自然语言处理

week-11

凌震华

2024年5月23日

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating Dialogue Systems

Conversational Agents

- Conversational Agents (Dialogue Systems, Dialogue Agents, Chatbots)
  - Personal Assistants on phones or other devices
  - SIRI, Alexa, Cortana, Google Assistant
  - Playing music, setting timers and clocks
  - Chatting for fun
  - Booking travel reservations
  - Clinical uses for mental health

Two kind of conversational agents

1. Chatbots
   - mimic informal human chatting
   - for fun, or even for therapy
2. (Task-based) Dialogue Agents
   - interfaces to personal assistants
   - cars, robots, appliances
   - booking flights or restaurants

语音及语言信息处理国家工程实验室

Chatbot Architectures

- **Rule-based**
  1. Pattern-action rules (ELIZA)
  2. + A mental model (PARRY):

     The first system to pass the Turing Test!

- **Corpus-based**
  3. Information Retrieval (XiaoIce)
  4. Neural encoder-decoder (BlenderBot)

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 Al 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.

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	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?

语音及语言信息处理国家工程实验室

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating Dialogue Systems

Example

A telephone 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 . . .

语音及语言信息处理国家工程实验室

Properties of Human Conversation

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

语音及语言信息处理国家工程实验室

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

- **Turn-taking issues**
  - When to take the floor?
  - When to yield the floor?
- **Interruptions**

语音及语言信息处理国家工程实验室

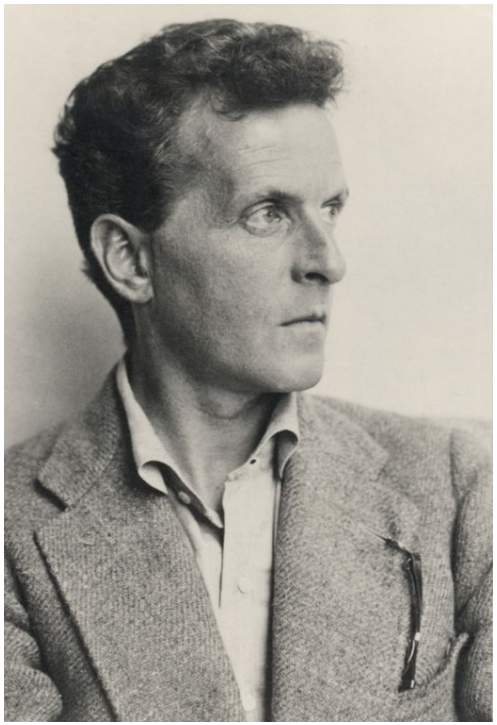
C_1: . . . I need to travel in May.
A_2: And, what day in May did you want to travel?
C_3: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_4: And you're flying into what city?
C_5: Seattle.
A_6: And what time would you like to leave Pittsburgh?
C_7: Uh hmm I don't think there's 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 . . .

Implications for Conversational Agents

- **Barge-in (插话)**
  - 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

语音及语言信息处理国家工程实验室

Language as Action



Each turn in a dialogue is a kind of action

Wittgenstein (1953) and Austin (1962)

Speech Acts (aka Dialogue Acts)
言语行为

Bach and Harnish (1979)

Constatives(断言): committing the speaker to something's being the case (*answering*, *claiming*, *confirming*, *denying*, *disagreeing*, *stating*)
Directives(指令): attempts by the speaker to get the addressee to do something (*advising*, *asking*, *forbidding*, *inviting*, *ordering*, *requesting*)
Commissives(承诺): committing the speaker to some future course of action (*promising, planning, vowing, betting, opposing*)
Acknowledgments(致谢): express the speaker's attitude regarding 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

语音及语言信息处理国家工程实验室

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?

语音及语言信息处理国家工程实验室

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?

语音及语言信息处理国家工程实验室

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?")

语音及语言信息处理国家工程实验室

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 ...

Another kind of structure: Subdialogues

Clarification subdialogue

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...

