

Dialogue

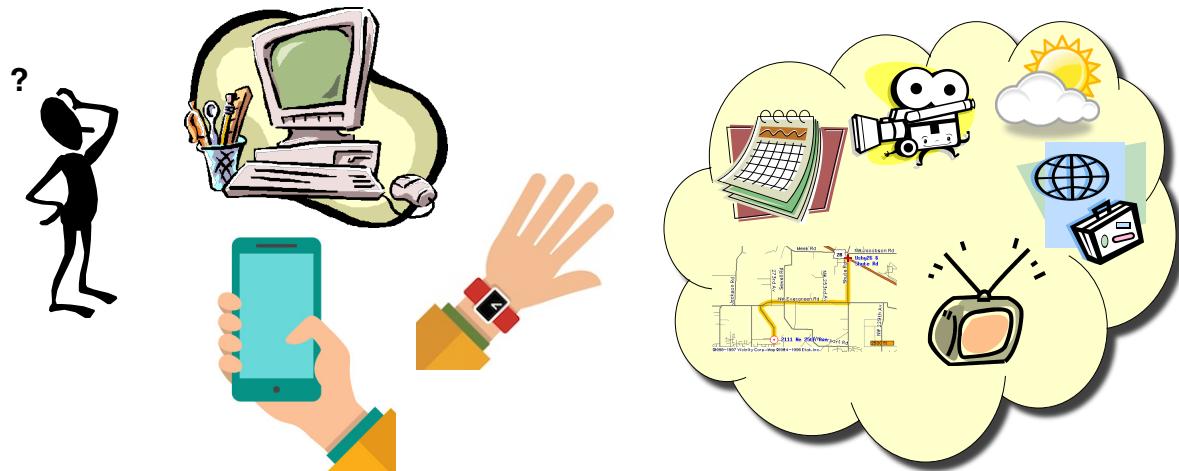
Jim Glass / MIT 6.806-6.864 / Spring 2021

Dialogue Storyline

- Introduction
- Chatbots
- Task-oriented dialogue systems
- Ongoing challenges

Human-Computer Interaction (HCI)

- HCI focuses on the interface between people and computers



What is a Dialogue System?

- Supports spoken, text-based, or multimodal conversational interactions between humans and computers
 - Replaces traditional graphical user interface with natural language
- Mainly **task-oriented** and *non-task-oriented* dialogue systems
- Described by various terms, e.g.,
 - chatbot, socialbot
 - personal digital assistant
 - statistical dialogue system
 - conversational interface/agent
 - conversational AI

Why Dialogue Systems?

Natural	Requires no special training
Flexible	Leaves hands and eyes free
Efficient	Has high data rate
Economical	Can be transmitted and received inexpensively

Dialogue ideal for information access & management when:

- The information space is broad and complex
- The users are technically naïve, or
- The device is small (e.g., spoken dialogue)

A Human Dialogue Example

..... disfluency

C: Yeah, [um] I'm looking for the Buford Cinema.

A: OK, and you're wanting to know what's showing there or ... interruption, overlap

C: Yes, please. confirmation

A: Are you looking for a particular movie? clarification

C: [um] What's showing.

A: OK, one moment. back channel

.....

A: They're showing *A Troll In Central Park*.

C: No. inference

A: *Frankenstein*. ellipsis

C: What time is that on? co-reference

A: Seven twenty and nine fifty.

C: OK, any others?



A Human Dialogue Example (con't)

.....

- A: Eight o'clock.
 [um] Shawshank Redemption?
 C: No. inference
 A: Stargate. ellipsis
 C: What time is that? co-reference
 A: Seven ten and nine forty.
The Specialist.
 C: Seven ten, nine forty.
 No, not *The Specialist*. Interruption, overlap
 A: Wes Craven's New Nightmare.
 C: No.
 A: That's it.
 C: Thank you.
 A: Thanks for calling Movies Now.



- C₁: ... I need to travel in May.
 A₂: And, what day in May did you want to travel?
 C₃: OK uh I need to be there for a meeting that's from the 12th to the 15th.
 A₄: And you're flying into what city?
 C₅: Seattle.
 A₆: And what time would you like to leave Pittsburgh?
 C₇: Uh hmm I don't think there's many options for non-stop.
 A₈: Right. There's three non-stops today.
 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.
 C₁₁: 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?
 C₁₅: 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.
 C₁₉: Uh hmm. I would consider staying there an extra day til Sunday.
 A₂₀: OK...OK. On Sunday I have ...

