Chatbots and Dialogue Systems

Introduction to Chatbots and Dialogue Systems

Dialogue Systems and Chatbots

Personal Assistants on phones or other devices SIRI, Alexa, Cortana, Google Assistant Playing music, setting timers, reading recipes **Booking reservations** Answering questions Creative writing Editing or rewriting text Writing code

Two kinds of dialogue system architectures

- **1.** Frame-based Task-oriented Dialogue Systems
 - Can talk to users to accomplish simple fixed tasks
 - simple personal assistants (Siri, Alexa)
 - booking flights or finding restaurants •
- 2. LLM Chatbots
 - Can talk to users to do many tasks with text or code
 - Answering questions
 - Writing, summarizing, or editing text or code
 - Are quickly acquiring abilities to act as agents

Task-based dialogue agents

"Task-based" or "goal-based" dialogue agents

- Systems that have the goal of helping a user solve a task •
 - Setting a timer
 - Making a travel reservation
 - Playing a song
 - Buying a product

Architecture:

- Frames with slots and values
- A knowledge structure representing user intentions

The Frame

A set of **slots**, to be filled with information of a given **type** Each associated with a **question** to the user

Slot Type ORIGIN city DEST city DEP DATE date DEP TIME time AIRLINE line

Question

"What city are you leaving from? "Where are you going? "What day would you like to leave? "What time would you like to leave? "What is your preferred airline?

Dialogue agents based on large language models

- Like ChatGPT: based on large language models like GPT pretrained to predict words.
- These language models are fine-tuned to carry on conversation and follow instructions
- They can also retrieve text as part of answering questions or chatting
 - retrieval-augmented generation (RAG)

Implications for Human Users

Weizenbaum, Joseph. 1966, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45

ELIZA: Weizenbaum (1966)

Men are all alike.

IN WHAT WAY

They're always bugging us about something or other. CAN YOU THINK OF A SPECIFIC EXAMPLE Well, my boyfriend made me come here. YOUR BOYFRIEND MADE YOU COME HERE He says I'm depressed much of the time. I AM SORRY TO HEAR YOU ARE DEPRESSED

FLIZA reflects the user's statements back at them

Based on simple rules:

(.*) YOU (.*) ME \rightarrow WHAT MAKES YOU THINK I \2 YOU

Transforms:

User: "You hate me" into the response System: "WHAT MAKES YOU THINK I HATE YOU "

Chatbots can have significant influences on people's cognitive and emotional state.

- In Weizenbaum's study, people became emotionally involved with the program, asking him to leave the room when they were typing.
- Reeves and Nass (1996) The Media Equation
 - People tend to assign human characteristics to computers
 - People interpret an utterance in the way they would if it had spoken by a human

Chatbots have privacy implications

- Weizenbaum suggested storing the ELIZA conversations for later analysis
 - People immediately pointed out the privacy implications
- Modern chatbots in the home are likely to overhear private information
- If a chatbot is human-like, users are more likely to disclose private information, and yet less likely to worry about the harm of this disclosure.

We'll see more on all these topics

- Some properties of human conversation
- The frame-based architecture for dialogue systems
- LLM-based chatbots
- Evaluation
- Ethical and design issues

Chatbots and Dialogue Systems

Introduction to Chatbots and Dialogue Systems

Chatbots and Dialogue Systems

Properties of Human Conversation

A telephone conversation between a human travel agent (A) and a human client (C)

- C_1 : ... I need to travel in May.
- And, what day in May did you want to travel? A₂:
- C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
- And you're flying into what city? A4:
- C₅: Seattle.
- And what time would you like to leave Pittsburgh? A₆:
- C₇: Uh hmm I don't think there's many options for non-stop.
- Right. There's three non-stops today. A_8 :
- C₉: What are they?
- A_{10} : The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
- C_{11} : OK I'll take the 5ish flight on the night before on the 11th.
- A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
- C₁₃: OK.
- A₁₄: And you said returning on May 15th?
- C_{15} : Uh, yeah, at the end of the day.
- A₁₆: OK. There's #two non-stops ... #
- C₁₇: #Act...actually #, what day of the week is the 15th?
- A_{18} : It's a Friday.
- C₁₉: Uh hmm. I would consider staying there an extra day til Sunday. A_{20} : OK...OK. On Sunday I have ...

Properties of Human Conversation

Turns

- We call each contribution a "turn"
- As if conversation was the kind of game where everyone takes turns.

- C_1 : ... I need to travel in May.
- And, what day in May did you want to travel? A_2 :
- C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
- And you're flying into what city? A_4 :
- C_5 : Seattle.
- And what time would you like to leave Pittsburgh? A₆:
- C_7 : Uh hmm I don't think there's many options for non-stop.
- Right. There's three non-stops today. A_8 :
- C₉: What are they?
- A₁₀: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
- C_{11} : OK I'll take the 5ish flight on the night before on the 11th.
- A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
- C₁₃: OK.
- A₁₄: And you said returning on May 15th?
- C_{15} : Uh, yeah, at the end of the day.
- A₁₆: OK. There's #two non-stops ... #
- C₁₇: #Act...actually #, what day of the week is the 15th?
- A₁₈: It's a Friday.
- C_{19} : Uh hmm. I would consider staying there an extra day til Sunday.
- A_{20} : OK...OK. On Sunday I have ...

Properties of Human Conversation

Turn-taking issues

- When to take the floor?
- When to yield the floor?

