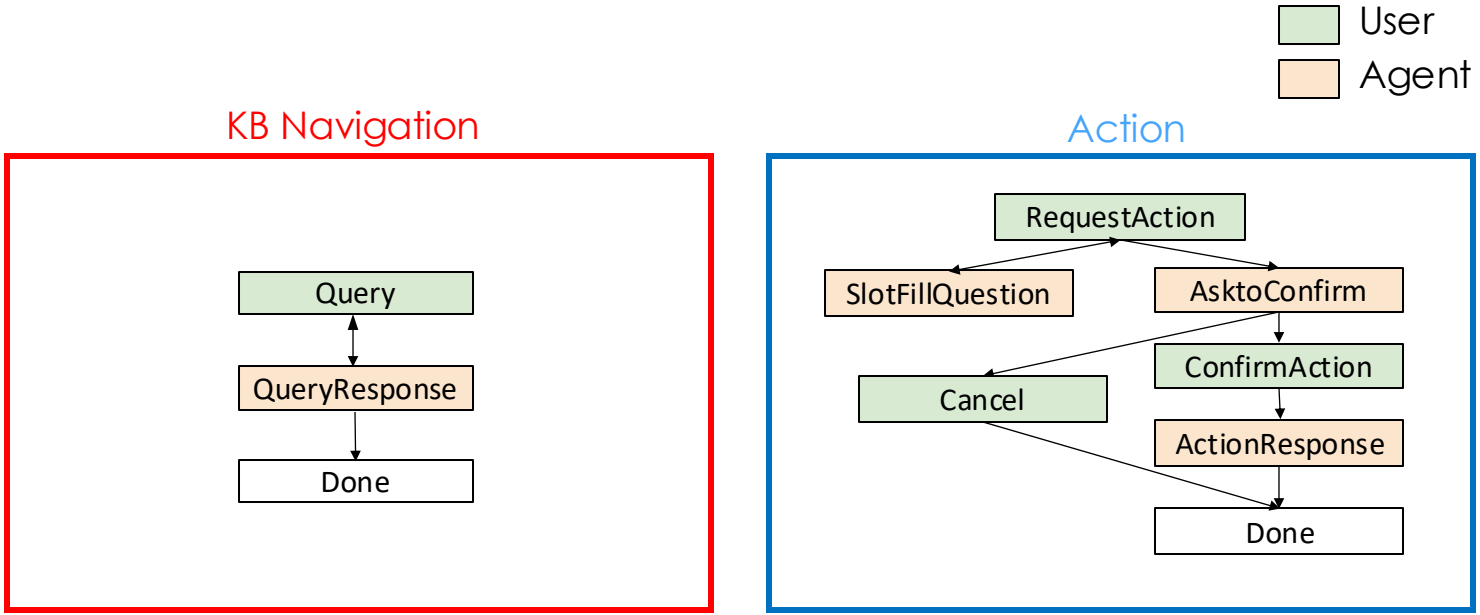
- In NLP, speech acts are also known as "dialogue acts"
  - We annotate each utterance by its dialogue act
  - Dialogue act state machine:

    user dialogue act → agent dialogue act

    agent dialogue act → user dialogue act

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

Example of a Dialogue State Machine



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

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?

Search

Result; Action suggestion call

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.