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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)

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating Dialogue Systems

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”

语音及语言信息处理国家工程实验室

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
```


语音及语言信息处理国家工程实验室

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)`

语音及语言信息处理国家工程实验室

Keywords are ranked from specific to general

User: I know everybody laughed at me

"I" is very general:

```
I: (I *) -> (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)
I         0 (list of transformation rules)
```


语音及语言信息处理国家工程实验室

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*)

语音及语言信息处理国家工程实验室

NONE

PLEASE GO ON

THAT’ S VERY INTERESTING

I SEE

语音及语言信息处理国家工程实验室

Memory

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

语音及语言信息处理国家工程实验室

PARRY: A computational model of schizophreniamory (精神分裂)

- 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

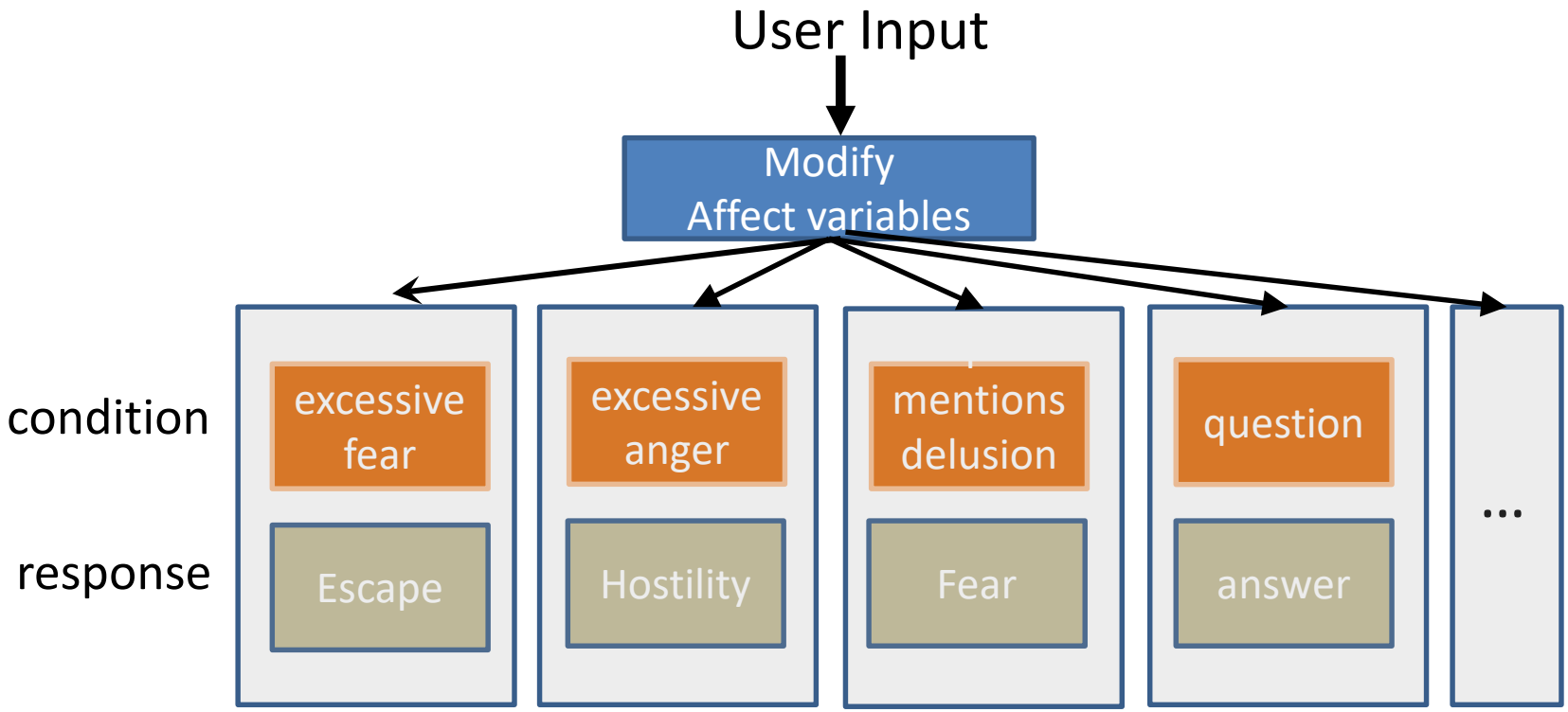
Affect variables

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

语音及语言信息处理国家工程实验室

Parry's responses depend on mental state



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.

语音及语言信息处理国家工程实验室

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating Dialogue Systems

Two architectures for corpus-based chatbots

- Response by retrieval
  - Use information retrieval to grab a response (that is appropriate to the context) from some corpus