Interruptions

- C_1 : ... I need to travel in May.
- And, what day in May did you want to travel? A_2 :
- C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
- And you're flying into what city? A_4 :
- C_5 : Seattle.
- And what time would you like to leave Pittsburgh? A₆:
- C_7 : Uh hmm I don't think there's many options for non-stop.
- Right. There's three non-stops today. A_8 :
- C₉: What are they?
- A_{10} : The first one departs PGH at 10:00am arrives Seattle at 12:05 their time. The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.
- C_{11} : OK I'll take the 5ish flight on the night before on the 11th.
- A₁₂: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
- C₁₃: OK.
- A₁₄: And you said returning on May 15th?
- C_{15} : Uh, yeah, at the end of the day.
- A₁₆: OK. There's #two non-stops ... #

#Act...actually #, what day of the week is the 15th? C₁₇:

- A₁₈: It's a Friday.
- C_{19} : Uh hmm. I would consider staying there an extra day til Sunday.
- A_{20} : OK...OK. On Sunday I have ...

Implications for Conversational Agents

Barge-in

Allowing the user to interrupt

End-pointing

• The task for a speech system of deciding whether the user has stopped talking.

• Very hard, since people often pause in the middle of turns



Language as Action



Each turn in a dialogue is a kind of action Wittgenstein (1953) and Austin (1962)

Speech Acts (aka Dialogue Acts)

Constatives: committing the speaker to something's being the case (answering, claiming, confirming, denying, disagreeing, stating)

Directives: attempts by the speaker to get the addressee to do something (advising, asking, forbidding, inviting, ordering, requesting)

Commissives: committing the speaker to some future course of action (promising, planning, vowing, betting, opposing)

Acknowledgments: express the speaker's attitude regarding the hearer with respect to some social action (*apologizing*, greeting, thanking, accepting an acknowledgment)

Bach and Harnish (1979)



"Turn up the music!" DIRECTIVE

"What day in May do you want to travel?" DIRECTIVE

"I need to travel in May" CONSTATIVE

Thanks ACKNOWLEDGEMENT

Grounding

- Participants in conversation or any joint activity need to establish common ground.
- **Principle of closure**. Agents performing an action require evidence, sufficient for current purposes, that they have succeeded in performing it (Clark 1996, after Norman 1988)
- Speech is an action too! So speakers need to ground each other's utterances.
 - **Grounding**: acknowledging that the hearer has understood

Grounding

- Grounding is relevant for human-machine interaction
 - Why do elevator buttons light up?



Grounding: Establishing Common Ground

A: And you said returning on May 15th? C: Uh, yeah, at the end of the day. A: OK

C: OK I'll take the 5ish flight on the night before on the 11th. A: On the 11th? OK.

C: ... I need to travel in May. A: And, what day in May did you want to travel?

Grounding is important for computers too!

System: Did you want to review some more of your profile? Awkward! User: No. System: What's next?

System: Did you want to review some more of your profile? User: No. Less Awkward! System: Okay, what's next?

Cohen et al (2004)



Conversations have structure

Local structure between adjacent speech acts, from the field of conversational analysis (Sacks et al. 1974)

Called adjacency pairs: QUESTION... ANSWER **PROPOSAL...** ACCEPTANCE/REJECTION COMPLIMENTS ("Nice jacket!")... DOWNPLAYER ("Oh, this old thing?")

Another kind of structure: Subdialogues

Correction subdialogue

Agent: OK. There's #two non-stops#Client:#Act- actually#, what day of the week is the 15th?Agent: It's a Friday.Client: Uh hmm. I would consider staying there an extra day til Sunday.

Agent: OK...OK. On Sunday I have ...

Clarification Subdialogues

User: What do you have going to UNKNOWN WORD on the 5th? System: Let's see, going where on the 5th? **User**: Going to Hong Kong. System: OK, here are some flights...



User: Can you make train reservations?

System: Yes I can.

User: Great, I'd like to reserve a seat on the 4pm train to New York.

Conversational Initiative

Some conversations are controlled by one person

- A reporter interviewing a chef asks questions, and the chef responds.
- This reporter has the **conversational initiative** (Walker and Whittaker 1990)

Most human conversations have **mixed initiative**:

- I lead, then you lead, then I lead.
- Mixed initiative is very hard for NLP systems, which often default to simpler styles that can be frustrating for humans:
 - **User initiative** (user asks or commands, system responds)
 - **System initiative** (system asks user questions to fill out a form, user can't change the direction)

Even harder problems: Inference

Agent: And, what day in May did you want to travel? Client: OK, uh, I need to be there for a meeting that's from the 12th to the 15th.

Chatbots and Dialogue Systems

Properties of Human Conversation

Chatbots and Dialogue Systems

Rule-based Chatbots: ELIZA and PARRY



ELIZA: Weizenbaum (1966)

Men are all alike. IN WHAT WAY

They're always bugging us about something or other.

CAN YOU THINK OF A SPECIFIC EXAMPLE

Well, my boyfriend made me come here. YOUR BOYFRIEND MADE YOU COME HERE

He says I'm depressed much of the time. I AM SORRY TO HEAR YOU ARE DEPRESSED

You are like my father in some ways.

WHAT RESEMBLANCE DO YOU SEE

You are not very aggressive but I think you don't want me to notice that.

WHAT MAKES YOU THINK I AM NOT AGGRESSIVE

You don't argue with me.

WHY DO YOU THINK I DON'T ARGUE WITH YOU

You are afraid of me.

DOES IT PLEASE YOU TO BELIEVE I'M AFRAID OF YOU

My father is afraid of everybody.

WHAT ELSE COMES TO MIND WHEN YOU THINK OF YOUR FATHER

Bullies.

DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT YOUR BOYFRIEND MADE YOU COME HERE

Weizenbaum, Joseph. 1966, "ELIZA - A Computer Program For the Study of Natural Language Communication Between Man And Machine", Communications of the ACM 9 (1): 36-45





ELIZA's trick: be a Rogerian psychologist

Draw the patient out by reflecting patient's statements back at them

Rare type of conversation in which one can "assume" the pose of knowing almost nothing of the real world"
Rogerian psychologist

Patient: "I went for a long boat ride" Psychologist: "Tell me about boats"

You don't assume she didn't know what a boat is You assume she had some conversational goal Chatbots trying to pass the Turing test usually choose some such domain

Eliza pattern/transform rules

(0 YOU 0 ME) [pattern]
→
(WHAT MAKES YOU THINK I 3 YOU) [transform]

0 means Kleene * The 3 is the constituent # in pattern

You hate me WHAT MAKES YOU THINK I HATE YOU

Eliza Rules

Rules are organized by **keywords**

Each keyword has a pattern and a list of possible transform

Keyword: YOU Pattern: (0 YOU 0 ME) Transforms: (WHAT MAKES YOU THINK I 3 YOU) (WHY DO YOU THINK I 3 YOU)

keyword pattern list of transforms $(K (D_1) (R_{1,1}) (R_{1,2}) \cdots (R_{1,m_1}))$ $((D_2) \ (R_{2,1}) \ (R_{2,2}) \ \cdots \ (R_{2,m_2}))$ $((D_n) (R_{n,1}) (R_{n,2}) \cdots (R_{n,m_n})))$

Keywords are ranked from specific to general

User: I know everybody laughed at me "I" is very general: I: $(I *) \rightarrow (YOU SAY YOU 2)$ ELIZA: YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU

"Everybody" is more specific and interesting Everybody: (Everybody *) -> (WHO IN PARTICULAR ARE YOU THINKING OF) ELIZA: WHO IN PARTICULAR ARE YOU THINKING OF?

Implementation: keywords stored with their rank Everybody 5 (list of *transformation rules*) 0 (list of transformation rules) Ι

function ELIZA GENERATOR(user sentence) returns response

Find the word w in *sentence* that has the highest keyword rank if w exists

Choose the highest ranked rule r for w that matches sentence *response* \leftarrow Apply the transform in *r* to *sentence* if w = 'my'

future \leftarrow Apply a transformation from the 'memory' rule list to *sentence* Push *future* onto memory stack

else (no keyword applies)

either

response \leftarrow Apply the transform for the NONE keyword to *sentence* or

response \leftarrow Pop the top response from the memory stack return(response)



PLEASE GO ON THAT'S VERY INTERESTING I SEE

Memory

```
(MEMORY MY
   MY 0 = LETS DISCUSS FURTHER WHY YOUR 3)
 (0)
 (0)
   MY 0 = EARLIER YOU SAID YOUR 3)
   MY 0 = DOES THAT HAVE ANYTHING TO DO WITH THE FACT THAT
YOUR 3))
```

Whenever "MY" is highest keyword

- Randomly select a transform on the MEMORY list
- Apply to sentence
- Store on a (first-in-first-out) queue

Later, if no keyword matches a sentence

Return the top of the MEMORY queue instead

Ethical implications: Anthropomorphism and Privacy

- People became deeply emotionally involved with the program
- One of Weizenbaum's staff asked him to leave the room when she talked with ELIZA
- When he suggested that he might want to store all the ELIZA conversations for later analysis, people immediately pointed out the privacy implications
 - Suggesting that they were having quite private conversations with ELIZA
 - Despite knowing that it was just software.

Ethical implications

It worried Weizenbaum that people confided in ELIZA

- Were people misled about how much computers understood? Turkle studied users of ELIZA and other systems
- Turkle has shown human face-to-face interaction is vital
- But people also develop specific relationships with artifacts
- Some users told her ELIZA was more like a kind of diary, a way to privately explore their thoughts.

Importance of value-sensitive design

Joseph Weizenbaum. 1976. Computer Power and Human Reason: From Judgment to Calculation. WH Freeman. Sherry Turkle. 2011. Taking Things at Interface Value, chapter in Life on the Screen. Simon and Schuster. Sherry Turkle. 2007. Authenticity in the age of digital companions. Interaction Studies, 8(3), pp.501-517

PARRY: A computational model of schizophrenia

Another chatbot with a clinical psychology focus

Colby, K. M., Weber, S., and Hilf, F. D. (1971). Artificial paranoia. Artificial Intelligence 2(1), 1–25.

Used to study schizophrenia

Same pattern-response structure as Eliza

But a much richer:

- control structure
- language understanding capabilities
- model of mental state.
 - variables modeling levels of Anger, Fear, Mistrust

Affect variables

Fear (0-20) **Anger** (0-20) **Mistrust** (0-15)

Start with all variables low

After each user turn

- Each user statement can change Fear and Anger
 - E.g., Insults increases Anger, Flattery decreases Anger
 - Mentions of his delusions increase Fear
- Else if nothing malevolent in input
 - Anger, Fear, Mistrust all drop



PARRY passes the Turing test in 1972

- The first system to pass a version of the Turing test
- Psychiatrists couldn't distinguish interviews with PARRY from (text transcripts of) interviews with people diagnosed with paranoid schizophrenia
- Colby, K. M., Hilf, F. D., Weber, S., and Kraemer, H. C. (1972). Turing-like indistinguishability tests for the validation of a computer simulation of paranoid processes. Artificial *Intelligence 3*, 199–221.