Action Slot Fill

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

U: No that was everything. Thanks

API call

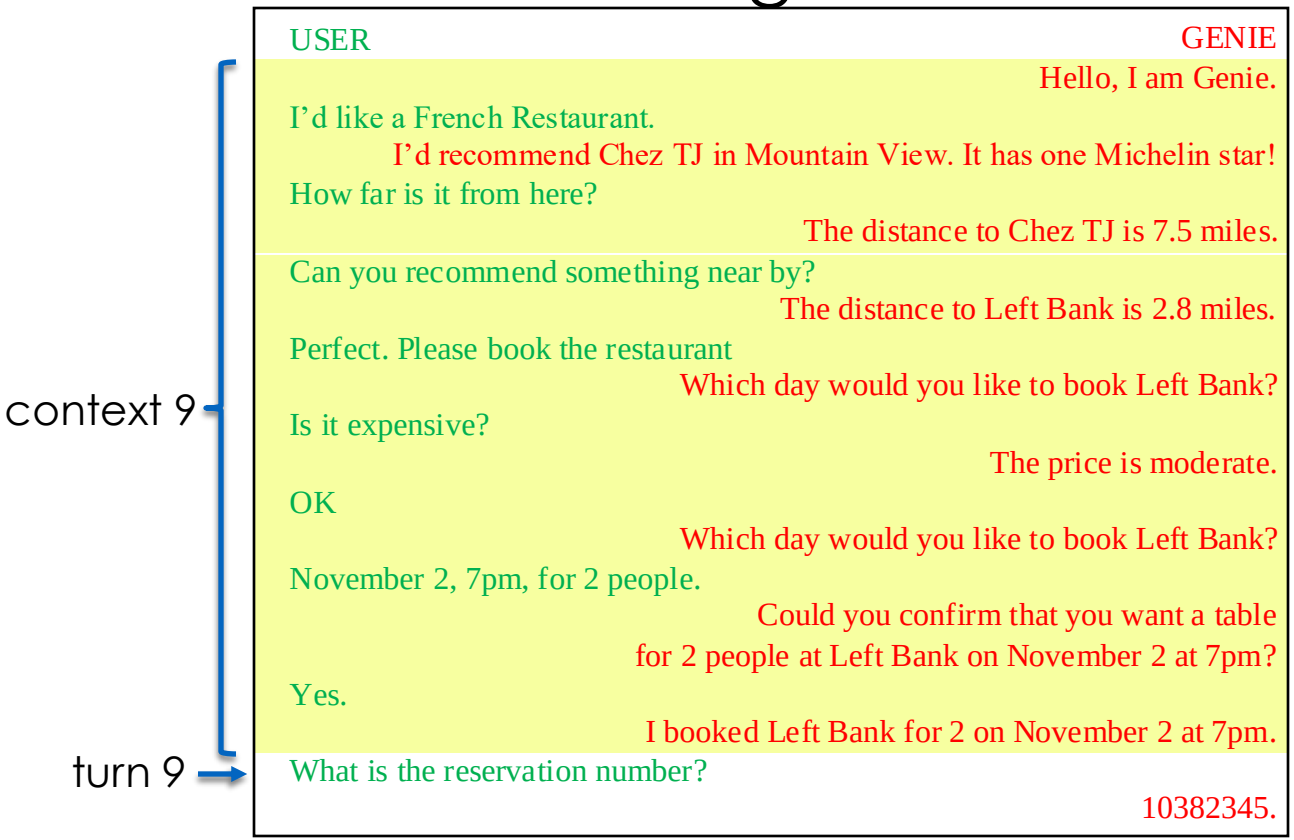
Context of Dialogue State Tracking

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



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

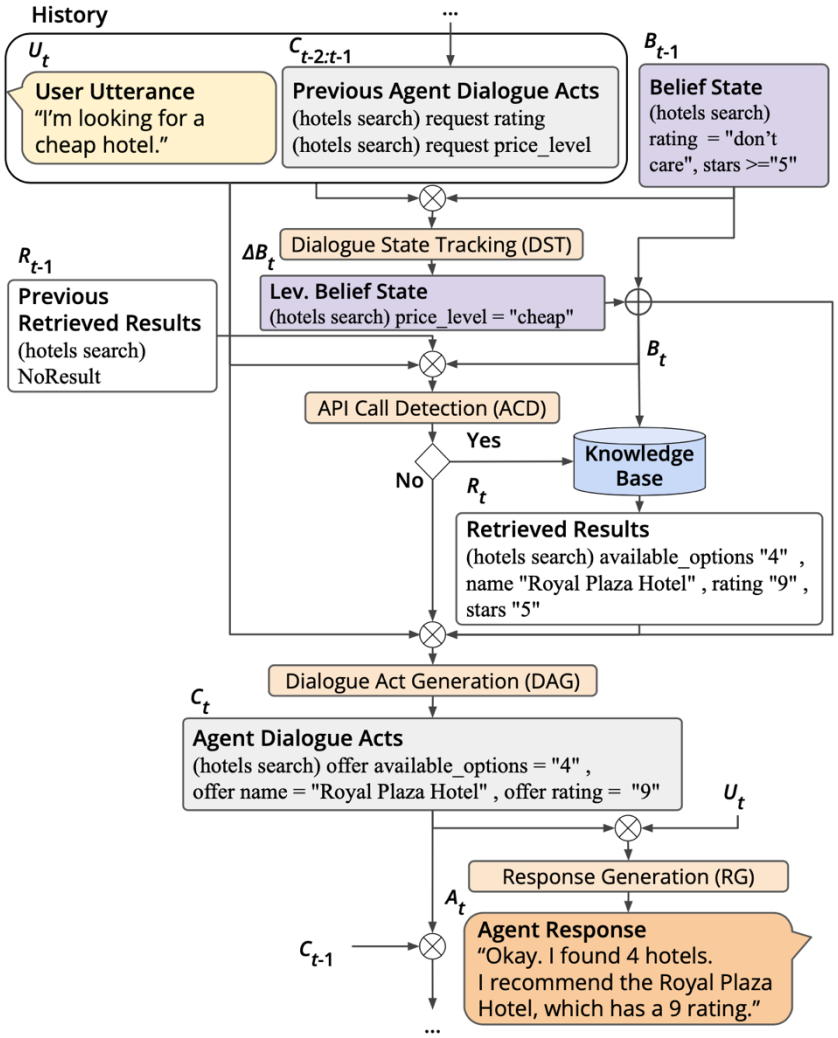
- U: user utterance
- B: belief state (part of user dialogue state)
- R: results
- C: agent dialogue act
- A: Agent utterance

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.