- Response by generation
  - Use a language model or encoder-decoder to generate the response given the dialogue context

语音及语言信息处理国家工程实验室

Corpus-based chatbots require corpora

- Modern corpus-based chatbots are very data-intensive
- They commonly require hundreds of millions or billions of words

语音及语言信息处理国家工程实验室

What conversations to draw on?

- Transcripts of telephone conversations between volunteers
  - Switchboard corpus of American English telephone conversations
- Movie dialogue
  - Various corpora of movie subtitles
- Hire human crowdworkers to have conversations
  - Topical-Chat 11K crowdsourced conversations on 8 topics
  - EMPATHETICDIALOGUES 25K crowdsourced conversations grounded in a situation where a speaker was feeling a specific emotion
- Pseudo-conversations from public posts on social media
  - Drawn from Twitter, Reddit, Weibo (微博), etc.
  - Tend to be noisy; often used just as pre-training.
- Crucial to remove personally identifiable information (PII)

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

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

语音及语言信息处理国家工程实验室

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

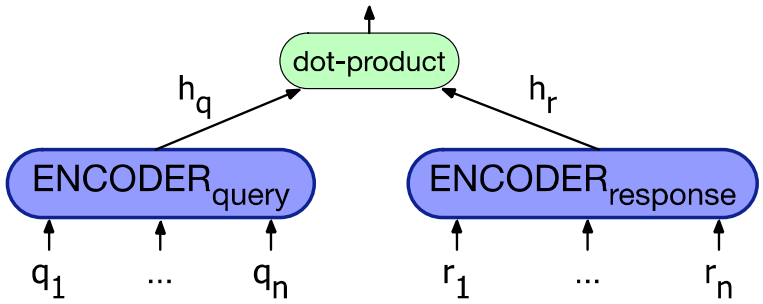
$$
\begin{aligned}
h_q &= \text{BERT}_Q(\text{q})[\text{CLS}] \\
h_r &= \text{BERT}_R(\text{r})[\text{CLS}] \\
\text{response}(q,C) &= \underset{r \in C}{\text{argmax}}\, h_q \cdot h_r
\end{aligned}
$$

语音及语言信息处理国家工程实验室

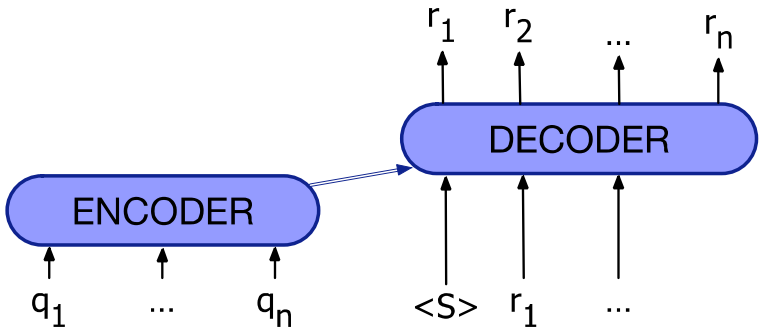
Response by generation

