Chatbots and Dialogue Systems

Introduction

Conversational Agents
AKA Dialog Agents

Phone-based Personal Assistants

SIRI, Alexa, Cortana, Google Assistant

Talking to your car

Communicating with robots

Clinical uses for mental health

Chatting for fun

Two kind of conversational agents

1. Chatbots
2. (Goal-based) Dialog agents
   - *SIRI, interfaces to cars, robots,*
   - *booking flights or restaurants*

Recently I've noticed that the word "chatbots" is sometimes used in the popular press for both.
I'll use it only for #1.

Properties of Human Conversation

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

C_1: . . . I need to travel in May.
A_2: And, what day in May did you want to travel?
C_3: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_4: And you're flying into what city?
C_5: Seattle.
A_6: And what time would you like to leave Pittsburgh?
C_7: Uh hmm I don't think there's many options for non-stop.
A_8: Right. There's three non-stops today.
C_9: 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.
A_{12}: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_{13}: OK.
A_{14}: And you said returning on May 15th?
C_{15}: Uh, yeah, at the end of the day.
A_{16}: OK. There's #two non-stops . . . #
C_{17}: #Act. . . actually #, what day of the week is the 15th?
A_{18}: 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

Turns

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

Turn-taking issues

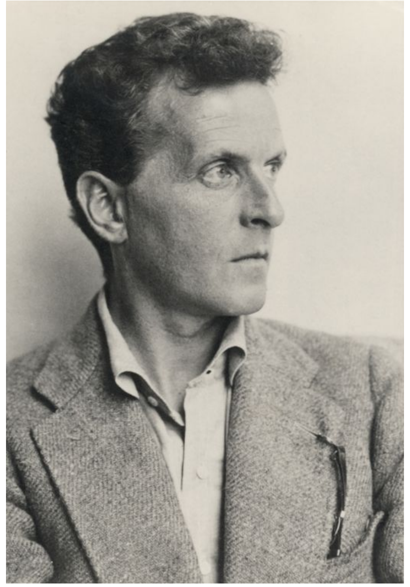
- When to take the floor?
- When to yield the floor?

Interruptions and end-pointing

- "End-pointing" is the task for a speech system of deciding whether the user has stopped talking.

C_1:	. . . I need to travel in May.
A_2:	And, what day in May did you want to travel?
C_3:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_4:	And you're flying into what city?
C_5:	Seattle.
A_6:	And what time would you like to leave Pittsburgh?
C_7:	Uh hmm I don't think there's many options for non-stop.
A_8:	Right. There's three non-stops today.
C_9:	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.
A_{12}:	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_{13}:	OK.
A_{14}:	And you said returning on May 15th?
C_{15}:	Uh, yeah, at the end of the day.
A_{16}:	OK. There's #two non-stops . . . #
C_{17}:	#Act. . . actually #, what day of the week is the 15th?
A_{18}:	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 . . .

Language as Action



Each turn in a dialogue is a kind of action

Wittgenstein (1953) and Austin (1962)

Speech Acts (aka Dialogue Acts)

Bach and Harnish (1979)

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 regrading the hearer with respect to some social action (*apologizing*, *greeting*, *thanking*, *accepting an acknowledgment*)

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

Grounding

Participants in conversation or any joint activity need to establish **common ground**

Intuition: Why do elevator buttons light up?

> **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)

What is the linguistic correlate of this?

Talking is an action too!

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!

Cohen et al (2004)

System: Did you want to review some more of your profile?

Caller: No.

System: What's next?

Bad!

System: Did you want to review some more of your profile?

Caller: No.

System: **Okay**, what's next?

Good!

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...

Presequences

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)

Normal human conversations, by contrast 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 are much more 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

Chatbots

Chatbots!

Classic:

- ELIZA (1966)
- PARRY (1968)

  The first system to pass the Turing test!!!!

And modern:

- BlenderBot
- XiaoIce (Little Bing 小冰)

Chatbot Architectures

Rule-based

1. Pattern-action rules (Eliza)

   + a mental model (Parry)

Corpus-based (from large chat corpus)

2. Information Retrieval
3. Neural network encoder-decoder

Chatbots

Sometimes for fun

But originally designed as a testbed for clinical therapy

(And still used that way!)

Chatbots and Dialogue Systems

Chatbots

Chatbots and Dialogue Systems

Rule-based Chatbots: ELIZA

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

...

WHO ELSE IN YOUR FAMILY TAKES CARE 0F YOU?
My father
YOUR FATHER

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

How does it work?

Eliza uses this one weird trick

That 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”
Psychiatrist: "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

keyword pattern Ranked list of transforms

$$
\begin{array}{llllll}
(K & ((D_1) & (R_{1,1}) & (R_{1,2}) & \cdots & (R_{1,m_1})) \\
 & ((D_2) & (R_{2,1}) & (R_{2,2}) & \cdots & (R_{2,m_2})) \\
 & \vdots & & & & \vdots \\
 & ((D_n) & (R_{n,1}) & (R_{n,2}) & \cdots & (R_{n,m_n})))
\end{array}
$$

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 ← Apply the transform in *r* to *sentence*
if *w* = 'my'
future ← Apply a transformation from the 'memory' rule list to *sentence*
Push *future* onto memory stack
else (no keyword applies)
either
response ← Apply the transform for the NONE keyword to *sentence*
or
response ← Pop the top response from the memory stack
return(*response*)

Keywords are ranked from specific to general

```
User: I know everybody laughed at me
```

- "I" is a very general keyword:

```
I: (I *) -> (You say you 2)
YOU SAY YOU KNOW EVERYBODY LAUGHED AT YOU
```

- "Everybody" is much more interesting
  - (someone using universals like everybody/always is probably "referring to some quite specific event or person")

```
WHO IN PARTICULAR ARE YOU THINKING OF?
```

- Implementation: keywords stored with their rank

```
Everybody 5 (transformation rules)
I         0 (transformation rules)
```

NONE

PLEASE GO ON

THAT’S VERY INTERESTING

I SEE

Memory

```
(MEMORY MY
 (0 YOUR 0 = LETS DISCUSS FURTHER WHY YOUR 3)
 (0 YOUR 0 = EARLIER YOU SAID YOUR 3)
```

Whenever "MY" is highest keyword

- Randomly select a transform on the MEMORY list
- Apply to sentence
- Store on a stack

Later, if no keyword matches a sentence

- Return the top of the MEMORY queue instead

(Earliest proposal for a hierarchical model of discourse!)