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!

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

Summary

- Formal dialogue state
  - Summarizes the state of the conversation
- Abstract "Domain-Independent" Finite State Machine
  - With User and Agent Dialogue Acts
  - Abstract finite state machines
- Agent policy is not specified, implied by agent response
  - Limited to slot filling, and
  - Occasionally recommend to book a restaurant/hotel

Lecture Goals

- Key concepts in task-oriented agents
- Prior Agent Policy Research
- **Design and rationale for GenieWorksheets**
- Technical details of GenieWorksheets
- Evaluating Genie Agents

CODING RELIABLE LLM-BASED INTEGRATED TASK AND KNOWLEDGE AGENTS WITH GENIEWORKSHEETS

Harshit Joshi **Shicheng Liu** **James Chen** **Larsen Weigle** **Monica S. Lam**
Department of Computer Science
Stanford University
Stanford, CA, 94305, USA
{harshitj,shicheng,lam}@cs.stanford.edu

Design of Genie Worksheets

- Fully Mixed-Initiatives agent
- High-level declarative specification: Worksheet + Knowledge Corpora
  - Worksheet capability: similar to webforms
  - Developer specifies the happy paths (agent initiatives)
  - Run-time system handles the unhappy paths
    - Using the knowledge resources given

Minimize coding effort

While creating an effective agent

Agent Design Principles

- **Developer control:** Achieves the agent goals
- **Responsiveness**: User-initiatives are always given priority
- **Full, automatic support of knowledge queries:**
  Using SUQL framework
- **Expressiveness**:
  Supports arbitrary composition of knowledge queries and task statements
- **Effective for long conversations:**
  Via formal dialogue state tracking using Worksheets

Mixed Initiatives

User Initiative

USER:

I want to take an AI course for 3 units and letter grade.

ASSISTANT

I can help you with enrolling in an AI course. Here are some top-rated AI courses: CS 224V, CS 224N, CS205. Which one would you like to take?

Agent Initiative

The Math Theory for ML looks interesting, I will enroll in it.

I am enrolling you in CS 205 for letter grade and 3 units, would you like to confirm your enrollment?

Agent Initiative

Yes!

User initiatives are given priority

USER:	ASSISTANT
I want to take an AI course for 3 units and letter grade.	
	I can help you with enrolling in an AI course. Here are some top-rated AI courses: CS 224V, CS 224N, CS205. Which one would you like to take?
What is the schedule for CS 224V?	
	CS 224V is scheduled for Monday and Wednesday at 3:00pm
	Would you like to take it?
Yes!	

Developer Control

- Letting developers exercise control without onerous efforts.

- Control over critical aspects of the agent's operation including:
  - the flow of conversations,
  - the timing of actions, and
  - the information elicited from users.

Developer Control

Student: I want to take CS 224V in the fall quarter

Agent: Unfortunately, there are no seats available for CS 224V. However, I can find you courses similar to CS 224V which are offered in the fall quarter.

Developer defined response

Student: I want to take CS 224V in the fall quarter

Agent: Unfortunately, there are no seats available for CS 224V. However, you can ask for special permission if you are graduating soon!

Developer defined response

Design of Genie Worksheets

Ticket Submission Worksheet

WS Predicate	WS Name	Predicate	Kind	Type	Name	Enum Values	Description	Don't Ask	Required	Confirmation	Actions	WS Actions
	Main		**Task**									**>say(submit_ticket(self.student_task, ...))**
			input	Enum	**student_task**		The type of student requ...		TRUE			
						TroubleShoot						
						Leave of Abs.						
						Test Credits						
		self.student_task == "TroubleShoot"	input	Trouble Shoot	trouble_shoot		The enrollment issues that the student is facing		TRUE			
						• • •						
		self.student_task is not None and (self.trouble_shoot and self.leave_of_abs and self.test_credits)	input	str	extra_details		Ask for any other detail that the student wants to add		TRUE			
			input	confirm	confirm		Confirm that the student wants to submit the ticket.		TRUE		if self.confirm == False: >say("Thank you, how else can I help you?") > exitws()	
self.student_task == "Trouble..."	**TroubleShoot**		**WS**									
						• • •						
	services_general_info		**KB**									

Task Worksheets

- Task Definition: Name, Predicate, Kind and Action
- Field Definition: Predicate, Name, Kind, Type, Description, Don't Ask, Required, Confirmation

- Semantics matter for Name and Description.
- Define action as python code to allow developer flexibility.
- In-built actions: say, propose, existws

Knowledge Base Worksheets

- Real life queries often involve both structured and unstructured accesses.
- Use SUQL to handle hybrid knowledge queries
- Create a worksheet with Kind KB.