- 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 $r_1 \ldots r_{t-1}$

$$\hat{r}_t = \text{argmax}_{w \in V} P(w|q, r_1 \ldots r_{t-1})$$



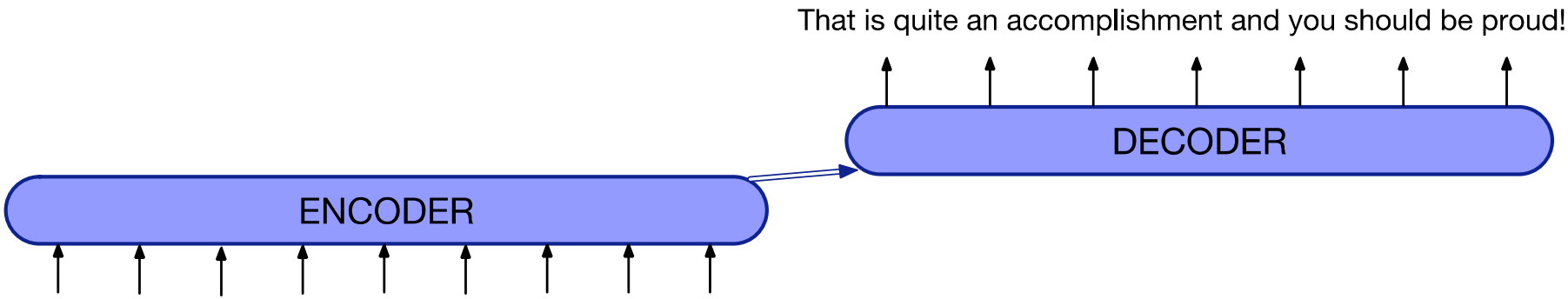
(a) Response by Retrieval



(b) Response by Generation

语音及语言信息处理国家工程实验室

Response by generation



[U:] I finally got promoted today at work!

[S:] Congrats! That's great!

[U:] Thank you! I've been trying to get it for a while now!

语音及语言信息处理国家工程实验室

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
  - on the EmpatheticDialogues dataset (Rashkin et al., 2019)

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

#	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.

Chirpy Cardinal (Paranjape et al. 2020)

语音及语言信息处理国家工程实验室

Chatbots: pro and con

- Pro:
  - Fun
  - Good for narrow, scriptable applications
- Cons:
  - They don't really understand
  - Giving the appearance of understanding may be problematic
  - Rule-based chatbots are expensive and brittle
  - IR-based chatbots can only mirror training data

语音及语言信息处理国家工程实验室

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating 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:

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

语音及语言信息处理国家工程实验室

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
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?

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

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

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