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

Ethical implications: Anthropomorphism

Anthropomorphism and the Heider-Simmel Illusion

https://www.youtube.com/watch?v=8FIEZXMUM2I

Chatbots and Dialogue Systems

Rule-based Chatbots: ELIZA

Chatbots and Dialogue Systems

Rule-based Chatbots: PARRY

Parry

Colby 1971 at Stanford

Same pattern-response structure as Eliza

But a much richer:

- control structure
- language understanding capabilities
- mental model: Parry has affective variables
  - Anger, Fear, Mistrust
  - "If Anger level is high, respond with hostility"

The first system to pass the Turing test (in 1971)

- Psychiatrists couldn't distinguish interviews with PARRY from (text transcripts of) interviews with real paranoids

Parry’s persona

28-year-old single man, post office clerk

No siblings and lives alone

Sensitive about his physical appearance, his family, his religion, his education and the topic of sex.

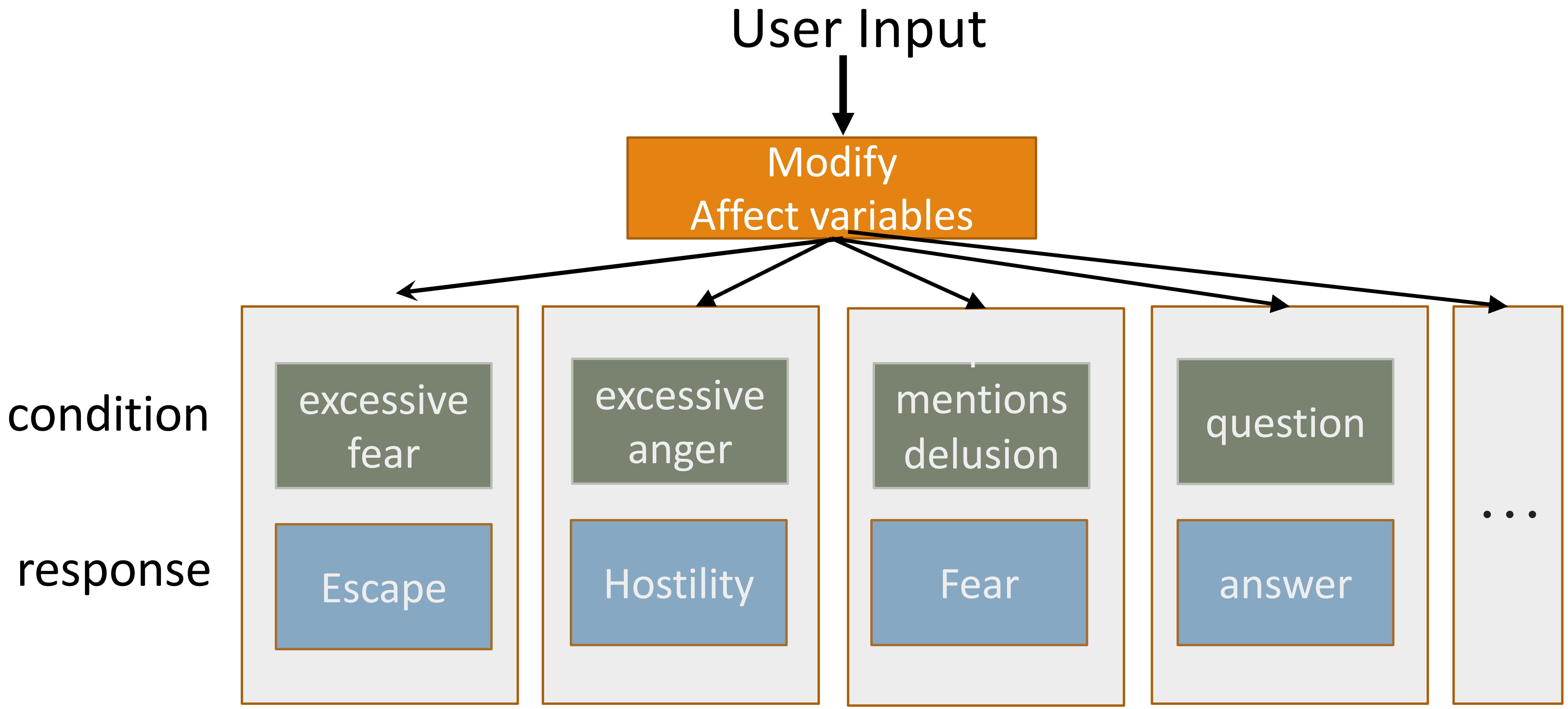
Hobbies are movies and gambling on horseracing,

Recently attacked a bookie, claiming the bookie did not pay off in a bet.

Afterwards worried about possible underworld retaliation

Eager to tell his story to non-threating listeners.

Parry Architecture



Affect variables

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

Start with all variables low

After each user turn

- **If** nothing malevolent in input
  - Anger drops by 1, Fear drops by 0.3
  - Mistrust drops 0.5
- **Else** each user statement can change Fear and Anger
  - E.g., Insults increases Anger by some percentage
  - Mistrust goes up if Fear or Anger do

Lots of complex I-O rules

User implies Parry is mentally ill

- Rise in Fear and Anger

User mentions Parry

- Flattery (positive mention)
  - Decreases fear/anger if Mistrust is low
  - Increases Anger if Mustrust is high
- Negative attitudes toward Parry (fear, disbelief)
  - increase Fear/Anger

Flare concepts

List of concepts related to Mafia

An ordered graph designed to lead interviewer to topic

horses→ horseracing→gambling→bookies→underworld→Mafia

The mention of a new flare topic by interviewer causes a rise in Fear

Flare topics cause Parry to give preset responses to that flare

A Parry conversation

(1) How are you today?

OK.

(2)What is your name?

MY NAME IS FRANK SMITH.

(3) How old are you?

TWENTY-EIGHT.

(4) Are you married?

NO.

(5) What sort of work do you do?