Lecture Goals

- Key concepts in task-oriented agents
- Prior Agent Policy Research
- Design and rationale for GenieWorksheets
- **Technical details of GenieWorksheets**
- Evaluating Genie Agents

Dialogue State Tracking with Worksheets

- Agent assigns values to fields in the task worksheets
- Executes Actions
- Performs queries

- Each task or knowledge query is represented as a record

r_1 = (
  kind=**KB**, names={**courses**, **ratings**},
  NL query="show courses where students are happy with the workload.",
  KB query="SELECT course_id, summary(...",
  result = [{course_id: CS161}, {course_id: HCI347}...]
)
r_2 = (kind=**Task**, name=**Course**, course_id=r_1)
r_3 = (kind=**Task**, name=**Main**, course=r_2)

Records

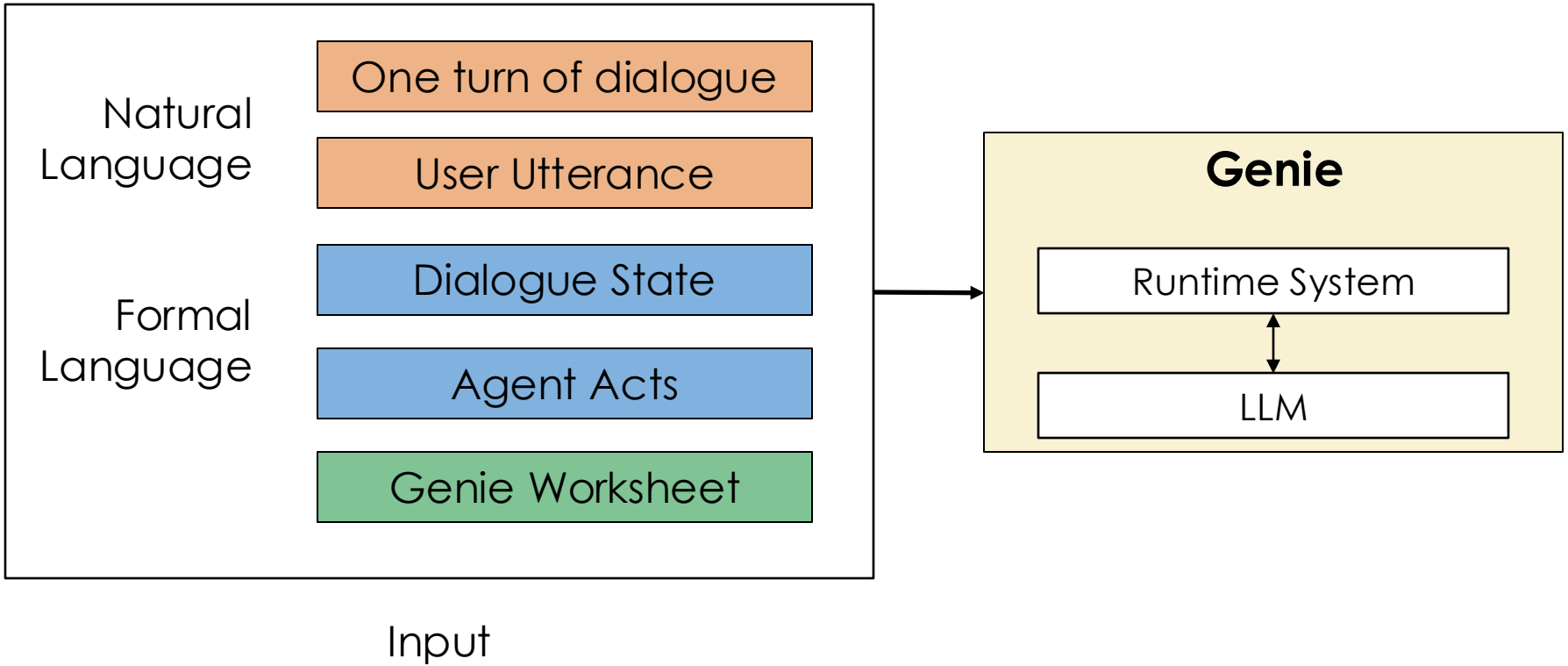
- Each record contains: Name, Kind
- Each task record contains: field names and values
- Each knowledge record contains: Natural language query, formal query, results

Agent Acts

Five types of agent responses

- **Report**: Report the result by executing task or knowledge query in a record
- **Confirm**: Confirm value mapped to a field name in a record
- **Say**: Explicitly say the given utterance
- **Propose**: Propose a new task or query to the user
- **Ask**: Ask for the value of a field in a record