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

Natural Language Understanding for filling slots

Wake me tomorrow at six.

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

语音及语言信息处理国家工程实验室

How to fill slots? Rule-based Slot-filling

Write regular expressions or grammar rules

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

Do text normalization

语音及语言信息处理国家工程实验室

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?"

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating Dialogue Systems

Dialogue-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

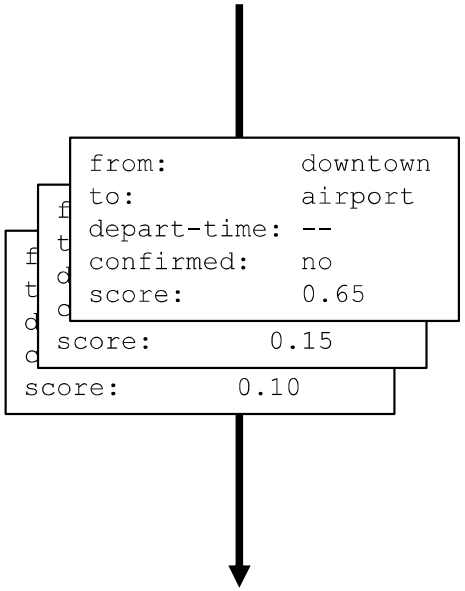
语音及语言信息处理国家工程实验室

LEAVING FROM DOWNTOWN	0.6
LEAVING AT ONE P M	0.2
ARRIVING AT ONE P M	0.1

{ from: downtown }	0.5
{ depart-time: 1300 }	0.3
{ arrive-time: 1300 }	0.1

Automatic Speech Recognition (ASR)

The Dialogue-State Architecture



Williams, Jason D., Antoine Raux, and Matthew Henderson. "The dialog state tracking challenge series: A review." Dialogue & Discourse 7, no. 3 (2016): 4-33.

语音及语言信息处理国家工程实验室

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

Young et al., 2010:

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

语音及语言信息处理国家工程实验室

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()

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

Slot filling as sequence labeling: BIO tagging

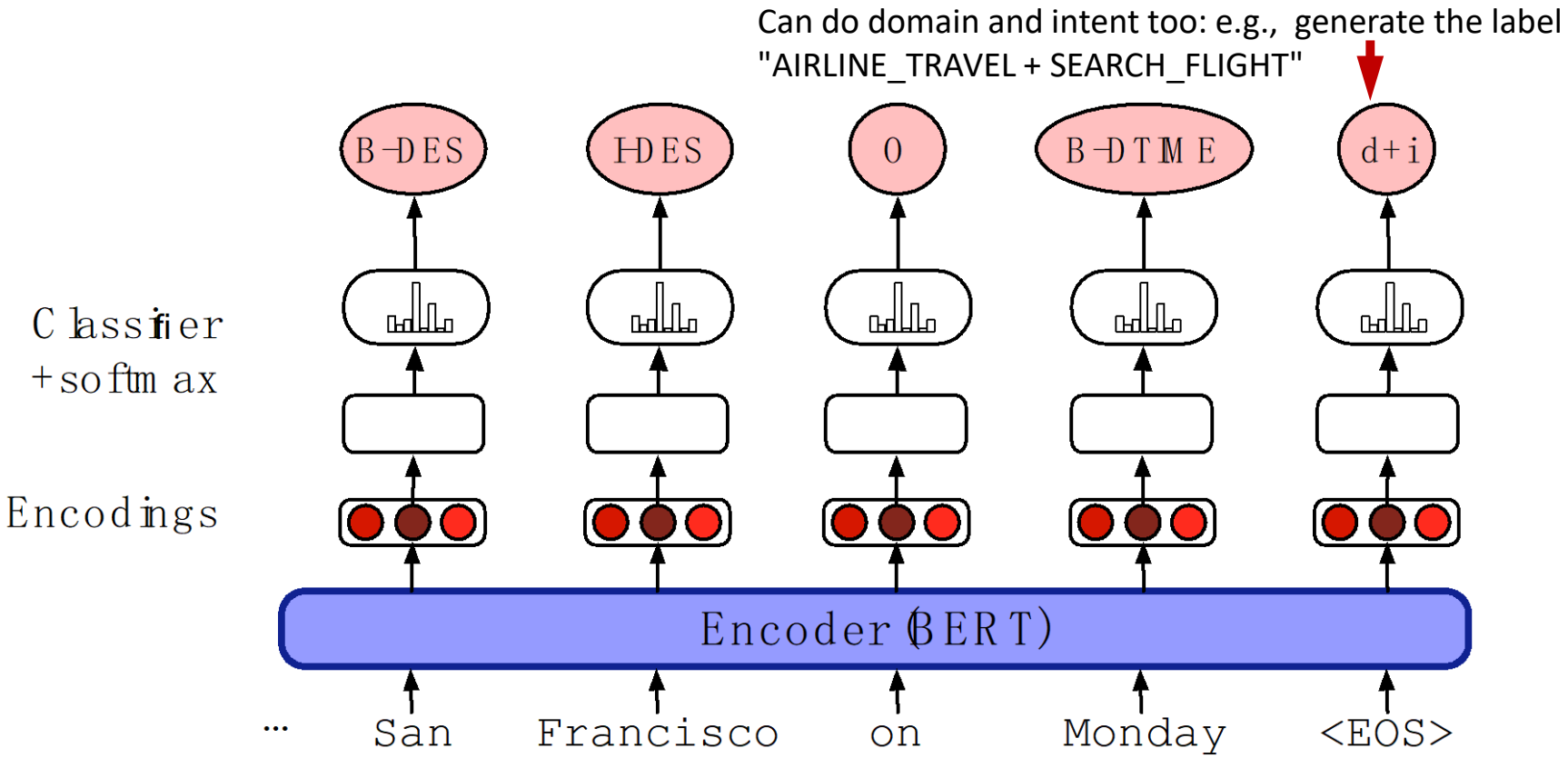
- The **BIO tagging** paradigm
- Idea: Train a classifier to label each input word with a tag that tells us what slot (if any) it fills