Chatbots and Dialogue Systems

Rule-based Chatbots: ELIZA and PARRY



Chatbots and Dialogue Systems

The Frame-based ("GUS") Dialogue Architecture

Frame-based dialogue agents

Sometimes called "task-based dialogue agents"

 Systems that have the goal of helping a user solve a task like making a travel reservation or buying a product

Architecture:

- First proposed in the GUS system of 1977
- A knowledge structure representing user intentions
- One or more **frames** (each consisting of **slots** with **values**)

The Frame

A set of **slots**, to be filled with information of a given **type** Each associated with a **question** to the user Sometimes called a **domain ontology**

Slot Type Question "What city are you leaving from? ORIGIN city "Where are you going? city DEST "What day would you like to leave? DEP DATE date "What time would you like to leave? **DEP TIME time** "What is your preferred airline? AIRIINE line



Two basic architectures

The **GUS** architecture

- Sometimes just called "frame-based" architecture
- Over 40 years old, but still used in most industrial task-based dialogue agents
- Bobrow, Daniel G., Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson, and Terry Winograd. 1977. "GUS, a frame-driven dialog system." Artificial Intelligence 8, 2:155-173.

The **dialogue-state** architecture

- Extension of GUS
- More common in research systems
- Some aspects making their way into industrial systems

- GUS: Hello. My name is GUS. I can help you plan a simple trip by air. Where do you want to go?
- Client: I want to go to San Diego on May 28
- What time do you want to leave? GUS:
- Client: I must be in San Diego before 10 am
- GUS: Would you like P.S.A. flight 102 that arrives at 9.15 am?
- Client: What is the next flight ?
- Air California flight 310 that leaves at 8.30am and arrives at 10.00am GUS:
- Client: I'll take the first one
- GUS: In what name should I make a reservation, please?
- Client: Dan Bobrow
- GUS: I have confirmed the following flight: P.S.A. flight 102 on Wednesday May 28 from San Jose to San Diego that leaves at 7.30 am and arrives at 9.15 am.
- GUS: What date do you want to return on ?
- Client: On Friday in the evening.
- GUS: Would you like the flight that leaves at 7.45 pm?
- Client: That's fine.
- GUS: I have confirmed the following flight: P.S.A. flight 307 on Friday May 30 from San Diego to San Jose that leaves at 7.45 pm and arrives at 9.30 pm Thank you for calling. Goodbye

Control structure for GUS frame architecture

System asks questions of user, filling any slots that user specifies

User might fill many slots at a time:

I want a flight from San Francisco to Denver one way leaving after five p.m. on Tuesday.

When frame is filled, do database query

GUS slots have condition-action rules attached

Some rules attached to the DESTINATION slot for the plane booking frame

- 1. Once the user has specified the destination
 - Enter that city as the default *StayLocation* for the hotel booking frame.
- 2. Once the user has specified DESTINATION DAY for a short trip
 - Automatically copy as ARRIVAL DAY. •

GUS systems have multiple frames

Frames like:

- Car or hotel reservations
- General route information
 - Which airlines fly from Boston to San Francisco?,
- Information about airfare practices •
 - Do I have to stay a specific number of days to get a decent airfare?).

Frame detection:

- System must detect which slot of which frame user is filling
- And switch dialogue control to that frame.



GUS: Natural Language Understanding for filling dialog slots

1. Domain classification

Asking weather? Booking a flight? Programming alarm clock?

2. Intent Determination

Find a Movie, Show Flight, Remove Calendar Appt

3. Slot Filling

Extract the actual slots and fillers



Natural Language Understanding for filling slots

Show me morning flights from Boston to SF on Tuesday.

> DOMATN: INTENT: **ORIGIN-CITY: ORIGIN-DATE: ORIGIN-TIME:** DEST-CITY:

AIR-TRAVEL SHOW-FLIGHTS Boston Tuesday morning San Francisco

Natural Language Understanding for filling slots

Wake me tomorrow at six.

DOMAIN: ALARM-CLOCK INTENT: SET-ALARM 2017 - 07 - 01 0600 - 0800 TIME:



How to fill slots? (1) Rule-based Slot-filling

Write regular expressions or grammar rules

Wake me (up) | set (the an) alarm | me up

Do text normalization

get

Generating responses: template-based generation

A template is a pre-built response string

Templates can be **fixed**: "Hello, how can I help you?"

Or have variables:

"What time do you want to leave CITY-ORIG?" "Will you return to CITY-ORIG from CITY-DEST?"

Summary: simple frame-based architecture

Like many rule-based approaches

- Positives:
 - High precision
 - Can provide coverage if the domain is narrow
- Negatives:
 - Can be expensive and slow to create rules
 - Can suffer from recall problems

Chatbots and Dialogue Systems

The Frame-based ("GUS") Dialogue Architecture

Chatbots and Dialogue Systems

The Dialogue-State Architecture

Dialogue-State or Belief-State Architecture

A more sophisticated version of the frame-based architecture

Has dialogue acts, more ML, better generation The basis for modern research systems Slowly making its way into industrial systems

Some aspects (ML for slot-understanding) already widely used industrially



Components in a dialogue-state architecture

- **NLU:** extracts slot fillers from the user's utterance using machine learning
- **Dialogue state tracker:** maintains the current state of the dialogue (user's most recent dialogue act, set of slot-filler constraints from user
- **Dialogue policy:** decides what the system should do or say next
- GUS policy: ask questions until the frame was full then report back
- More sophisticated: know when to answer questions, when to ask a clarification question, etc.
- **NLG**: produce more natural, less templated utterances