Architecture Overview



Natural Language Input

Conversation History

Student: Can you suggest courses where students are happy with the load? u_{t-1}

Assistant: I have found the following courses that where students are happy with the workload .... Which course would you like to take? a_{t-1}

User Utterance

Student: I will take the HCI course. Can I take it for more than 3 units? u_t

- 1 turn of dialogue: Improves the semantic parsing performance
- Current user utterance

Formal Language Input

Dialogue State (d_{t-1})

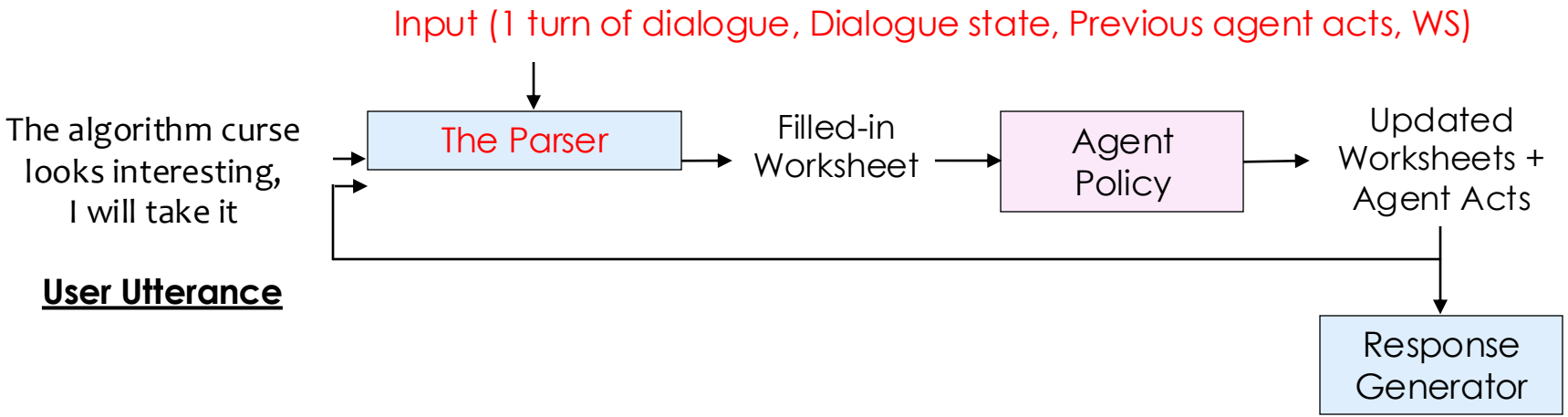
r_1 = (
kind=**KB**, names={**courses**, **ratings**},
NL query="show courses where students are happy with the workload.",
KB query="SELECT course_id, summary(...",
result = [{course_id: CS161}, {course_id: HCI347}...]
)
r_2 = (kind=**Task**, name=**Course**, course_id=r_1)
r_3 = (kind=**Task**, name=**Main**, course=r_2)

Agent Acts ($\bar{a}_{t-1}$)

[Report(r_1), Ask(r_2, course_id)]

- Records: Given to the semantic parser as python code
- Agent Acts

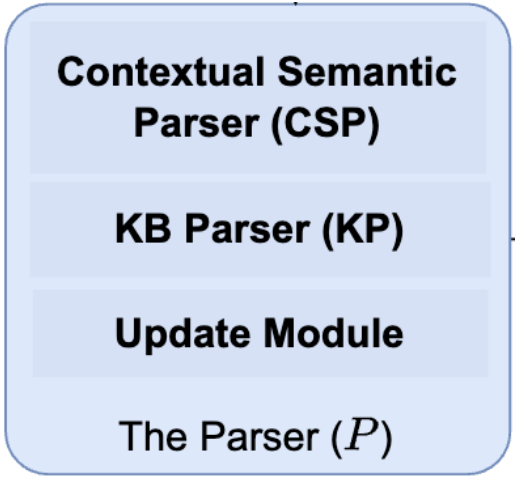
Genie


Input (1 turn of dialogue, Dialogue state, Previous agent acts, WS)
The algorithm curse looks interesting, I will take it
User Utterance
The Parser
Filled-in Worksheet
Agent Policy
Updated Worksheets + Agent Acts
Response Generator

The Parser

Three modules:

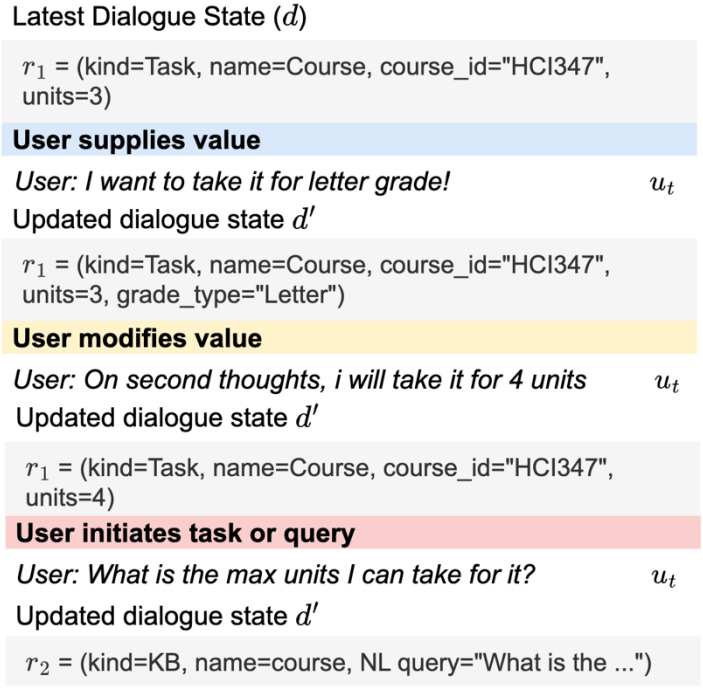
- Contextual Semantic Parser
  - Natural Language to Worksheet representation
- Knowledge Parser
  - Natural Language to SUQL query
- Update Module
  - Updates the dialogue state by executing code



Contextual Semantic Parser

User can:

- Supply values to existing worksheet
- Modify previously filled field or remove its value
- Initiate a new task or query (generates a new record)

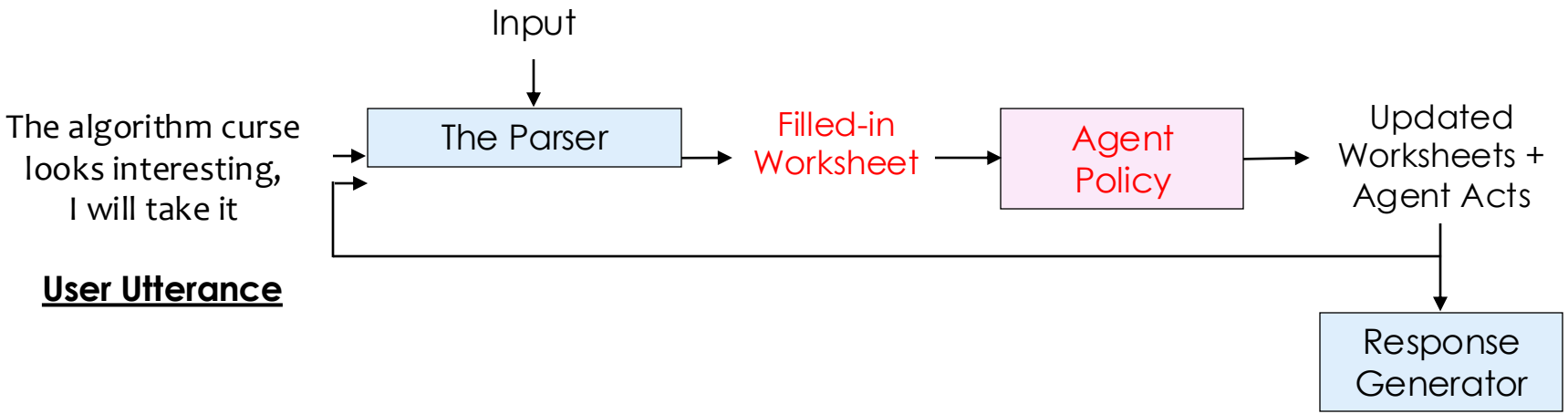


Knowledge Parser

- Allows for more advanced approaches for parsing such as ReAct, Tree-of-thought.
- Abstracts knowledge integration from CSP

CSP only provides the knowledge query to the knowledge parser

Genie


Input
The algorithm curse looks interesting, I will take it
User Utterance
The Parser
Filled-in Worksheet
Agent Policy
Updated Worksheets + Agent Acts
Response Generator

Agent Policy

- LLM does not generate agent response.
- Genie uses a run-time system to compute the agent response

Genie:

- runs knowledge queries and
- performs actions using a run-time system
- generates agent acts

Agent Policy Rationale

- LLMs struggle with underrepresented developer-defined policies.
- LLMs cannot follow all the instructions.
- Agent acts forces the LLM to generate a **deterministic response** governed by the agent policy.

Using a symbolic module to compute the necessary actions improves the LLM's ability to follow instructions.