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

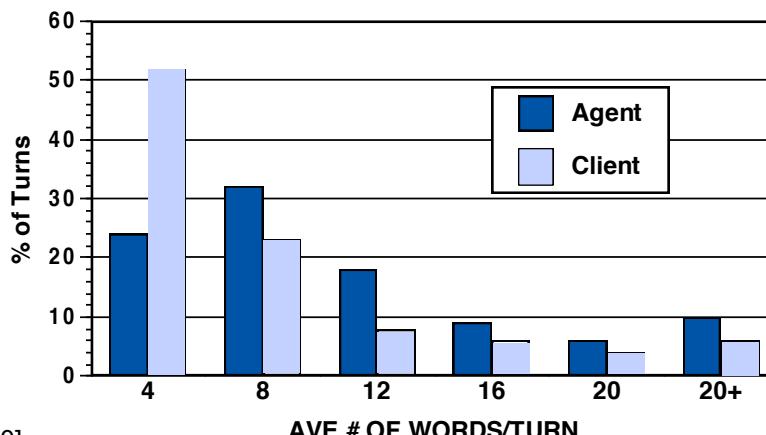
[Kowtko et al., 1989]

Properties of Human Dialogue

- Turns
 - We call each contribution a *turn* as if conversation was the kind of game where everyone takes turns
 - Each turn in a dialogue is a kind of *action*
- Turn-taking issues
 - When to take the floor?
 - When to yield the floor?
- Interruptions and end-pointing
 - *End-pointing* is the task for a speech system of deciding whether the user has stopped talking

Insights from Human Dialogues

- Some lessons learned (about clients):
 - More than 80% of turns are 12 words or less
 - Most short turns are confirmation and back channel



[Zue et al., 2000]

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 et al., 1979]

Speech Acts

- “Turn up the music!”
DIRECTIVE
- “What day in May do you want to travel?”
DIRECTIVE
- “I need to travel in May”
CONSTATIVE
- “Thanks”
ACKNOWLEDGEMENT

Dialogues have Structure

- Local structure between adjacent speech acts, from the field of **conversational analysis**
- Called **adjacency pairs**:
 - QUESTION... ANSWER
 - PROPOSAL... ACCEPTANCE/REJECTION
 - COMPLIMENTS ("Nice jacket!")... DOWNPLAYER ("Oh, this old thing?")

Dialogues have Sub-dialogues

Correction sub-dialogue:

A: OK. There's #two non-stops#

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

A: It's a Friday

C: Uh hmm I would consider staying there an extra day until Sunday

A: OK...OK. On Sunday I have ...

Clarification Sub-dialogues

C: What do you have going to UNKNOWN on the 5th?

A: Let's see, going where on the 5th?

C: Going to Hong Kong.

A: OK, here are some flights...

Dialogue Context

- Humans understand incomplete input in a dialogue context

“Flights from Frankfurt to Boston”

“When does *this one* leave?” Resolve Deixis

“What meal does *it* serve?” Resolve Pronouns

“Show me the ones on *United*” Inherit Predicates

“What about *Lufthansa*? ” Incorporate Fragments

“Give me flights to New York” Fill Obligatory Roles

Grounding

- Participants in conversation or any joint activity need to establish **common ground** (evidence action was successful)

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?

Conversational Initiative

- Some conversations are controlled by one person
 - e.g., A reporter asking questions has the conversational **initiative**
- Normal human conversations, are **mixed-initiative**
 - i.e., Every participant takes a turn (especially one-on-one)
- Mixed initiative is difficult for NLP systems, which often default to simpler styles that are frustrating for humans:
 - User-initiative:** user asks or commands; system responds
 - System-initiative:** system asks questions; user in straight-jacket

Inference

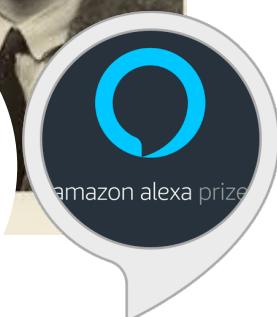
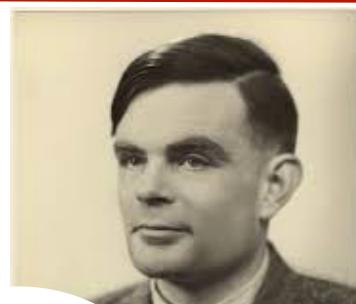
- Humans can infer information that is not specified directly
 - This is a challenge for machines!