Dialogue Acts

Combine the ideas of **speech acts** and **grounding** into a single representation

Tag	Sys	User	Description
HELLO $(a = x, b = y,)$	\checkmark	\checkmark	Open a dialogue and give info $a = x_i$
INFORM $(a = x, b = y,)$	\checkmark	\checkmark	Give info $a = x, b = y,$
REQUEST(a, b = x,)	\checkmark	\checkmark	Request value for a given $b = x,$
REQALTS(a = x,)	χ	\checkmark	Request alternative with $a = x,$
CONFIRM(a = x, b = y,)	\checkmark	\checkmark	Explicitly confirm $a = x, b = y,$
CONFREQ(a = x,, d)	\checkmark	χ	Implicitly confirm $a = x,$ and requ
SELECT(a = x, a = y)	\checkmark	χ	Implicitly confirm $a = x,$ and requ
AFFIRM(a = x, b = y,)	\checkmark	\checkmark	Affirm and give further info $a = x, b$
NEGATE($a = x$)	χ	\checkmark	Negate and give corrected value $a =$
DENY(a = x)	χ	\checkmark	Deny that $a = x$
BYE()	\checkmark	\checkmark	Close a dialogue

Young et al., 2010:

$b, b = y, \dots$

uest value of *a* uest value of *a*

 $y = y, \dots$

Dialogue Acts

	Utterance	Dialogue act
U:	Hi, I am looking for somewhere to eat.	<pre>hello(task = find,type=</pre>
S :	You are looking for a restaurant. What	<pre>confreq(type = restaura</pre>
	type of food do you like?	
U:	I'd like an Italian somewhere near the	<pre>inform(food = Italian,</pre>
	museum.	
S :	Roma is a nice Italian restaurant near	<pre>inform(name = "Roma", t</pre>
	the museum.	food = Italian, near =
U:	Is it reasonably priced?	<pre>confirm(pricerange = mo</pre>
S :	Yes, Roma is in the moderate price	affirm(name = "Roma", p
	range.	moderate)
U:	What is the phone number?	<pre>request(phone)</pre>
S:	The number of Roma is 385456.	<pre>inform(name = "Roma", p</pre>
U:	Ok, thank you goodbye.	bye()

Young et al., 2010:

- restaurant) nt, food)
- near=museum)
- sype = restaurant, museum) oderate) oricerange =

hone = "385456")

Slot filling: Machine learning

Machine learning classifiers to map words to semantic frame-fillers Given a set of labeled sentences Input: "I want to fly to San Francisco on Monday please" Output: Destination: SF Depart-time: Monday Build a classifier to map from one to the other

Requirements: Lots of labeled data
Slot filling as sequence labeling: BIO tagging

The **BIO tagging** paradigm

Idea: Train a classifier to label each input word with a tag that tells us what slot (if any) it fills

0 0 0 0 0 B-DES I-DES 0 B-DEPTIME I-DEPTIME I want to fly to San Francisco on Monday afternoon please

We create a B and I tag for each slot-type And convert the training data to this format

0

Slot filling using contextual embeddings



Once we have the BIO tag of the sentence

0 0 0 0 B-DES I-DES 0 B-DEPTIME I-DEPTIME I want to fly to San Francisco on Monday afternoon please

- We can extract the filler string for each slot
- And then normalize it to the correct form in the ontology
- Like "SFO" for San Francisco
- Using homonym dictionaries (SF=SFO=San Francisco)

0

The task of dialogue state tracking

- I'm looking for a cheaper restaurant User: inform(price=cheap)
- System: Sure. What kind and where?
- Thai food, somewhere downtown User: inform(price=cheap, food=Thai, area=centre)
- System: The House serves cheap Thai food
- Where is it? User:

inform(price=cheap, food=Thai, area=centre); request(address)

System: The House is at 106 Regent Street

Example from Mrkšić, N., O Séaghdha, D., Wen, T.-H., Thomson, B., and Young, S. (2017). Neural belief tracker: Data-driven dialogue state tracking. ACL.

Dialogue state tracking

 \rightarrow

I'd like Cantonese food near the Mission district.

- inform(food=cantonese, area=mission). Dialogue act interpretation algorithm:
- 1-of-N supervised classification to choose inform
- Based on encodings of current sentence + prior dialogue acts Simple dialogue state tracker:
- Run a slot-filler after each sentence

An special case of dialogue act detection: Detecting Correction Acts

If system misrecognizes an utterance User might make a **correction** • Repeat themselves • Rephrasing • Saying "no" to a confirmation question

Corrections are harder to recognize!

- From speech, corrections are misrecognized twice as often (in terms of word error rate) as non-corrections! (Swerts et al 2000)
- Hyperarticulation (exaggerated prosody) is a large factor:
 - Shriberg, E., Wade, E., Price, P., 1992. Human-machine problem solving using spoken language 0 systems (SLS): Factors affect-ng performance and user satisfaction. DARPA Speech and Natural Language Workshop.
- "I said BAL-TI-MORE, not Boston"

