CS224v

Conversational Virtual Assistants with Deep Learning

Lecture 1: Introduction

Monica Lam

CA: Paridhi Maheshwari, Jian Vora
My Background

- Director of Stanford Open Virtual Assistant Lab (OVAL)
- Started working on virtual assistants in 2015
- Created the first conversational virtual assistant
  - Uses a contextual neural semantic parser for conversations
- Created the first inter-operating, privacy-protecting assistant
  - Popular Science’s Best of What’s New Award in Security in 2019
- Publications on https://oval.cs.stanford.edu
The Future of Voice

Apple Knowledge Navigator Video (1987)
What New Technology is Needed?
A Companion to the Elderly

- Motivation
  - 27% of individuals over 60 in the US live alone
  - Social isolation $\rightarrow$ depression
    $\rightarrow$ 50% increased risk of dementia
- A companion that has a long-term relationship
  - Help the user with digital tasks
  - Chitchatting about current affairs and personal history
    - Email contains important memory and photos
Future Assistant Capabilities

- Correct answers & actions
- Purposeful conversation
- Chitchat conversation
- Common knowledge
- Human values: Anti-toxicity, anti-biased
- Social intelligence
Thorndike (1920) defined social intelligence as: “the ability to understand and manage men and women, boys and girls -- to act wisely in human relations”

1. Verbal and non-verbal fluency
2. Knowledge of social rules and roles
3. Active listening (listen attentively to a speaker, understand what they're saying, respond and reflect on what's being said, and retain the information for later.)
4. Understanding how other people’s emotions work
5. Playing social roles efficiently
6. Self-image and impression management

https://observatory.tec.mx/edu-news/social-intelligence
Exciting Times for NLP Research

- 2012-2017: Basic ideas of deep learning for NLP
- 2017-2022: Transformers to 1.7T parameter large language models
- 2022-2027: Apple Knowledge Navigator
  Elderly’s companion
  Lots of interesting research opportunities
What You Will Learn in This Course?

• State of the art in conversational agents
  • Lectures: You learn it
  • Homework: You practice it
  • Independent project: You advance it!

• Projects: from basic AI to working conversational assistants

• Research apprenticeship:
  • Many suggested ideas: You can propose ones too
  • Supervision outside of lecture hours
  • Opportunity for research assistantship in Winter 2023

<table>
<thead>
<tr>
<th></th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>15%</td>
</tr>
<tr>
<td>Assignment</td>
<td>25%</td>
</tr>
<tr>
<td>Final Project</td>
<td>60%</td>
</tr>
</tbody>
</table>
Lecture Overview

1. Inspiration: Knowledge navigator
2. Overview on the state of the art
3. Overall architecture inspired by neuroscience
4. Core technical concepts
5. This course
Current Commercial State of the Art

• Conversational chatbots are brittle

Welcome back, Monica.

• Virtual assistants understand single commands, not conversations
1. Task-Oriented Agents

- **Agent initiative:**
  the agent guides the conversation

- **Natural Language Understanding (NLU):**
  Intent classification
  - Intent: a small fixed set of possible intents
  - slot-value pairs: values to slots

User Utterance
I’d like to transfer $100.00

<table>
<thead>
<tr>
<th>Intent &amp; Slots</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transfer Money</td>
</tr>
<tr>
<td>From</td>
</tr>
<tr>
<td>To</td>
</tr>
<tr>
<td>Amount</td>
</tr>
<tr>
<td>$100.00</td>
</tr>
</tbody>
</table>
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 –

NLU: intent + slots
ReserveAction

Domain-specific rule-based policy

ElicitSlot
Name = “Terun”
Food = “pizza”

ShowResults
Recommend

Fixed set of follow-up intents

Hard-coded sentences
Dialogue Trees: Pros & Cons

Pros: Control over conversation

Cons:
1. **Intent & slots: lack expressiveness**
   - What are the top 3 restaurants either near the airport or the Fisherman’s Wharf

2. **Tedious: exponentially many paths (navigation of knowledge bases)**

3. **Developer cannot anticipate all possible utterances**

---

User Initiative

**User**

- I want to transfer money

- 12512032

Agent Initiative

**Agent**

- What account would you like to transfer money to?