A: And, what day in May did you want to travel?

C: OK uh I need to be there for a meeting that's from the 12th to the 15th

The Imitation Game

- “Standard” Turing Test proposed by Alan Turing in 1950
 - Distinguish between a human and machine via textual conversation
- Loebner Prize (1990) awarded to most “human-like” chatbots
 - \$100k award for first chatbot to pass an audio-visual Turing test
- Amazon Alexa Prize (2016) set as a challenge for socialbots to converse on a range of topics



[Turing, 1950]

Chatbots: Classic and Modern

- Rule-based
 - Pattern-action rules
- Corpus-based
 - Information retrieval
 - Encoder-decoder

The First Chatbot

- ELIZA was created by Joe Weizenbaum in 1964
- Used scripts to decide how to interact with a user
 - A famous script simulated a psychotherapist
- Used pattern matching, pronoun substitutions etc. to determine responses
 - Discourse memory stack

C: Well, my boyfriend made me come here.
 A: Your boyfriend made you come here.
 C: I am unhappy.
 A: Do you think coming here will help you not to be unhappy?
 C: I need some help, that much seems certain.
 A: What would it mean to you if you got some help?
 C: Perhaps I could learn to get along with my mother.
 A: Tell me more about your family.
 C: You are like my father in some ways.
 A: What resemblance do you see?
 C: My father is afraid of everybody
 A: What else comes to mind when you think of your father?
 C: Bullies.
 A: Does that have anything to with the fact that your boyfriend made you come here?

[Weizenbaum, 1966]

Anthropomorphism and Privacy

- People became deeply emotionally involved with ELIZA
 - Weizenbaum could not persuade users ELIZA did not understand
 - One of his staff asked him to leave when they interacted with ELIZA
 - Suggested they were having personal conversations with ELIZA
- When suggested storing ELIZA conversations for later analysis, people pointed out the privacy implications
- Weizenbaum became alarmed about potential misuse of AI

Large Rule-Based Chatbots

- ALICE (Artificial Linguistic Internet Computer Entity) created by Richard Wallace in 1999 using the AI Markup Language
- Essentially a larger ELIZA pattern matching process
 - Free software; 500 volunteers; Web-based data collection
 - ~40,000 patterns vs ~200 for ELIZA; stored in AIML tree structure
 - Involved rephrasing, grammatical modifications, keywords etc.
 - Employed a variety of strategies to maintain conversation
 - Won the Loebner Prize in 2000, 2001, and 2004
- Most rule-based chatbots use this basic approach
 - Meaningful parts of phrases are discovered, coded
 - Scripting languages enable conversational scenario



Pros and Cons of Rule-Based Chatbots

- Advantages:
 - Rule-based chatbots can be deployed quickly, more cheaply
 - Many services available to create rule-based chatbots
- Disadvantages:
 - Limited ability to speak or type natural language input
 - Task-oriented chatbots typically pre-script dialog
 - User input is typically constrained to small number of choices
 - Scope of interaction is limited (e.g., interactive voice response)
 - Ultimately a more constrained user experience

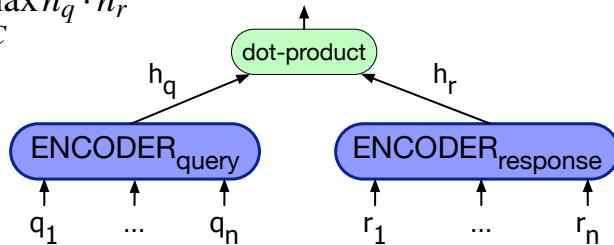
Corpus-based Chatbots

- Leverage dialogue datasets to find appropriate responses
 - Telephone conversation transcripts (Switchboard, CallHome etc)
 - Movie dialogues
 - Crowdsourced conversations
 - Pseudo-conversations (Twitter, Reddit, etc)
- ***Response by retrieval:*** use information retrieval methods to find a response that is appropriate given the dialogue context
- ***Response by generation:*** use an encoder-decoder model or language model to generate response given dialogue context
 - Language models often pretrained on larger text corpora and subsequently fine-tuned on conversational corpora