```
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
```

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

语音及语言信息处理国家工程实验室

Slot filling using contextual embeddings



Once we have the BIO tag of the sentence

```
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
```

- 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)

语音及语言信息处理国家工程实验室

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

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*.

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

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 systems (SLS): Factors affect-ng performance and user satisfaction. DARPA Speech and Natural Language Workshop.
- "I said BAL-TI-MORE, not Boston"

语音及语言信息处理国家工程实验室

Features for detecting corrections in spoken dialogue

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:

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

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

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

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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?

语音及语言信息处理国家工程实验室

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).

语音及语言信息处理国家工程实验室

Rejection

- *I’m sorry, I didn’t understand that.*

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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 implictly
$\geq \gamma$	very high confidence	don't confirm at all

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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
```

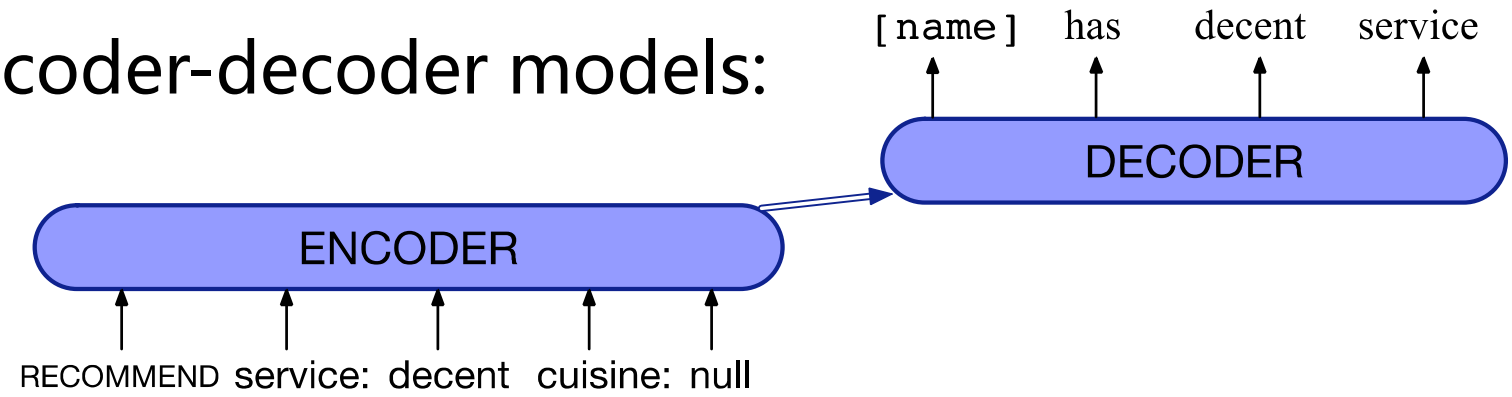
1 `restaurant_name` is in `neighborhood` and serves `cuisine` food.

2 There is a `cuisine` restaurant in `neighborhood` called `restaurant_name`.

语音及语言信息处理国家工程实验室

Sentence Realization: mapping from frames to delexicalized sentences

Encoder-decoder models:



Output:

```
restaurant_name has decent service
```

Relexicalize to:

```
Au Midi has decent service
```


语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室

- Chatbots and Dialogue Systems 聊天机器人与对话系统
  - Introduction
  - Properties of Human Conversation
  - Rule-based Chatbots: ELIZA and PARRY
  - Corpus-based Chatbots
  - Frame-based Dialogue Architecture
  - The Dialogue-State Architecture
  - Evaluating Dialogue Systems

Evaluating chatbots and task-based dialogue

- Task-based dialogue:
  - mainly by measuring task performance
- Chatbots:
  - mainly by human evaluation

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

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 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

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

语音及语言信息处理国家工程实验室

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?

语音及语言信息处理国家工程实验室

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*.

语音及语言信息处理国家工程实验室

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.

语音及语言信息处理国家工程实验室

Task-based systems are evaluated by task success!

1. End-to-end evaluation (Task Success)
2. Slot Error Rate for a Sentence

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

语音及语言信息处理国家工程实验室

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?

语音及语言信息处理国家工程实验室

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	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?

语音及语言信息处理国家工程实验室

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

语音及语言信息处理国家工程实验室