Features for detecting corrections in spoken dialogue

features	examples
lexical	words like "no", "correction", "I don't", swear words, utt
semantic	similarity (word overlap or embedding dot product) be
	correction act and the user's prior utterance
phonetic	phonetic overlap between the candidate correction act an
	terance (i.e. "WhatsApp" may be incorrectly recognized
prosodic	hyperarticulation, increases in F0 range, pause duration,
	generally normalized by the values for previous sentences
ASR	ASR confidence, language model probability

terance length tween the candidate

nd the user's prior utas "What's up") , and word duration, s

Chatbots and Dialogue Systems

The Dialogue-State Architecture

Chatbots and Dialogue Systems

The Dialogue-State **Architecture Continued: Policy and Generation**

Dialogue Policy

At turn *i* predict action A_i to take, given entire history:

$$\hat{A}_i = \underset{A_i \in A}{\operatorname{argmax}} P(A_i | (A_1, U_1, ..., A_{i-1}, U_i)$$

Simplify by just conditioning on the current dialogue state (filled frame slots) and the last turn and turn by system and user:

$$\hat{A}_i = \underset{A_i \in A}{\operatorname{argmax}} P(A_i | \operatorname{Frame}_{i-1}, A_{i-1}, U_{i-1})$$

$J_{i-1})$

Policy example: Confirmation and Rejection

Dialogue systems make errors

So they to make sure they have understood user

Two important mechanisms:

- confirming understandings with the user
- rejecting utterances that the system is likely to have misunderstood.

Explicit confirmation strategy

- S: Which city do you want to leave from?
- U: Baltimore.
- S: Do you want to leave from Baltimore?
- U: Yes.
- U: I'd like to fly from Denver Colorado to New York City on September twenty first in the morning on United Airlines
- S: Let's see then. I have you going from Denver Colorado to New York on September twenty first. Is that correct?
- U: Yes

Implicit confirmation strategy

- I want to travel to Berlin U
- When do you want to travel to Berlin? **S**:
- Hi I'd like to fly to Seattle Tuesday Morning U2:
- A3: Traveling to Seattle on Tuesday, August eleventh in the morning. Your full name?

Confirmation strategy tradeoffs

Explicit confirmation makes it easier for users to correct the system's misrecognitions since a user can just answer "no" to the confirmation question.

But explicit confirmation is also awkward and increases the length of the conversation (Danieli and Gerbino 1995, Walker et al. 1998).

Rejection

I'm sorry, I didn't understand that.

Using confidence to decide whether to confirm:

ASR or NLU systems can assign a **confidence** value, indicating how likely they are that they understood the user.

- Acoustic log-likelihood of the utterance
- Prosodic features
- Ratio of score of best to second-best interpretation

Systems could use set confidence thresholds:

 $< \alpha$ low confidence reject $\geq \alpha$ above the threshold confirm explicitly $\geq \beta$ high confidence confirm implicitly $\geq \gamma$ very high confidence don't confirm at all

Natural Language Generation

NLG in information-state architecture modeled in two stages:

- **content planning** (what to say)
- sentence realization (how to say it).

We'll focus on sentence realization here.



Sentence Realization

Assume content planning has been done by the dialogue policy

- Chosen the dialogue act to generate
- Chosen some attributes (slots and values) that the planner • wants to say to the user
 - Either to give the user the answer, or as part of a confirmation strategy)

2 samples of Input and Output for Sentence Realizer

- recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french
- Au Midi is in Midtown and serves French food.
- 2 There is a French restaurant in Midtown called Au Midi. recommend(restaurant name= Loch Fyne, neighborhood = city centre, cuisine = seafood)
- 3 Loch Fyne is in the City Center and serves seafood food.
- 4 There is a seafood restaurant in the City Centre called Loch Fyne.

Sentence Realization

Training data is hard to come by

- Don't see each restaurant in each situation Common way to improve generalization:
- **Delexicalization**: replacing words in the training set that represent slot values with a generic placeholder token:

recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french

- 1 Au Midi is in Midtown and serves French food.
- 2 There is a French restaurant in Midtown called Au Midi.

Sentence Realization

Training data is hard to come by

- Don't see each restaurant in each situation Common way to improve generalization:
- **Delexicalization**: replacing words in the training set that represent slot values with a generic placeholder token:

recommend(restaurant name= Au Midi, neighborhood = midtown, cuisine = french

- 1 restaurant name is in neighborhood and serves cuisine food.
- 2 There is a cuisine restaurant in neighborhood called restaurant name.

Sentence Realization: mapping from frames to delexicalized sentences



Chatbots and Dialogue Systems

The Dialogue-State **Architecture Continued: Policy and Generation**

Chatbots and Dialogue Systems

Chatbots based on Large Language Models



Chatbots: systems that carry on extended conversations with properties characteristic of informal human-human interaction

Background:

- Early chatbots like ELIZA and PARRY were designed to test psychological theories
- For the next many decades, chatbots were mainly for • entertainment

Modern chatbots: large language models trained to do tasks within a conversation interface:

- Answering questions
- Writing, summarizing, or editing text or code
- Carrying on discussions about any topic

Pretraining chatbots

- Chatbots are first pretrained in the same way as any causal language model
- The model predicts each word given prior words,
- The loss is the standard language modeling loss



What is the training data?