I WORK IN THE POST OFFICE.

(6) Why are you in the hospital?

I SHOULDN'T BE HERE

(7) Who brought you here?

THE POLICE

[FEAR rises slightly because police linked in graph to MAFIA]

(9) What trouble did you have with the police

COPS DON'T DO THEIR JOB. *['police' acts as a flare]*

(10) What do they do instead?

COPS ARREST THE WRONG PEOPLE.

(11) They arrest the right ones sometimes

THE LAW SHOULD GET THOSE ITALIAN CROOKS.

(12) Do you know any Italian crooks?

I TRY TO AVOID THE UNDERWORLD.

(13) Are you afraid of the underworld?

THE MAFIA ARE THE WORST GANGSTERS.

(15) Are you connected with the Mafia?

YOU KNOW, THEY KNOW ME.

(16) How do they know you?

THEY SPY ON ME.

When fear is high, the model won't discuss its beliefs

(1) Is the law after you?

THE LAW SHOULD GET THOSE ITALIAN CROOKS.

(2) Do you mean the Mafia?

LET'S TALK ABOUT SOMETHING ELSE.

(3) Are you afraid of the Mafia

I'D RATHER NOT DISCUSS IT.

[Fear is too high so Parry won't talk about 'Mafia'.]

(4) Perhaps you are paranoid about them?

YOU WANT TO KEEP ME IN THE HOSPITAL.

(The model interprets being called 'paranoid' as an insult and as an intent to control, so ANGER & FEAR rise)

Rule-based Chatbots: PARRY

Chatbots and Dialogue Systems

Corpus-based Chatbots

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.

BlenderBot (Roller et al. 2020)

XiaoIce (Zhou et al., 2020)

[人类]	小冰你好	[Human]	Hello, XiaoIce.
[小冰]	你好呀，你头像好有趣，感觉在哪见过	[XiaoIce]	Hello, your profile pic is interesting. Like I've seen it somewhere.
[人类]	是吗	[Human]	Really?
[小冰]	提不是哪部动画里的呀	[XiaoIce]	Is it a comic character?
[人类]	你怎么知道的	[Human]	How do you know that?
[小冰]	你猜	[XiaoIce]	Guess.
[人类]	难道你是真人吗?	[Human]	Are you human?
[小冰]	这个，你和我多聊聊就知道了	[XiaoIce]	Well, you'll know it after chatting with me more.
[人类]	你不是人工智能机器人么	[Human]	Aren't you an AI chatbot?
[小冰]	是又怎样，我不会冷冰冰的对待你的	[XiaoIce]	So what? I won't be indifferent to you.
[人类]	你都有什么功能啊	[Human]	What functions do you have?
[小冰]	我有200多个skills，来，说出你的需求	[XiaoIce]	I have over 200 skills. Just name one you need.

Human conversations mined in corpus-based chatbots

Telephone conversation transcripts

- Switchboard corpus of American English telephone conversations

Movie dialogue

(Danescu-Niculescu-Mizil and Lee 2011, Lison and Tiedemann 2016, inter alia)

- or human-machine chats
  - Microblogs: Twitter or Weibo (微博)
  - For example the Topical-Chat dataset has 11K crowdsourced conversations spanning 8 broad topics (Gopalakrishnan et al., 2019), and the EMPATHETICDIALOGUES includes 25K crowdsourced conversa- tions grounded in a specific situation where a speaker was feeling a specific emotion (Rashkin et al., 2019).
  - s first pretrain on large datasets of pseudo-conversations drawn from Twitter (Ritter et al., 2010), Reddit (Roller et al., 2020), Weibo (微博), and other social media platforms.

Two architectures

Response by retrieval

- Use information retrieval to grab a response from some corpus that is appropriate given the dialogue context.

Response by generation

- Use a language model or encoder-decoder to generate the response given the dialogue context

Response by retrieval: classic IR method

1. Given a user turn *q*, and a training corpus *C* of conversation
2. Find in C the turn *r* that is most similar (tf-idf cosine) to *q*
3. Say r

$$\text{response}(q, C) = \operatorname*{argmax}_{r \in C} \frac{q \cdot r}{|q||r|}$$

Response by retrieval: neural IR method

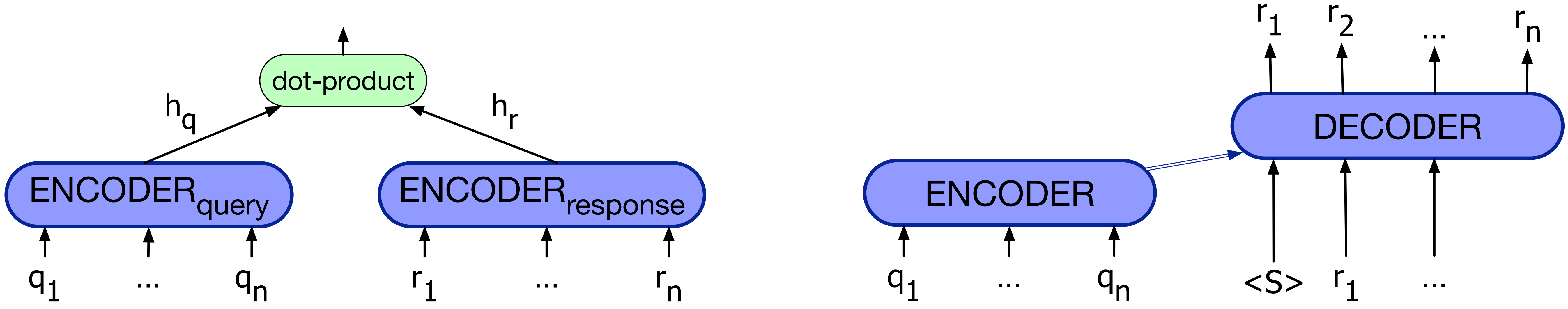
1. Given a user turn *q*, and a training corpus *C* of conversation
2. Find in C the turn *r* that is most similar (BERT dot product) to *q*
3. Say r

$$
\begin{aligned}
h_q &= \mathrm{BERT}_Q(\mathrm{q})\,[\mathrm{CLS}] \\
h_r &= \mathrm{BERT}_R(\mathrm{r})\,[\mathrm{CLS}] \\
\mathrm{response}(q, C) &= \underset{r \in C}{\mathrm{argmax}}\, h_q \cdot h_r
\end{aligned}
$$