- Got it. How much money would you like to transfer?

**We need mixed-initiative agents**
2. Large Language Models

- OpenAI's GPT-3 Generative Pre-Trained Transformers, 2020
  - Dataset: 45 TB of text data on the internet
  - Unsupervised training: to predict the next word
  - 175B parameters
    (Human brain: 86B neurons)
# Large Language Models

<table>
<thead>
<tr>
<th>Year</th>
<th>Model name</th>
<th># parameters (Billion)</th>
<th>Language(s)</th>
<th>Institution</th>
</tr>
</thead>
<tbody>
<tr>
<td>2020</td>
<td>GPT-3</td>
<td>175</td>
<td>English</td>
<td>OpenAI</td>
</tr>
<tr>
<td></td>
<td>Jurassic-1</td>
<td>178</td>
<td>English</td>
<td>AI21 Labs</td>
</tr>
<tr>
<td></td>
<td>Gopher</td>
<td>280</td>
<td>English</td>
<td>DeepMind</td>
</tr>
<tr>
<td></td>
<td>LaMDA</td>
<td>137</td>
<td>English</td>
<td>Google</td>
</tr>
<tr>
<td></td>
<td>Yuan</td>
<td>245</td>
<td>Chinese</td>
<td>Inspur</td>
</tr>
<tr>
<td></td>
<td>PanGu-a</td>
<td>200</td>
<td>Chinese</td>
<td>Huawei</td>
</tr>
<tr>
<td></td>
<td>HyperCLOVA</td>
<td>82</td>
<td>Korean</td>
<td>Naver</td>
</tr>
<tr>
<td></td>
<td>OPT</td>
<td>175</td>
<td>English</td>
<td>Meta</td>
</tr>
<tr>
<td></td>
<td>PaLM</td>
<td>540</td>
<td>English</td>
<td>Google</td>
</tr>
<tr>
<td></td>
<td>BLOOM</td>
<td>176</td>
<td>46 languages</td>
<td>BigScience</td>
</tr>
<tr>
<td></td>
<td>Megatron-Turing NLG</td>
<td>530</td>
<td>English</td>
<td>Microsoft &amp; NVIDIA</td>
</tr>
<tr>
<td></td>
<td>YaLM</td>
<td>100</td>
<td>English, Russian</td>
<td>Yandex</td>
</tr>
<tr>
<td></td>
<td>Wu Dao</td>
<td>1750</td>
<td>Chinese, English, …</td>
<td>Tsinghua</td>
</tr>
</tbody>
</table>

**Note:** The models listed are representative of the advancements in large language models from 2020 to 2022.
Let’s Buy 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.

Can you tell me what day and time you would like to fly?
User: July 7th, 9am
Agent: I can book a flight for you departing at 9am on July 7th.

Would you like me to book that for you?
User: What is the flight number?
Agent: The flight number is UA123.

Fact check:
United UA 123
London to Washington
LLMs: Pros and Cons

• Pros: Plausible responses with fluency and embellishment
  Demonstrate common knowledge, social intelligence

• Cons:
  1. Hallucination, no understanding of semantics, not grounded with external resources
     • Facts from live knowledge bases or APIs
     • Performing actions (e.g. turn off the light)
  2. No control over low-level expressions (including human values)
  3. No control over high-level choices: purpose, long term memory

Cannot use as is!
3. Semantic Parsing

User Utterance
Tell me the top 3 restaurants either near the airport or the Fisherman's Wharf

Neural Semantic Parser

Meaning

• Translate natural language into its meaning (formally represented)
• Sequence-to-sequence transformation
• Used by Alexa for 1st party commands
• AMRL representation is not executable and proprietary
Semantic Parsing $\rightarrow$ Executable Code