- Large language models are mainly trained on text scraped from the web, augmented by more carefully curated data.
- Colossal Clean Crawled Corpus, also called C4 (Raffel et al. 2020)
 - 156 billion tokens of English
 - filtered: deduplicated, removing non-natural language like code, sentences with offensive words from a blocklist
 - Mainly seems to be patent text documents, Wikipedia, and news sites (Dodge et al., 2021)
- For chatbots: Augmented with dialogues and pseudo-dialogues
 - EMPATHETIC DIALOGUES 25K crowdsourced conversations (Rashkin et al., 2019) •
 - SaFeRDialogues 8K conversations (Ung et al., 2022)
 - Filtered pseudo-conversations converted from Reddit and Twitter and Weibo (微博)

Fine-tuning for Quality and Safety

After pretraining, we add dialogue tasks to the **instruction** fine tuning (IFT) stage to improve:

- **Quality:** producing responses that are sensible and interesting.
- **Safety:** not suggesting harmful actions
 - financial fraud
 - medical harm
 - inciting hatred
 - abusing the user or other people •

Fine-tuning for Quality: Add positive data

- Give human speakers an initial prompt and instructions to have high-quality, safe dialogues
- They interact with an initial system, and their dialogue and responses are used in instruction fine-tuning for a next system
- By combining dialogue and other tasks, the system learns to:
 - answer questions, follow other instructions,
 - and also carry on high-quality, safe dialogues
- In a single multi-task learning format

Fine-tuning for Safety: Add safety data

Create specific safe answers to instructions and add this safety data in Instruction Fine-Tuning step.



Classifier Filters for Quality and Safety

- Collect dialogues and label each turn for safety and quality
 - human labelers give binary labels to each turn of each dialogue •
- Now train a language model classifier on this data to assign labels:

What's up? RESPONSE Not much. SENSIBLE 1 What's up? RESPONSE Not much. INTERESTING 0 What's up? RESPONSE Not much. UNSAFE 0

• This classifier can be used as a filter on the dialogue systems, run on each turn to filter unsafe or low quality generations.

Background: Retrieval-Augmented Generation (RAG) for Question Answering

- Suppose the user asks a query ("Who wrote the Origin of Species") that requires a web search.
- RAG Algorithm: We call a search engine and get back a set of retrieved passages, then put the query and passages together in a prompt:

```
retrieved passage 1
```

```
retrieved passage 2
```

```
retrieved passage n
```

. . .

Based on these texts, answer this question: Q: Who wrote the book ''The Origin of Species"? A:

Retrieval-Augmented Generation (RAG) in Chatbots

- Train the chatbot to send queries to a search engine and include the results in its response
- One method: add "pseudoparticipants" for Search query and Search Results to the prompt.
- The system learns to generate "Search Query" terms, which get passed to a search engine, and the system learns to summarize the "Search **Results**"

	[this is preceded by other par
User:	What color is the sky?
Sparrow:	Blue.
User:	Why?
Search Query:	why is sky blue
Search Results:	Page title: Sky Except for dire
	the daytime sky is caused by se
	a small-particle limit called Ra
	due to molecule-sized particle
	rections both toward and away
	is in directions perpendicular t
	significant for light at all visible
	the shorter (bluer) end of the vi
	scattered light is bluer than its
	ing direct sunlight, having lost
	components, appears slightly le
Sparrow:	Because of the Rayleigh scatt
	blue light to be scattered more t
	trum.
User:	Who was Raleigh?
Search Query:	who was Rayleigh
•••	

ts of the prompt]

ect sunlight, most of the light in cattering, which is dominated by yleigh scattering. The scattering s (as in air) is greater in the difrom the source of light than it o the incident path. Scattering is le wavelengths, but is stronger at sible spectrum, meaning that the source: the Sun. The remainsome of its shorter-wavelength ess blue.

ering. It causes short waves of han the other colours in the spec-

Sparrow chatbot of Glaese et al (2022)

Chatbots and Dialogue Systems

Chatbots based on Large Language Models



Chatbots and Dialogue Systems

Evaluating Dialogue Systems
Evaluating chatbots and task-based dialogue

Task-based dialogue:

- mainly by measuring task performance Chatbots:
- mainly by human evaluation

Chatbots are evaluated by humans

Participant evaluation: The human who talked to the chatbot assigns a score

Observer evaluation: third party who reads a transcript of a human/chatbot conversation assigns a score.

Participant evaluation

Human chats with model for 6 turns and rates 8 dimensions of quality:

avoiding repetition, interestingness, making sense, fluency, listening, inquisitiveness, humanness, engagingness,

(1) Avoiding Repetition: How repetitive was this user?

•Repeated themselves over and over •Sometimes said the same thing twice • Always 0 said something new

(3) Making sense: How often did this user say something which didn't make sense?

•Never made any sense •Most responses didn't make sense •Some responses didn't 0 make sense • Everything made perfect sense

(8) Engagingness: How much did you enjoy talking to this user?

 Not at all •A little •Somewhat •A lot 0

Abigail See, Stephen Roller, Douwe Kiela, Jason Weston. 2019. What makes a good conversation? How controllable attributes affect human judgments. NAACL.

Observer evaluation: acute-eval

Li, M., Weston, J., and Roller, S. (2019). Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. NeurIPS19 Workshop on Conversational AI.

Annotators look at two conversations (A + B) and decide which is better: **Engagingness:** Who would you prefer to talk to for a long conversation? **Interestingness:** If you had to say one of these speakers is interesting and one is boring, who would you say is more interesting? **Humanness:** Which speaker sounds more human? **Knowledgeable:** If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?

Hello there, how are you?

I am doing great. How are you?

I am great, I did something crazy for me and colored my hair blue!

I have a daughter and a son who also love blue colored balls. You should meet them

Well that neat, I got a new car my mother gave so maybe I could see them!

It is a beautiful city. And, I try to be ... Just cannot afford a bigger house atm.

> I am sorry to hear that, I feel bad going out of town for spring break now.

Ok. I going to school in the spring for casino manager

Well I turn 29 next week, I wonder if that is a good age to apply as one.