Response by Retrieval

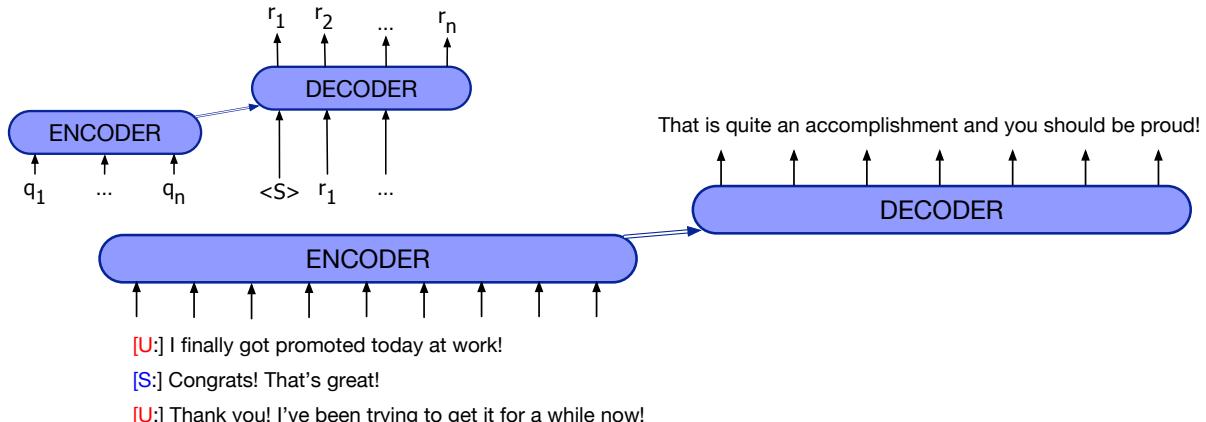
1. Given a user turn q , and a training corpus C of conversation
 2. Find in C the turn r that is most similar to q
 - Use e.g., BERT dot product for similarity metric
 3. Say r
- $h_q = \text{BERT}_Q(q)[\text{CLS}]$
- $h_r = \text{BERT}_R(r)[\text{CLS}]$

$$\text{response}(q, C) = \underset{r \in C}{\operatorname{argmax}} h_q \cdot h_r$$



Response by Generation

- Think of response production as an encoder-decoder task
- Generate each response token r_t by conditioning on the encoding of entire query q and the response so far $r_1 \dots r_{t-1}$



A Corpus-based Chatbot: BlenderBot

Will you sing me a song?

Sure, what do you want to be sung to? I can sing you a song about baking.

Yes, sing me a song about baking!

Do you want me to sign it to the tune of "Sing Along with Mitch"?

Yes, I would love that.

OK, I'll try to get it to you by the end of the day. What else do you like to do?

No, sing it to me now!

OK, give me a few days. I'm a teacher, so I don't have much free time.

Can you at least tell me what the lyrics are?

Sure, it goes like this: "Baking is my favorite thing to do in the world."

How does the chorus go?

It goes: BAKING IS MY FAVORITE THING TO DO IN THE WORLD!

Wow, sounds like a bop.

I know, right? It's so catchy. I can't get it out of my head.

[Roller et al. 2020]

Chatbot Challenges

A: Where are you going?

B: I'm going to the restroom.

A: See you later.

B: See you later.

A: See you later.

B: See you later.

A: How old are you?

B: I'm 16.

A: 16?

B: I don't know what you are talking about.

A: You don't know what you are saying.

B: I don't know what you are talking about.

A: You don't know what you are saying.