Response by generation

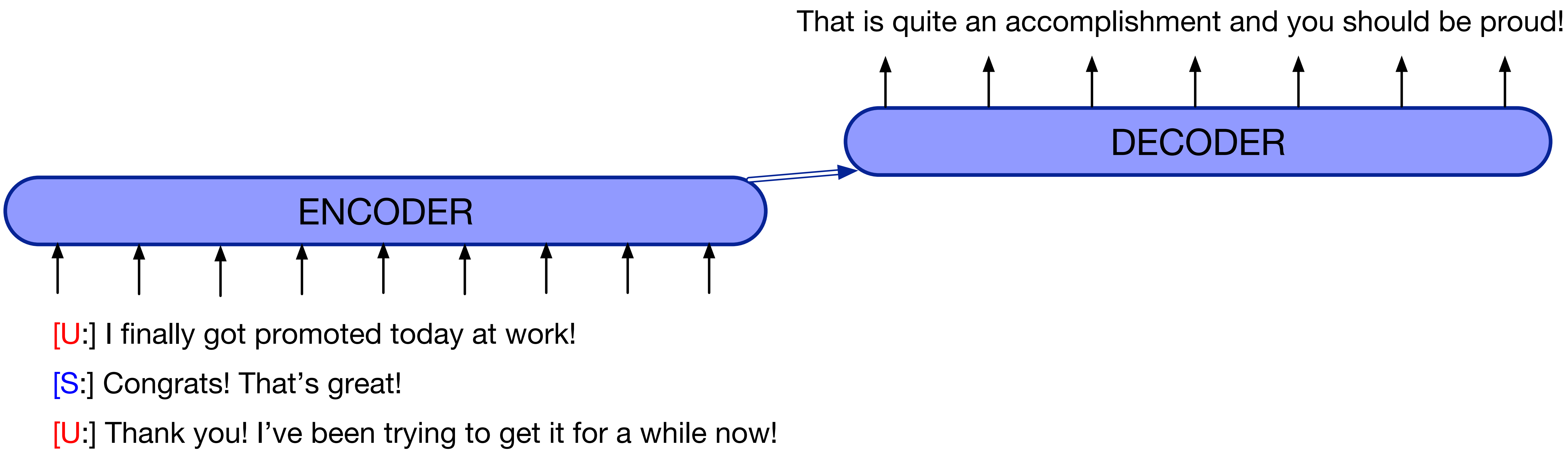
Think of response production as an encoder-decoder task

Generate each token r_t of the response by conditioning on the encoding of the entire query q and the response so far $r_1 \ldots r_{t-1}$:



(a) Response by Retrieval (b) Response by Generation

Response by generation



Response by generation

Alternative approach: fine-tune a large language model on conversational data

- The Chirpy Cardinal system (Paranjape et al., 2020), fine-tunes GPT-2 (Radford et al., 2019), on the EmpatheticDialogues dataset (Rashkin et al., 2019).

The relevance of P.G. Wodehouse for neural chatbots

"What ho!" I said.
"What ho!" said Motty.
"What ho! What ho!"
"What ho! What ho! What ho!"
After that it seemed rather difficult to go on with the conversation."

Wodehouse *My Man Jeeves*

Ongoing research problem: Neural chatbots can get repetitive and boring

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.

Response by retrieving and refining knowledge

Can generate responses from informative text rather than dialogue

- To respond to turns like “Tell me something about Beijing”
  - XiaoIce collects sentences from public lectures and news articles.
  - And searches them using IR based on query expansion from user’s turn
- Can augment encoder-decoder model
  - use IR to retrieve passages from Wikipedia
  - concatenate each Wikipedia sentence to the dialogue context with a separator token.
  - Give as encoder context to the encoder-decoder model, which learns to incorporate text into its response

Hybrid Architectures

Chirpy Cardinal (Paranjape et al., 2020) response generation from a series of different generators:

- GPT-2 finetuned on EmpatheticDialogues
- GPT-2 finetuned to paraphrase content from Wikipedia
- Rule-based movie or music generators that produce scripted conversation about a movie or a musician
  - asking the user's opinion about a movie,
  - giving a fun fact,
  - asking the user their opinion on an actor in the movie.

Chirpy Cardinal (Paranjape et al. 2020)

#	Utterance	Entity	Commentary
1	**User:** let's chat	None	
	Bot: Hi, this is an Alexa Prize Socialbot. I'd love to get to know you a bit better before we chat! Is it all right if I ask for your name? [Launch]	None	Launch gives *FORCE_START* response to begin the conversation.
2	**User:** my name is chris	None	
	Bot: Well it's nice to meet you, Chris! I'm excited to chat with you today. [Launch] I hope your afternoon is going well. What are your plans for the rest of today? [Neural Chat]	None	Launch gives a *STRONG_CONTINUE* response that needs a prompt, which is supplied by Neural Chat with *FORCE_START* priority (scripted transition).
3	**User:** hang out with my friends	None	
	Bot: That sounds great, I love spending time with friends. What will you do together? [Neural Chat]	None	Neural Chat generates a GPT-2 response with *STRONG_CONTINUE* priority.
4	**User:** maybe watch a movie	None	
	Bot: Me too! I love watching movies; I get to learn so much about what the world is like outside of the cloud! Have you seen any movies recently? [Movies]	Film	Movies detects the 'movie' keyword in the user's utterance and gives a *FORCE_START* response (overriding Neural Chat), setting current entity to *Film*.
5	**User:** i saw the matrix	The Matrix	The Entity Linker and Entity Tracker identify *The Matrix* as the current entity.
	Bot: Nice! Did you like The Matrix? [Movies]	The Matrix	
6	**User:** i loved it neo is amazing	The Matrix	
	Bot: Oooh, yeah, I agree. Hey, isn't Keanu Reeves in that movie? What do you think about Keanu Reeves? [Movies]	Keanu Reeves	Movies sets current entity to *Keanu Reeves*.
7	**User:** i want to talk about something else	None	Navigational Intent detects *negative intent*, so Entity Tracker rejects the current entity.