- **ThingTalk** is a programming language
- **Complete:**
  - Anything the virtual assistant can do can be represented in ThingTalk
  - Beyond queries (APIs, control flow, etc)
- **Simple:** Execute the code to get the results

Neural Semantic Parser (Codex)

User Utterance

Tell me the top 3 restaurants either near the airport or the Fisherman's Wharf

Meaning

**ThingTalk**

```
sort(desc rating of @com.yelp()
filter location == new Location("airport")
| | location == new Location("fisherman's wharf")) [1:3]
```
Semantic Parsing: Pros & Cons

Pros: Grounding of semantics with external resources

Cons:
1. How to extend to conversations?
   • Only used commercially for single commands and queries
2. How to get training data?
   • There are many paths and variations
   • It is hard to annotate natural language
3. How to combine with common knowledge and common sense
   • E.g. “book me a restaurant on Valentine’s Day”
### Summary

<table>
<thead>
<tr>
<th>Techniques</th>
<th>Knowledge Navigator</th>
<th>Task-oriented Chatbots</th>
<th>Virtual Assistants</th>
<th>Neural Chatbot</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct answers/actions</td>
<td>xxx</td>
<td></td>
<td></td>
<td>LLMs</td>
</tr>
<tr>
<td>Purposeful conversation</td>
<td>xxx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chitchat conversation</td>
<td>xxx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Common knowledge</td>
<td>xxx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Human values</td>
<td>xxx</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Social intelligence</td>
<td>xxx</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

How to get the best of all worlds?
Lecture Overview

1. Inspiration: Knowledge navigator
2. Overview on the state of the art
3. Overall architecture inspired by neuroscience
4. Core technical concepts
5. This course
Let’s Turn to Our Brain for Clues
Neuroscience (for kids)

Wernicke’s Area: Language comprehension
- Damage: can only produce word salad

Broca’s Area: Speech production and articulation
(Speech Center)
- Damage: can understand, cannot produce speech

http://faculty.washington.edu/chudler/functional.html
Neuroscience

Prefrontal Cortex: Problem solving, emotions, complex thoughts
Damage: Perfect understanding, but talks without inhibition / attention
(speaks whatever that comes to mind)
... like the LLMs
Speaking Without Attention/Inhibition

• Your experience?

• Implication for conversational assistants?
Prefrontal Cortex (PFC)

- Executive functions
  - Attention and inhibitory control
  - Monitoring internal and external states
  - Planning and use of long-term memory
  - Flexible and adaptive thinking, including multi-tasking
- PFC: top-down control to language/motor/perceptual cortices
- Let the developer exert executive controller in a program!
Architecture

Programmable Executive Controller

User NL → Large Language Model (LLM) → Semantic Parser → TT → External DB/API

Agent NL

control
Lecture Overview

1. Inspiration: Knowledge navigator
2. Overview on the state of the art
3. Overall architecture inspired by neuroscience
4. Core technical concepts
5. This course
## Summary of Pros and Cons

<table>
<thead>
<tr>
<th>Approach</th>
<th>Pros</th>
<th>Cons</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Dialogue Tree: GenieScript</td>
<td>Developer control of conversations</td>
<td>Intents lack expressiveness</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lacks user initiative</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Too many paths</td>
<td></td>
</tr>
<tr>
<td>(2) Semantic Parsing</td>
<td>Truth with grounding</td>
<td>Training data acquisition</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single commands</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Too narrow</td>
<td></td>
</tr>
<tr>
<td>(3) LLMs</td>
<td>Common knowledge, social norms and intelligence</td>
<td>Planning</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correctness, hallucination</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lacks attention, inhibition. Long-term memory,</td>
<td></td>
</tr>
</tbody>
</table>
Dialogue Plan

Greet

Find a flight (KB)

Book a flight (API)

Successful?

Rent a car (API)

Bye

Knowledge Base Navigation

- Products: stores, menus, flights, songs
- Financial databases, stock quotes, building facilities
- Schema.org
Expanding a Dialogue Plan

Genie provides a *comprehensive, generic* state machine for DB navigation.
Expanding a Dialogue Plan

Find a flight (KB)

Book a flight (API)

Successful?