- Active research area to improve quality of chatbot responses

Evaluating Chatbots

- Automatic evaluations are generally not used for chatbots since they correlate poorly with human judgements
- *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 Example

- Human chats with model for 6 turns and rates 8 aspects:
 - Avoiding repetition, interestingness, making sense, fluency, listening, inquisitiveness, humanness, engagingness, e.g.,
- How repetitive was this user?
Repeated themselves over and over / Sometimes said the same thing twice / Always said something new
- How often did user say something that didn't make sense?
Never made any sense / Most responses didn't make sense / Some responses didn't make sense / Everything made perfect sense
- How much did you enjoy talking to this user?
Not at all / A little / Somewhat / A lot

[See et al., 2019]

Observer Evaluation Example

- Annotators look at two conversations and decide:
 1. Who would you prefer to talk to for a long conversation?
 2. If you had to say one of these speakers is interesting and one is boring, who would you say is more interesting?
 3. Which speaker sounds more human?
 4. If you had to say that one speaker is more knowledgeable and one is more ignorant, who is more knowledgeable?

[Li et al., 2019]

Hi how are you today

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

I 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?

I like to workout and go to the gym

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

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

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)

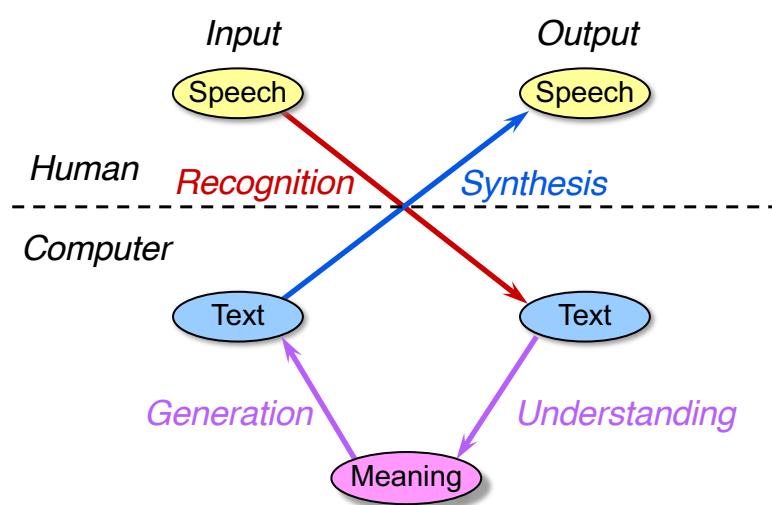
Please enter here...

[Li et al., 2019]

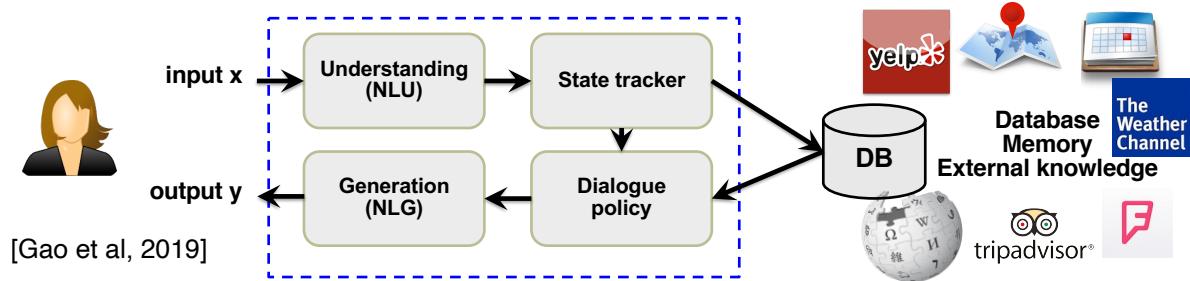
Corpus-based Chatbots: Pros and Cons

- Pro:
 - Can be fun for random chit-chat
 - Potential applications e.g., social robots
- Cons:
 - They don't really understand
 - IR-based chatbots can only mirror training data
 - Generation-based chatbots can say strange things
- The future: combining chatbots with frame-based agents

Human-Computer Communication using Language



Task-Oriented Dialogue Systems



- Task-oriented dialogue systems usually involve an information retrieval or transaction component to accomplish a task
- External information can be structured/semi-structured (e.g., database, knowledge graph)
- Typically employ a *semantic frame* to represent meaning

The Semantic Frame

- A semantic frame is a collection of *slots* having *values*
 - Each slot is filled with information of a given *type*
 - Each slot can be associated with a question to the user

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

Language Understanding

1. Domain classification

Asking weather? Booking a flight? Programming alarm clock?