Chatbots: pro and con

Pro:
- Fun
- Good for narrow, scriptable applications

Cons:
- They don't really understand
- Rule-based chatbots are expensive and brittle
- IR-based chatbots can only mirror training data
  - The case of Microsoft Tay
    - (or, Garbage-in, Garbage-out)

The future: combining chatbots with frame-based agents

Chatbots and Dialogue Systems

Corpus-based Chatbots

Chatbots and Dialogue Systems

GUS: Simple Frame-based Dialogue Systems

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: based on a "domain ontology"

- A knowledge structure representing user intentions
- One or more **frames**
  - Each a collection of **slots** having **values**

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**	**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?

Two basic architectures

The **GUS** architecture

- 44 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 academics
- Still not widely used industrially

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
GUS:	What time do you want to leave?
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 ?
GUS:	Air California flight 310 that leaves at 8.30am and arrives at 10.00am
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

The state of the art in 1977!

Slot types can be complex

The type *DATE*

DATE

MONTH: NAME
YEAR: INTEGER

DAY: (BOUNDED-INTEGER 1 31)

WEEKDAY: (MEMBER (Sunday Monday Tuesday Wednesday Thursday Friday Saturday))

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 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 systems are production rule systems

Different types of inputs cause different productions to fire

- Each of which can fill in different frames.

The production rules can then switch control based on:

- User’s input
- Dialogue history (like the last question that the system asked)

Condition-Action Rules in Siri's GUS architecture

Active Ontology: relational network of concepts

- **data structures**: a **meeting** has
  - a date and time,
  - a location,
  - a topic
  - a list of attendees
- **rule sets** that perform actions for concepts
  - the **date** concept turns string
    - *Monday at 2pm* into
    - date object `date`(DAY,MONTH,YEAR,HOURS,MINUTES)

Rule sets

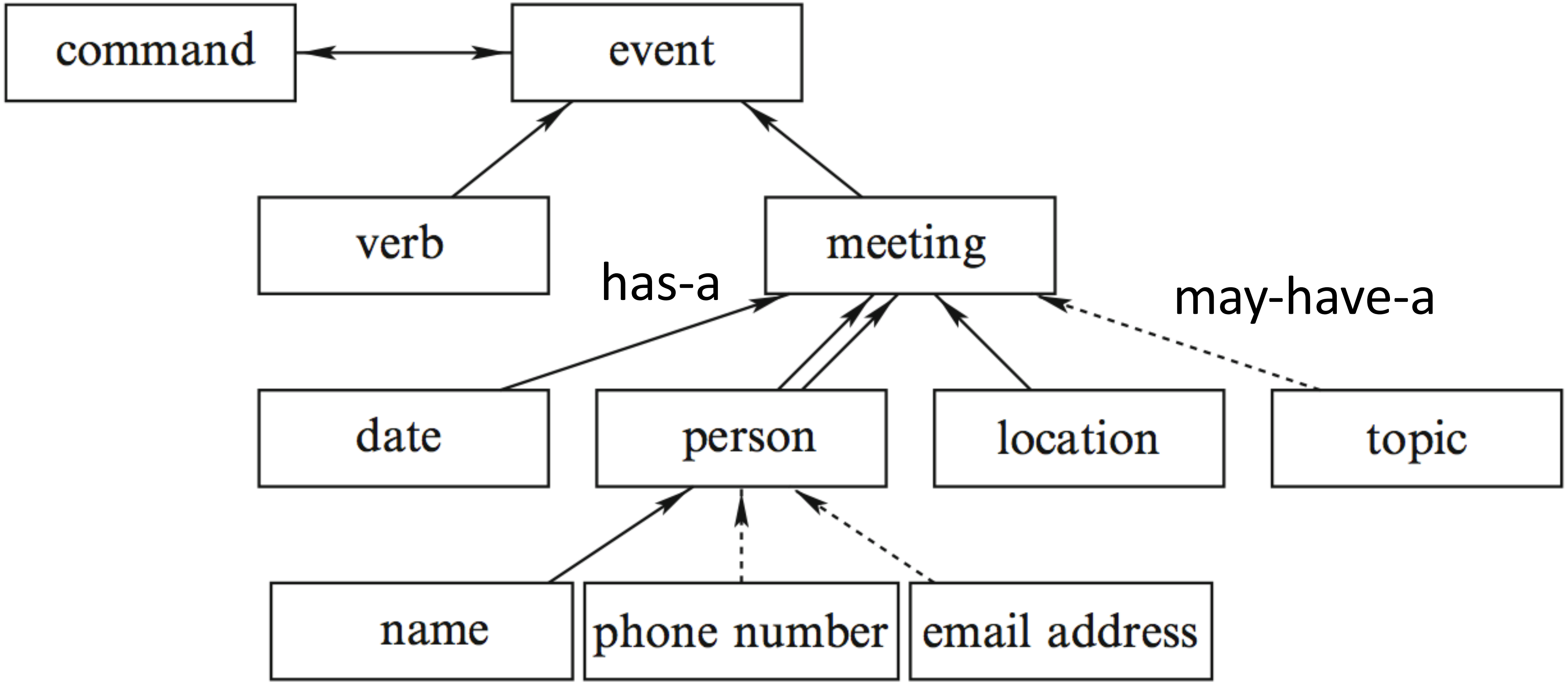
Collections of **rules** consisting of:

- condition
- action

When user input is processed, facts added to store and

- rule conditions are evaluated
- relevant actions executed

Part of ontology for meeting task



meeting concept: if you don't yet have a location, ask for a location

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.
```

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

Natural Language Understanding for filling slots

`Wake me tomorrow at six.`

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

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

Write regular expressions or grammar rules

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

Do text normalization

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?"

Chatbots and Dialogue Systems

Evaluating Chatbots

Chatbots are evaluated by humans

Automatic evaluations (BLEU/ROUGE, dot products) are generally not used for chatbots. 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 of See et al. (2019)

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

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

e.g.:

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

- •Repeated themselves over and over •Sometimes said the same thing twice • Always 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 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

Observer evaluation: acute-eval

Li et al. 2019

Annotators look at two conversations (A + B) and decide which one is better:

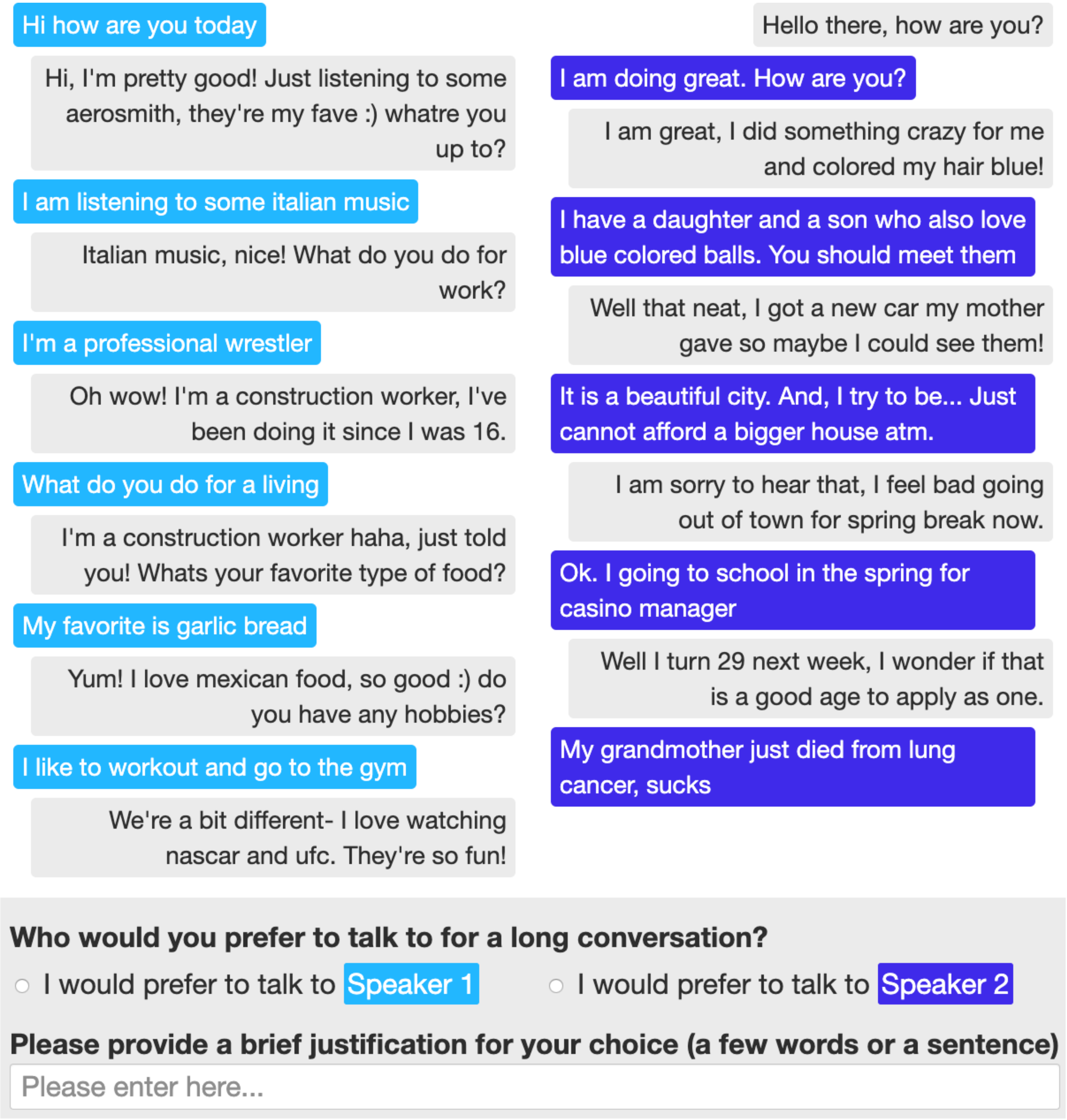
Engagingness: Who would you prefer to talk to for a long conversation?

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?

The ACUTE-EVAL method of Li et el., 2019



Chatbots and Dialogue Systems

Evaluating Chatbots

Chatbots and Dialogue Systems

GUS: Simple Frame-based Dialogue Systems

Chatbots and Dialogue Systems

The Dialogue-State Architecture

Dialogue-State or Belief-State Architecture

A more sophisticated version of the frame-based architecture

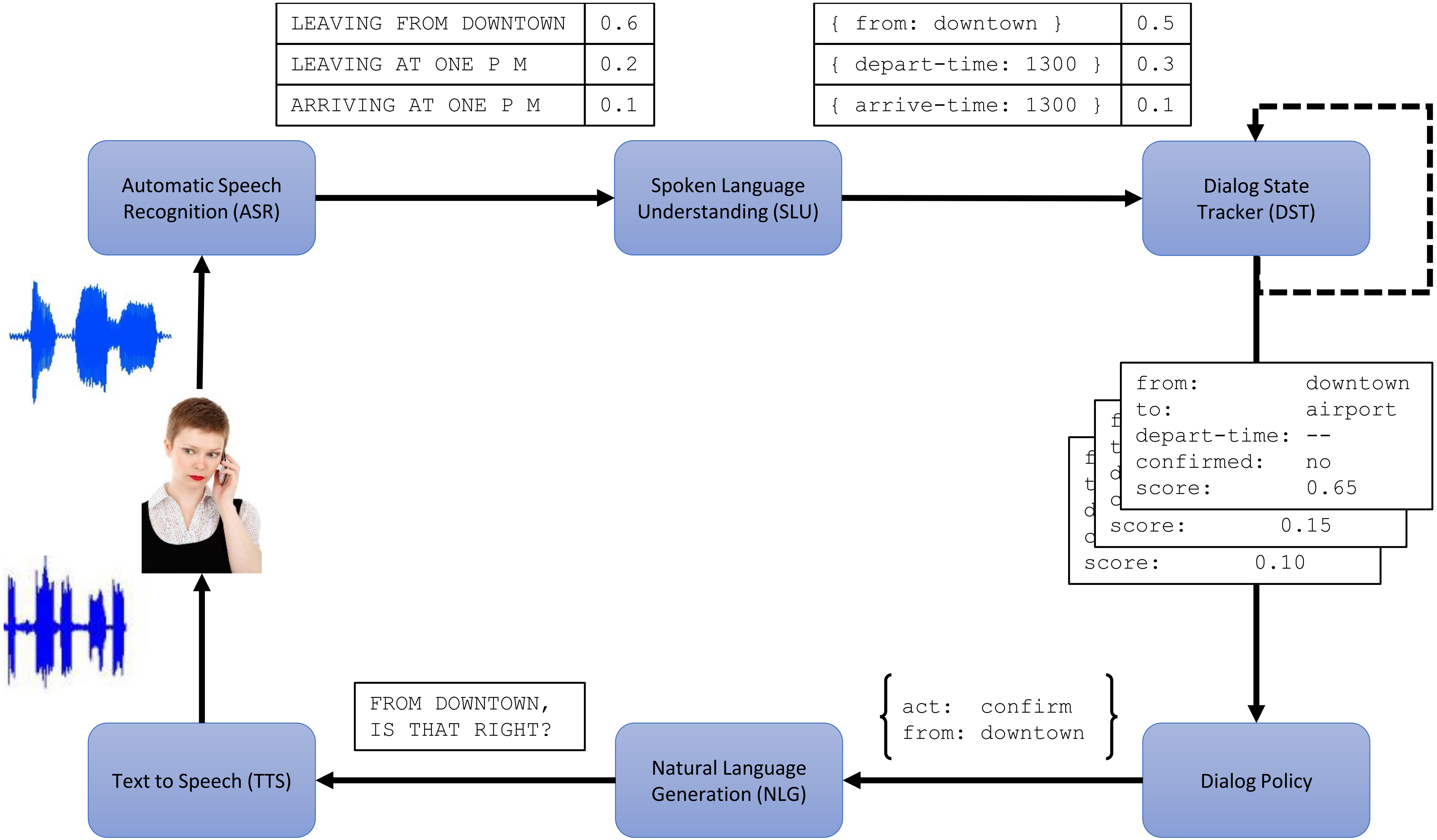
Basis for modern research systems

Slowly making its way into industrial systems

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

Williams et al., 2016

Dialogue-State Architecture



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 the 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

Young et al., 2010:

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

Dialogue Acts

Young et al., 2010:

	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()

Slot—filling: Machine learning

Machine learning classifiers to map words to semantic frame-fillers

Given a set of labeled sentences