Rent a car (API)

Actions
- Change passwords, return products, bank transfers
- Book hotels, appointments, buy insurance
- IoT commands

Genie provides a comprehensive, generic state machine for DB navigation

User
Agent

Greet

SearchRequest

SlotFillQuestion

ProposeRefine

ProposeN

ProposeOne

SearchRefine

SearchQuestion

InfoQuestion

ProvideInfo

End

Successful?

y

Rent a car (API)

n

Bye

y

Rent a car (API)

n
Expanding a Dialogue Plan

Genie provides a comprehensive, generic state machine for DB navigation and API execution.
A Restaurant Assistant

Request
I’d like to find a moderately priced restaurant

ProposeOne
I have Terun. It’s a moderately priced restaurant that serves pizza.

ProposeN
I found Terun and Coconuts. Both are moderately priced.

AskAction
I like that. Can you help me book it? I need it for 3 people.

SearchRefine
I don’t like pizza. Do you have something Caribbean?

InfoQuestion
Can you tell me the address of Terun?

Large Language Models
Automatically uses large language models to generate domain-specific training data & train
I don't like to be near the trains. Do you have something by the lake?

I'd like to find a moderately priced hotel.

I like that. Can you help me book it? I need it for 2 nights.

I have Best Western. It's a moderately priced hotel that is by the train station.

I found Best Western and Holiday Inn. Both are moderately priced.

I don't like to be near the trains. Do you have something by the lake?

Can you tell me the address of Best Western?
Do you have something more recent?

Request
What are the hits by Taylor Swift?

ProposeOne
I have *Shake it Off* from 2014.

ProposeN
I found *Shake it off* and *You Belong With Me.*

AskAction
I like that. Play the song please.

SearchRefine
Do you have something more recent?

InfoQuestion
Can you tell me the year of release?
No-Code Mixed-Initiative Agent

User

I want to transfer money

Agent

How can I help you?

What account would you like to transfer money to?

12512032

Got it. How much money would you like to transfer?

How much money do I have in my account?

You have $1,050

How much money would you like to transfer?
Conversational Assistant Architecture

Programmable Executive Controller

User NL

Large Language Model (LLM)

External DB/API

TT

Semantic Parser

Control: Prompting
[thoughtfulness, long term memory, goals]
Fine-tuning: learned 2nd nature capability

Agent NL
## Summary of Topics

<table>
<thead>
<tr>
<th>Approach</th>
<th>Pros</th>
<th>Cons</th>
<th>Solutions</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Dialogue Tree:</td>
<td>Developer control of conversations</td>
<td>Intents lack expressiveness</td>
<td>Use Thingtalk from (2)</td>
</tr>
<tr>
<td>GenieScript</td>
<td></td>
<td>Lacks user initiative</td>
<td>Combine with (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Too many paths</td>
<td>Generic DB state machines</td>
</tr>
<tr>
<td>(2) Semantic Parsing</td>
<td>Truth with grounding</td>
<td>Training data acquisition</td>
<td>Synthesize from db schemas, using LLMs</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Single commands</td>
<td>Conversational semantics</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Too narrow</td>
<td>Combine with (3)</td>
</tr>
<tr>
<td>(3) LLMs</td>
<td>Common knowledge, Social intelligence</td>
<td>Planning</td>
<td>Control from (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correctness, hallucination</td>
<td>Combine with (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lacks attention, inhibition,</td>
<td>Thoughtfulness (Current research)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Long-term memory,</td>
<td></td>
</tr>
</tbody>
</table>
Conversational Virtual Assistant