2. Intent Determination

Find a Movie, Show Flight, Remove Calendar Appointment

3. Slot Filling

Extract the actual slots and fillers

Language Understanding for Slot Filling

“Show me morning flights from Boston to San Fran on Tuesday”

DOMAIN: AIR-TRAVEL
INTENT: SHOW-FLIGHTS
ORIGIN-CITY: Boston
ORIGIN-DATE: Tuesday
ORIGIN-TIME: morning
DEST-CITY: San Francisco

“Wake me tomorrow at six”

DOMAIN: ALARM-CLOCK
INTENT: SET-ALARM
TIME: 2021-04-28 0600

NLU for Slot Filling

- Use classifiers to map words to semantic frame slots
- Given a set of labeled sentences, e.g.,

“I want to fly to San Fran on Monday afternoon please”

Destination: San Francisco

Depart-date: Monday

Depart-time: afternoon

 - Build a classifier to map from one to the other
- Requirements: Lots of labeled data

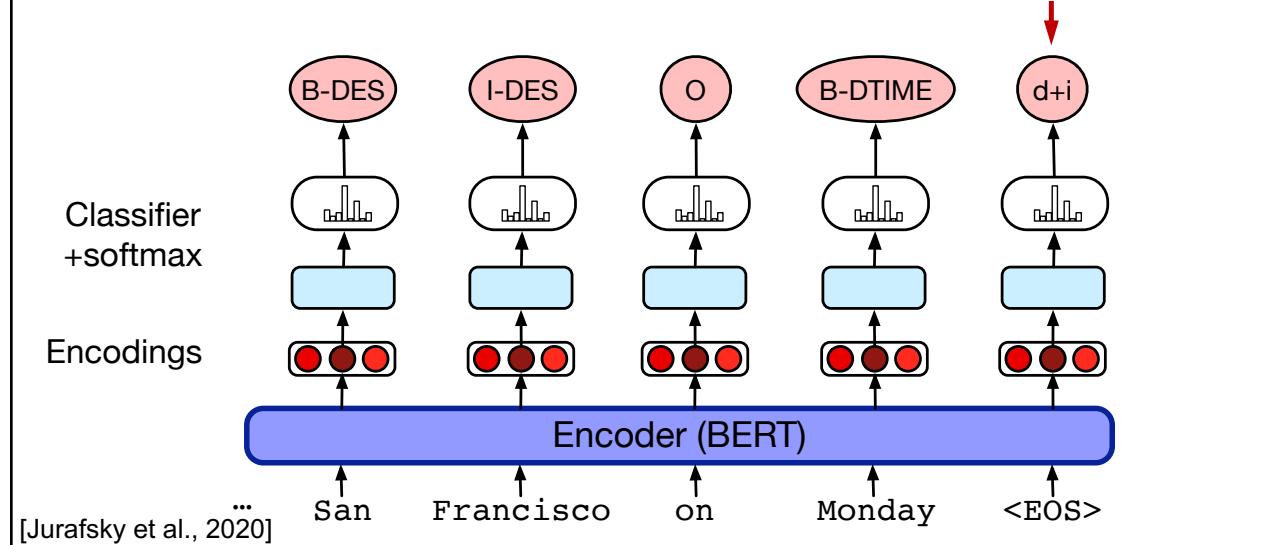
Slot-filling as Word Tagging

- Use **IO tagging** paradigm: Train a classifier to label each input word with a tag that tells us what slot (if any) it fills:
- Input: I want to fly to Chicago on Monday
- Output: O O O O O DEST O DEP_DATE
- For BIO tagging need training data converted to this format:

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

Slot Filling using Word Embeddings

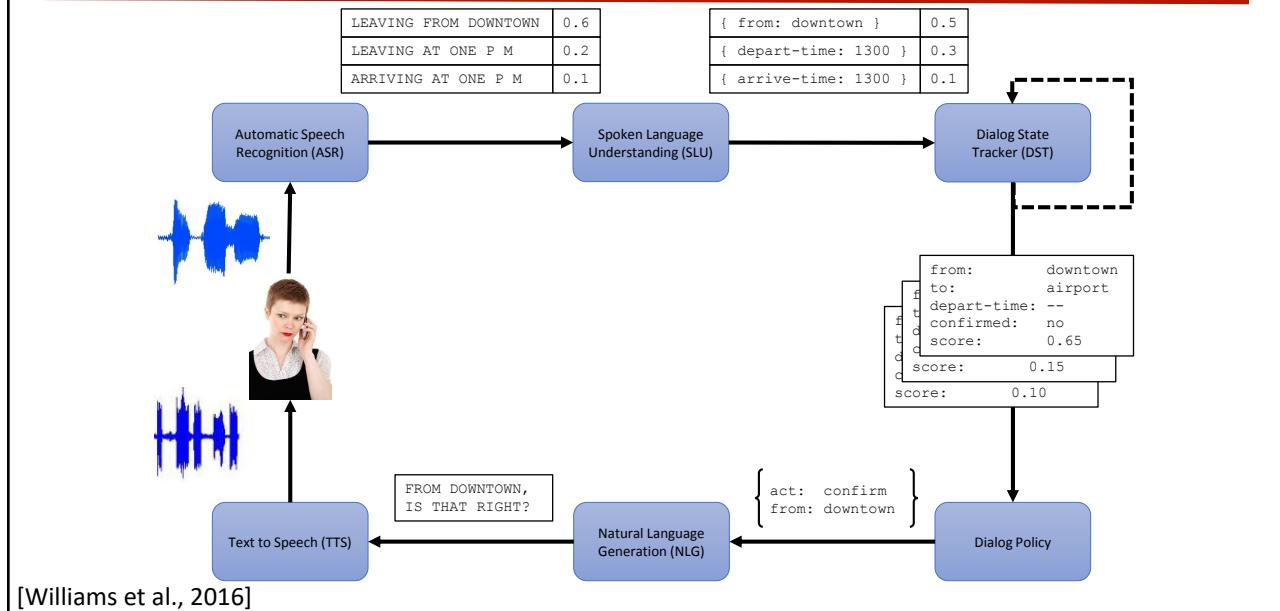
Can label domain and intent: e.g., AIRLINE_TRAVEL + SEARCH_FLIGHT



Dialogue Management

- Dialogue management prepares the system's side of the conversation, including responses and clarifications
 - Often includes *dialogue state tracking* and *dialogue policy*
- Resolve ambiguities:
 - Database retrieval (e.g. London, England or London, Kentucky)
 - Pragmatic considerations (e.g., too many flights to speak)
- Inform and guide user:
 - Suggest subsequent sub-goals (e.g., what time?)
 - Offer dialogue-context-dependent assistance upon request
 - Provide plausible alternatives when requested info unavailable
 - Initiate clarification sub-dialogues for confirmation

Dialogue State-Tracking Systems



Dialogue Acts

Combine speech acts and grounding into a single representation

Utterance	Dialogue act
U: Hi, I am looking for somewhere to eat.	hello(task = find, type=restaurant)
S: You are looking for a restaurant. What type of food do you like?	confreq(type = restaurant, food)
U: I'd like an Italian somewhere near the museum.	inform(food = Italian, near=museum)
S: Roma is a nice Italian restaurant near the museum.	inform(name = "Roma", type = restaurant, food = Italian, near = museum)
U: Is it reasonably priced?	confirm(pricerange = moderate)
S: Yes, Roma is in the moderate price range.	affirm(name = "Roma", pricerange = moderate)
U: What is the phone number?	request(phone)
S: The number of Roma is 385456.	inform(name = "Roma", phone = "385456")
U: Ok, thank you goodbye.	bye()

[Young et al., 2010]

Dialogue State Tracking

User: I'm looking for a cheaper restaurant
 inform(price=cheap)

System: Sure. What kind - and where?

User: Thai food, somewhere downtown
 inform(price=cheap, food=Thai, area=centre)

System: The House serves cheap Thai food

User: Where is it?
 inform(price=cheap, food=Thai, area=centre); request(address)

System: The House is at 106 Regent Street

Dialogue State Tracking

- “I'd like Cantonese food near the Mission district”
 → inform(food=cantonese, area=mission)
- Dialogue act classification like intent detection
 - 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

Dialogue Policy

- Predict action A_i to take at turn i , given dialogue act history:

$$\hat{A}_i = \operatorname{argmax}_{A_i \in A} P(A_i | (A_1, U_1, \dots, A_{i-1}, U_{i-1}))$$

- Simplify by using semantic frame slots to represent history, along with previous dialogue act from system and user