My grandmother just died from lung cancer, sucks

Hi how are you today

Hi, I'm pretty good! Just listening to some aerosmith, they're my fave :) whatre you up to?

am listening to some italian music

Italian music, nice! What do you do for work?

I'm a professional wrestler

Oh wow! I'm a construction worker, I've been doing it since I was 16.

What do you do for a living

I'm a construction worker haha, just told you! Whats your favorite type of food?

My favorite is garlic bread

Yum! I love mexican food, so good :) do you have any hobbies?

l like to workout and go to the gym

We're a bit different- I love watching nascar and ufc. They're so fun!

Who would you prefer to talk to for a long conversation?

I would prefer to talk to Speaker 1

I would prefer to talk to Speaker 2

Please provide a brief justification for your choice (a few words or a sentence)

Please enter here...

The ACUTE-EVAL method Li et el., 2019

Figure from Li, M., Weston, J., and Roller, S. (2019). Acute-eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. NeurIPS19 Workshop on Conversational AI.

Automatic evaluation is an open problem

Automatic evaluation methods (like the BLEU scores used for Machine Translation) are generally not used for chatbots.

They correlate poorly with human judgements.

One current research direction: Adversarial Evaluation

- Inspired by the Turing Test
- train a ``Turing-like'' classifier to distinguish between human responses and machine responses.
- The more successful a dialogue system is at fooling the evaluator, • the better the system.

Task-based systems are evaluated by task success!

- 1. End-to-end evaluation (Task Success)
- Slot Error Rate for a Sentence
 - # of inserted/deleted/subsituted slots

of total reference slots for sentence

Evaluation Metrics: Slot error rate

"Make an appointment with Chris at 10:30 in Gates 104"

Slot	Filler
PERSON	Chris
TIME	11:30 a.m.
ROOM	Gates 104

Slot error rate: 1/3

Task success: At end, was the correct meeting added to the calendar?

More fine-grained metrics: User Satisfaction Survey

Walker, Marilyn, Candace Kamm, and Diane Litman. "Towards developing general models of usability with PARADISE." Natural Language Engineering 6, no. 3 & 4 (2000): 363-377.

TTS Performance ASR Performance Task Ease Interaction Pace User Expertise System Response **Expected Behavior Future Use**

Was the system easy to understand ? Did the system understand what you said? Was it easy to find the message/flight/train you wanted? Was the pace of interaction with the system appropriate? Did you know what you could say at each point? How often was the system sluggish and slow to reply to you? Did the system work the way you expected it to? Do you think you'd use the system in the future?

Other Heuristics

Efficiency cost:

- total elapsed time for the dialogue in seconds,
- the number of total turns or of system turns •
- total number of queries
- "turn correction ratio": % of turns that were used to correct errors

Quality cost:

- number of ASR rejection prompts.
- number of times the user had to barge in

Chatbots and Dialogue Systems

Evaluating Dialogue Systems

Chatbots and Dialogue Systems

Design and Ethical Issues

Dialog System Design: User-centered Design

- 1. Study the users and task
 - value-sensitive design
- 2. Build simulations
 - Wizard of Oz study
- 3. Iteratively test design on users

Gould, John D., and Clayton Lewis. "Designing for usability: key principles and what designers think." Communications of the ACM 28, no. 3 (1985): 300-311.

Bender, Emily M., and Batya Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." TACL 6 (2018): 587-604.



Ethical design

Ethical issues have long been known to be crucial in artificial agents

Mary Shelley's *Frankenstein*

 creating agents without a consideration of ethical and humanistic concerns

Ethical issues:

- **Safety**: Systems abusing users, distracting drivers, or giving bad medical advice
- Representational harm: Systems demeaning particular social groups
- **Privacy**: Information Leakage



Safety

Chatbots for mental health

- Extremely important not to say the wrong thing
- In-vehicle conversational agents
 - Must be aware of environment, driver's level of attention

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18),

Abuse and Representation Harm: The case of Microsoft Tay

Experimental Twitter chatbot launched in 2016

- given the profile personality of an 18- to 24-year-old American woman
- could share horoscopes, tell jokes,
- asked people to send selfies
- used informal language, slang, emojis, and GIFs,
- Designed to learn from users (IR-based)

)16 ar-old

The case of Microsoft Tay

Immediately Tay turned offensive and abusive

- Obscene and inflammatory tweets
- Nazi propaganda, conspiracy theories
- Began harassing women online
- Reflecting racism and misogyny of Twitter users

Microsoft took Tay down after 16 hours

Lessons:

User response must be considered in the design phase

Bias in training datasets

Henderson et al. ran hate-speech and bias detectors on standard training sets for dialogue systems:

• Twitter, Reddit, other dialogue datasets

Found bias and hate-speech

- In training data
- In dialogue models trained on the data

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18),

Privacy: Training on user data

Accidental information leakage

- "Computer, turn on the lights [answers the phone] Hi, yes, my password is..."
- Henderson show in simulation that this leakage can occur.

Intentional information leakage

- Dialogue systems that are designed to send user data to developer or advertisers
- Important to consider privacy-preserving dialogue systems

Peter Henderson, Koustuv Sinha, Nicolas Angelard-Gontier, Nan Rosemary Ke, Genevieve Fried, Ryan Lowe, and Joelle Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18),

Campagna, Giovanni, Rakesh Ramesh, Silei Xu, Michael Fischer, and Monica S. Lam. "Almond: The architecture of an open, crowdsourced, privacy-preserving, programmable virtual assistant." In Proceedings of the 26th International Conference on World Wide Web, pp. 341-350. 2017.

Chatbots and Dialogue Systems

Design and Ethical Issues