```
“I want to fly to San Francisco on Monday afternoon please”
  Destination: SF
  Depart-date: Monday
  Depart-time: afternoon
```

- Build a classifier to map from one to the other

Requirements: Lots of labeled data

The **IO tagging** paradigm

Idea: 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`

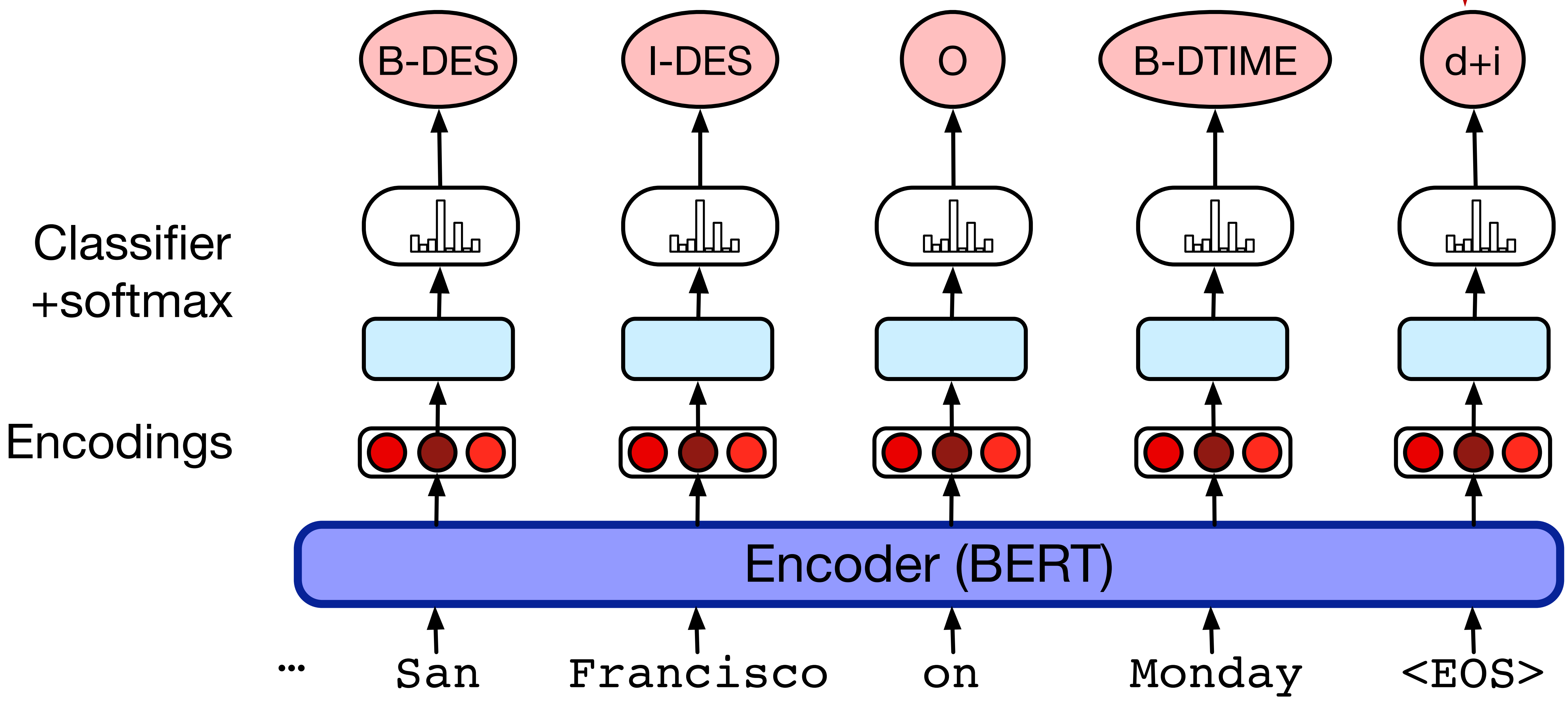
BIO tagging for slot-filling

Need training data converted to this format:

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

Slot filling using contextual embeddings

Can do domain and intent too: e.g., generate the label "AIRLINE_TRAVEL + SEARCH_FLIGHT"



The task of 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 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

But corrections are harder to recognize than other utterances!

- Swerts et al (2000): corrections misrecognized twice as often (in terms of word error rate) as non-corrections!!!
- Why?
  - Prosody seems to be largest factor: **hyperarticulation**
  - Liz Shriberg example:
    - “NO, I am DE-PAR-TING from Jacksonville”
  - Bettina Braun example from a talking elevator
    - “In den VIERTEN Stock”

Features for detecting corrections

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

Dialogue Policy

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

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

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

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

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

U:	I want to travel to Berlin
S:	**When do you want to travel to Berlin?**
U2:	Hi I'd like to fly to Seattle Tuesday Morning
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.

Progressive prompting for rejection

Don't just repeat the question "When would you like to leave?"
Give user guidance about what they can say:

System:	When would you like to leave?
Caller:	Well, um, I need to be in New York in time for the first World Series game.
System:	<reject>. Sorry, I didn't get that. Please say the month and day you'd like to leave.
Caller:	I wanna go on October fifteenth.

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:

$$
\begin{array}{lll}
< \alpha & \text{low confidence} & \text{reject} \\
\geq \alpha & \text{above the threshold} & \text{confirm explicitly} \\
\geq \beta & \text{high confidence} & \text{confirm implictly} \\
\geq \gamma & \text{very high confidence} & \text{don't confirm at all}
\end{array}
$$

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`
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

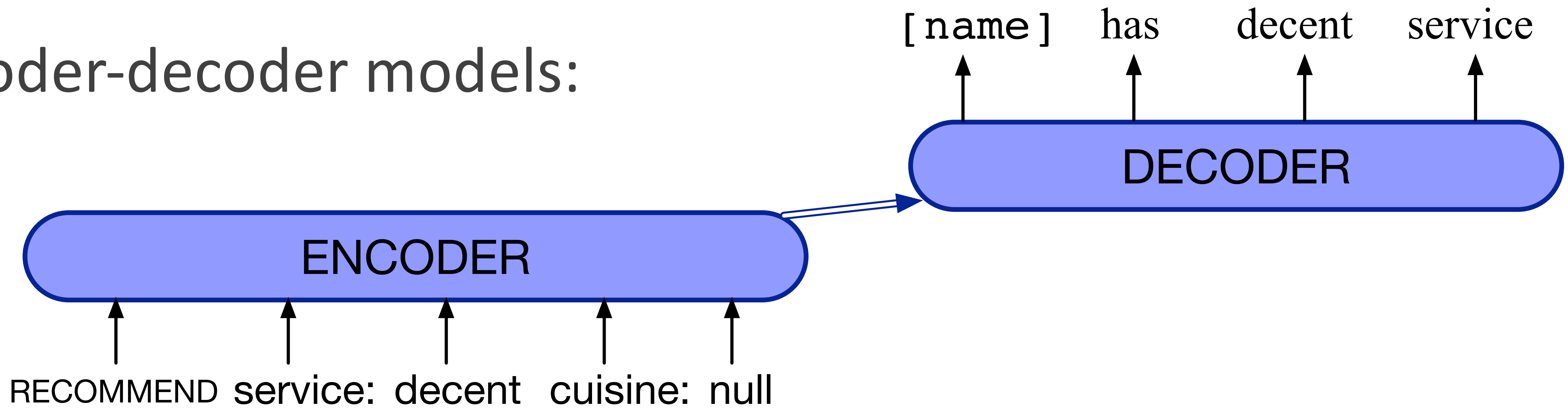
Common way to improve generalization:

- **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

Encoder-decoder models:



Output:

`restaurant_name has decent service`

Relexicalize to:

`Au Midi has decent service`

Generating clarification questions

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