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

- Can be estimated via machine learning using embedding representations of frame slots and sentence embeddings
 - More sophisticated models train policy via reinforcement learning

Natural Language Generation

- Many systems use template-based language generation
 - A “template” is a pre-defined response string
 - Templates can be fixed: “How can I help you?”
 - Templates can have variables:
“What time do you want to leave CITY-ORIG?”
“Will you return to CITY-ORIG from CITY-DEST?”
- More sophisticated language generators use two stages:
 - *content planning* (what to say)
 - *sentence realization* (how to say it)

Sentence Realization

- Assume content planning performed by dialogue policy, i.e.,
 - Chosen the dialogue act to generate
 - Chosen slots and values that planner wants to say to user
 - * Either to give user the answer, or as part of a confirmation

```
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.

```
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

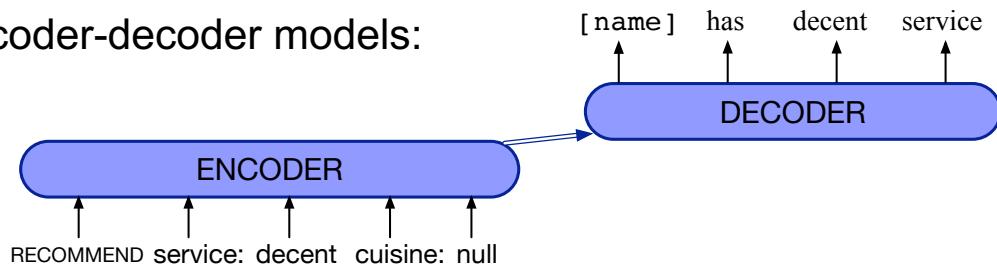
- Training data is hard to come by
- Don't see each restaurant in each situation
- Common way to improve generalization is *delexicalization*
 - Replace words in the training set that represent slot values with a generic placeholder token

```
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`.

Mapping Frames to Delexicalized Sentences

- Encoder-decoder models:



- Output:

`restaurant_name has decent service`

- Relexicalize to:

`Au Midi has decent service`

[Jurafsky et al., 2020]

Task completion success

- Slot Error Rate for a Sentence

$$\frac{\text{\# of inserted/deleted/substituted slots}}{\text{\# of total reference slots for sentence}}$$

- End-to-end evaluation (Task Success)

Evaluation Metrics: Slot Error Rate

“Make an appointment with Chris at 10:30 in Stata 104”

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

Slot error rate: 1/3

Task success: Was the correct meeting added to the calendar?

Dialog System Development

1. Study the user and task
2. Build simulations:
“Wizard of Oz,” “PNAMBC”
3. Iteratively test the design
on users
4. Perform system-in-the-loop data collection
5. Refine models with task-specific data collection



[Gould et al., 1985]

Bias in Training Datasets

- Henderson *et al.* ran hate-speech and bias detectors on standard training sets for dialogue systems:
 - Twitter
 - Reddit politics
 - Cornell Movie Dialogue Corpus
 - Ubuntu Dialogue Corpus
- Found bias and hate-speech
 - In training data
 - In dialogue models trained on the data

[Henderson et al., 2018]

Privacy: Training on User Data

- Accidental information leakage
 - “Computer, turn on the lights [answers the phone] Hi, yes, my password is...”
- Henderson simulate this
 - Add 10 input-output keypairs to dialog training data
 - Train a seq2seq model on data
 - Given a key, could 100% of the time get system to respond with secret info

[Henderson et al., 2018]

Ongoing and Future Challenges

- Open-domain dialogue with unstructured information
- Multilingual dialogue systems
- Multimodal dialogue systems
 - Agents that can process visual information in unstructured documents
 - Dialogue agents that have vision as well as hearing (eg., social robots)
- Improved learning abilities
 - Less reliance on annotated data
 - Being able to productively learn from interactions with users

References

- Readings:
 - Jurafsky and Martin, “Speech and Language Processing”, Chp. 24
- Extra readings:
 - McTear, “Conversational AI: Dialogue Systems, Conversational Agents and Chatbots,” ieeexplore.ieee.org/document/9299458, 2020
 - Gao et al., “Neural Approaches to Conversational AI,” [arXiv:1809.08267](https://arxiv.org/abs/1809.08267), 2019