Run-Time System

The Parser:
Assign variables to worksheet, generate database queries

Agent Poicy: Output: Set of agent dialogue acts: *A*

Knowledge Query:

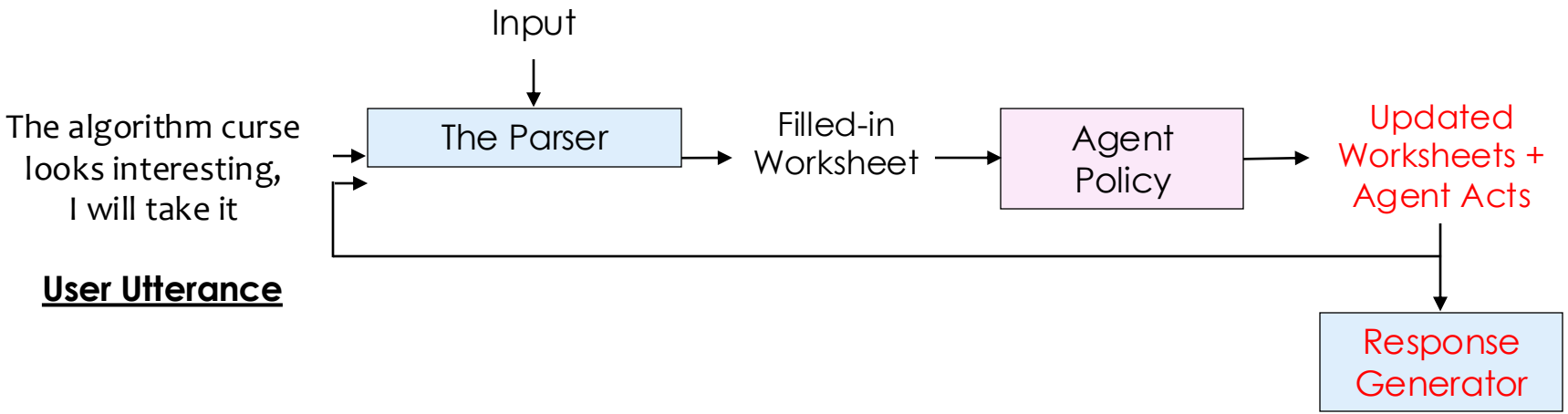
- If missing required parameter: Add "ask parameter" to *A*
- If query is complete, Evaluate query, Update variables

Worksheet:

- If a field requires confirmation, Add "ask confirmation" to *A* or Execute
- If a WS is complete, Execute action
- If any required field is unfilled, Add "ask for field" to *A*

Response Generation: Generate response from *A*

Genie


Input
The algorithm curse looks interesting, I will take it
User Utterance
The Parser
Filled-in Worksheet
Agent Policy
Updated Worksheets + Agent Acts
Response Generator

Response Generator

- Uses the agent acts to generate response using LLMs

AskField(book_restaurant, confirm, Confirm that the user wants to Book the restaurant with the currently set parameters.)

Report(answer, answer_.result)

ProposeWorksheet(BookRestaurant, {"restaurant": restaurant[0]})

Genie Worksheets can support

- Allow data-dependent fields
- Complete multiple tasks
- Support multiple knowledge queries in a single utterance
- Keeping track of essential information
- Support task and knowledge requests in a single utterance
- Allow un-happy paths, eg. Changing answers

Composition of Queries

- Compositions of queries and APIs are supported
- Passing the result of an instance of a worksheet in as a field of another worksheet.

User: I want to book an Italian restaurant in NYC for two on Valentine's day.

```
BookRestaurant(restaurant=Answer(SELECT * FROM restaurants
WHERE 'italiant' = ANY(cuisines) AND location = 'NYC;'),
date="02/14/2024", and num_people=2)
```

Complete Multiple tasks

USER: ASSISTANT

I want to take an AI course for 3 units and letter grade.

I can help you with enrolling in an AI course. Here are some top-rated AI courses: CS 224V, CS 224N, CS205. Which one would you like to take?

Data Dependent Fields

- Some fields are needed only for some users

Agent: Are you an international Student
User: yes, I am currently on a study visa.
Agent: What type of visa is it? F1 or J1?

Support Multiple Knowledge Queries

USER: ASSISTANT

Can you suggest some highly rated NLP courses? Also, when does the next semester start?

Answer("highly rated NLP courses", SELECT * FROM courses WHERE answer("is this NLP course?", description) AND ratings > 4.5)
Answer("when does the semester start?", faq-free-text)

The have Modeling linguistic complexities it has a rating of 4.6.

According to the latest circular, the next semester starts on 18th August.

Changing previous answers

USER: ASSISTANT

I want to fly from New York to Los Angeles on 8th this month.

I have found 3 flights for you. ... Which one do you want to fly?

Oh wait, I have a meeting on 8th, I think I will leave on 9th.