Music
Spotify

Podcasts
Spotify

Radio
TuneIn

News
SmartNews

Restaurants
Yelp

Search
Bing

Weather
Yr.no

Jokes
Dad Jokes

Reminder

Timer

Internet of Things

Appliances
Thermostat
Switch
Light
Fan
Door
Lock
Window Cover
Vacuum Cleaner

Sensors
Temperature
Motion
Illuminancne
Humidity
Flood
Battery
Ultra-violet

Podcasts
Spotify

Radio
TuneIn

News
SmartNews

Restaurants
Yelp

Search
Bing

Weather
Yr.no

Jokes
Dad Jokes

Reminder

Timer
Open-Source, Private Assistant
Demo on https://oval.cs.stanford.edu

1. On the web: genie.stanford.edu  
   Genie Demo

2. Baidu
   smartspeaker development kit

3. Home Assistant
   Open-source Home Automation Project
   - 100,000+ users
   - A local, private gateway controlling over 1000 IoT devices
   - Genie distributed as voice interface

Other partners: Picovoice, SmartNews, Yelp
Lecture Overview

1. Inspiration: Knowledge navigator
2. Overview on the state of the art
3. Overall architecture inspired by neuroscience
4. Core technical concepts
5. This course
This Course

State of the art:
• You learn it: Lectures
• You practice it: Homeworks
• You advance it: Projects

<table>
<thead>
<tr>
<th></th>
<th>Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Participation</td>
<td>15%</td>
</tr>
<tr>
<td>Assignment</td>
<td>25%</td>
</tr>
<tr>
<td>Final Project</td>
<td>60%</td>
</tr>
</tbody>
</table>
You Learn It: Lecture Design

<table>
<thead>
<tr>
<th>LLMs</th>
<th>Fluency; Common knowledge; Social intelligence</th>
</tr>
</thead>
</table>
| **Conversational Semantic Parser** | Fine-tune pre-trained networks to translate conversation to formal representation  
Synthesize training data with pre-trained networks  
Multilinguality |

| Executive Controller | High-level dialogue flow  
Attn & inhibition:  
- Control LLMs via fine-tuning & prompting  
Monitoring internal and external states  
(Long-term memory - TBD) |

Readings: https://cs224v.stanford.edu
You Practice it: Homework Design

Every student gets hands-on experience with the strengths of today’s 2 main technologies:

**Week 1: Social intelligence**
What is the social intelligence of LLMs (GPT3)?

**Week 2: Factual correctness**
Can semantic parsing answer questions from Wikidata?
You Advance it: Projects

- Project discussion starts on Wed!
- Many interesting quarterly projects
  - Convergence of 3 major developments
    (Dialogue flow + Semantic Parsing + LLMs → research opportunities
- Best projects were led by students who consulted with me (weekly) outside class
- Opportunity for independent research & research assistantship in Winter 2023.
<table>
<thead>
<tr>
<th>Weeks</th>
<th>Monday</th>
<th>Wednesday</th>
<th>Due</th>
<th>Project</th>
</tr>
</thead>
<tbody>
<tr>
<td>9/26, 28</td>
<td>Course Intro</td>
<td>Project Ideas</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/3, 5</td>
<td>NLP Intro</td>
<td>DB Queries (1)</td>
<td>HW1</td>
<td>Intent</td>
</tr>
<tr>
<td>10/10, 12</td>
<td>DB Queries (2)</td>
<td>Project discussion</td>
<td>HW2</td>
<td>Abstract</td>
</tr>
<tr>
<td>10/17, 19</td>
<td>Dialogue (1)</td>
<td>Dialogue (2)</td>
<td></td>
<td>Proposal</td>
</tr>
<tr>
<td>10/24, 26</td>
<td>Dialogue (3)</td>
<td>GenieScript</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10/31, 11/2</td>
<td>LLMs</td>
<td>Multi-lingual Agents</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/7, 9</td>
<td>Error Discovery</td>
<td>Virtual Assistant Anatomy</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/14, 16</td>
<td>Mixed Modal</td>
<td>Text-based QA</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11/28, 30</td>
<td>Speech</td>
<td>Advanced Topics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12/5, 7</td>
<td>Final Project</td>
<td>Final Project</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>