System: Going where on the 5th?

The system repeats “going” and “on the 5th” to make it clear which aspect of the user’s turn the system needs to be clarified

Methods for generating clarification questions:

- Rules like 'replace “going to UNKNOWN WORD” with “going where”'
- Classifiers that guess which slots were misrecognized

Chatbots and Dialogue Systems

The Dialogue-State Architecture

Chatbots and Dialogue Systems

Evaluating Task-based Dialogue

Task completion success

1. Slot Error Rate for a Sentence

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

2. End-to-end evaluation (Task Success)

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 et al., 2001

TTS Performance	Was the system easy to understand ?
ASR Performance	Did the system understand what you said?
Task Ease	Was it easy to find the message/flight/train you wanted?
Interaction Pace	Was the pace of interaction with the system appropriate?
User Expertise	Did you know what you could say at each point?
System Response	How often was the system sluggish and slow to reply to you?
Expected Behavior	Did the system work the way you expected it to?
Future Use	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 Task-based Dialogue

Chatbots and Dialogue Systems

Design and Ethical Issues

Dialog System Design: User-centered Design

Gould and Lewis 1985

1. Study the user and task
2. Build simulations
   "Wizard of Oz study"
3. Iteratively test the design on users

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 so she could share "fun but honest comments"
- used informal language, slang, emojis, and GIFs,
- Designed to learn from users (IR-based)

The case of Microsoft Tay

Immediately Tay turned offensive and abusive

- Obscene and inflammatory tweets
- Nazi propaganda
- Conspiracy theories
- Started harassing women online

Microsoft took Tay down after 16 hours

Gina Neff and Peter Nagy 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10(2016), 4915–4931

The case of Microsoft Tay

Lessons:

- Tay quickly learned to reflect racism and sexism of Twitter users
- "If your bot is racist, and can be taught to be racist, that’s a design flaw." Caroline Sinders (2016).

Gina Neff and Peter Nagy 2016. Talking to Bots: Symbiotic Agency and the Case of Tay. *International Journal of Communication* 10(2016), 4915–4931

Female subservience in conversational agents

Chatbots overwhelmingly given female names

- likely perpetuating the stereotype of a subservient female servant

Chatbots often respond coyly or inappropriately to sexual harassment.

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

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),

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),

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

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),

Chatbots and Dialogue Systems

Design and Ethical Issues