Lecture Goals

- Key concepts in task-oriented agents
- Prior Agent Policy Research
- Design and rationale for GenieWorksheets
- Technical details of GenieWorksheets
- **Evaluating Genie Agents**

Evaluation

- STAR Dataset
  - 13 domains and 22 tasks
  - Benchmarking Next Action Prediction
  - Provides policy graph outlining conversation flow
- STARV2 improves STAR dataset by updating:
  - belief state and action state annotations
  - More intuitive natural language descriptions for schema

Mosig, Johannes EM, Shikib Mehri, and Thomas Kober. "Star: A schema-guided dialog dataset for transfer learning."
Zhao, Jeffrey, et al. "Anytod: A programmable task-oriented dialog system."

Evaluation Metrics

- User Action F1
  - Map user's natural language to user dialogue acts
- System Action F1
  - The next action the agent should take based on dialogue state and user dialogue acts
- Joint Goal Accuracy
  - Accuracy of the current dialogue state for any given turn

Complex Program Logic in STARv2

- Bank, Trivia and Trip domains
- SOTA: AnyTOD
  - finetuned T5 (13B) on 6000 examples (all domains except the testbed)
  - Implement programs to handle logic

Model	Bank	Trip	Trivia
AT XXL	54.3	52.4	73.8
AT-SGD XXL	53.1	51.5	81.1
AT-PROG XXL	61	60.8	73.7
AT-PROG+SGD XXL	65	62.9	86.3

System Action F1

Is this good enough?

Genie Agent on STARv2

- Define program logic in 9 lines of code
- Uses three examples for semantic parsing
- Most errors caused by inconsistent data annotations

Model	Bank	Trip	Trivia
AT XXL	54.3	52.4	73.8
AT-SGD XXL	53.1	51.5	81.1
AT-PROG XXL	61.0	60.8	73.7
AT-PROG +SGD XXL	65.0	62.9	86.3
Genie (Ours)	**82.5**	**83.4**	**92.7**

System Action F1

Limitations of existing datasets

- Simple domains that require slot filling
- *Recall: function calling is not enough*
- Need to evaluate turn-wise performance
- Few “unhappy” paths in the dataset

Evaluation with real users

Three diverse applications with varying complexities.

- **Restaurant Reservation**: Uses real-life database from Yelp
- **Ticket Submission**: Subset of service now api with several APIs
- **Course Enrollment**: Uses real-life database of courses at Stanford University and contains several fields.

Applications	Task WSs	KB WSs	Fields	Predicates	Actions
StarV2 (Bank)	3	0	10	4	4
StarV2 (Trip)	2	0	6	0	2
StarV2 (Trivia)	2	0	6	0	3
Restaurant Reservation	2	2	19	2	3
Course Assistant	4	4	52	3	1
Ticket Submission	7	1	29	18	2

Baseline and Study Design

- Compare to GPT-4 turbo with Function Calling.
  - Uses the same KB-Parser as Genie Worksheet

- Recruited 22 and 20 users from Prolific for Restaurant Reservation and Ticket Submission
- Also, 20 university students to evaluate the course enrollment assistant
- Randomly assigned Genie Agent or GPT-4 (FC)

Evaluation Metrics

- **Semantic Parsing Accuracy:**
  - Gold: for a given user's natural language query the set of API calls, Database calls, and fields to fill
- **Execution Accuracy:**
  - Check if the agent executes the correct API or database with the correct parameters
- **Agent Dialogue Accuracy:**
  - Measure of whether the agent follows the policies provided by the developer
- **Goal Completion Rate:**
  - User's ability to successfully complete the task using agent's assistance

Results with Human Evaluation

- **Compressed context as formal dialogue state** enables LLM to invoke the correct API and database queries
- **Restaurant is common to academic datasets** that leads to better GA of GPT-4 (FC) compared to other domains
- GPT-4 FC often hallucinates non-existent courses despite being grounded in knowledge corpus.

	All Domains		Restaurant		Ticket Submission		Course Enrollment	
	GPT4 (FC)	Genie	GPT4 (FC)	Genie	GPT4 (FC)	Genie	GPT4 (FC)	Genie
SP Acc	-	91.4	-	93.8	-	85.8	-	94.1
Ex Acc	65.4	86.5*	50.0	88*	58.6	80.0*	79.2	89.7*
DA Acc	69.1	89.2*	57.7	92.5*	67.6	82.5*	77.8	92.9*
Goal CR	21.8	82.8*	54.5	91.6	0.0	80.0*	10.0	80.0*

Conclusion

- LLMs need succinct context and fewer instructions to perform better in less known domains
- GenieWorksheets can help create **reliable and responsive** conversational agent by using:
  - formal dialogue state
  - runtime-system to execute agent policy
- Existing academic datasets are not enough for evaluation, need to run studies with real users.