

**SIT330-770: Natural Language Processing**

Week 10 – Dialogue Systems & Conversational AI

Dr. Mohamed Reda Bouadjenek

School of Information Technology, Faculty of Sci Eng & Built Env

[reda.bouadjenek@deakin.edu.au](mailto:reda.bouadjenek@deakin.edu.au)

**DEAKIN UNIVERSITY**

1

**SIT330-770: Natural Language Processing**

Week 10. 1 - Introduction to Chatbots and Dialogue Systems

Dr. Mohamed Reda Bouadjenek

School of Information Technology, Faculty of Sci Eng & Built Env

**DEAKIN UNIVERSITY**

2

**Conversational Agents (AKA Dialogue Systems AKA Dialogue Agents AKA 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

3

**Two kind of conversational agents**

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

4

**Chatbot Architectures**

- Rule-based**
  - Pattern-action rules ([ELIZA](#))
  - + A mental model ([PARRY](#)):  
The first system to pass the Turing Test!
- Corpus-based**
  - Information Retrieval (Xiaolce)
  - Neural encoder-decoder (BlenderBot)

5

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 sing 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)

6

Xiaoice (Zhou et al., 2020)

The screenshot shows a conversation between a human and the Xiaoice chatbot. The human asks about the bot's profile picture, if it's a comic character, and its functions. The bot responds with its name, a guess, and information about its skills.

[人类] 小冰你好	[Human] Hello, Xiaoice
[小冰] 你好呀，你头像好有趣，感觉	[Xiaoice] your profile pic is
在哪儿见过	interesting. Like I've seen it somewhere.
[人类] 是吗	[Human] Really?
[小冰] 没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	[Xiaoice] So what? I won't be cold and lifeless.
[人类] 你都有什么功能啊	[Human] What functions do you have?
[小冰] 我有200多个技能，来，说出	[Xiaoice] I have over 200 skills. Just name one you need.

7

### 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

8

### 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?

9

SIT330-770: Natural Language Processing

Week 10, 2 - Properties of Human Conversation

Dr. Mohamed Reda Bouadjenek

School of Information Technology,  
Faculty of Sci Eng & Built Env

The slide features a red robot with a circular head and a cylindrical body, interacting with a smartphone. The phone screen shows a messaging interface with several messages exchanged between users.

10

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

C<sub>1</sub>: ...I need to travel in May.  
 A<sub>2</sub>: And, what day in May did you want to travel?  
 C<sub>3</sub>: Oh, I need to be there for a meeting that's from the 12th to the 15th.  
 A<sub>4</sub>: And you're flying into what city?  
 C<sub>5</sub>: Seattle.  
 A<sub>6</sub>: And what time would you like to leave Pittsburgh?  
 C<sub>7</sub>: Uh hum I don't think there's many options for non-stop.  
 A<sub>8</sub>: Right. There's three non-stops today.  
 C<sub>9</sub>: What are they?  
 A<sub>10</sub>: 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<sub>11</sub>: I'll take the 8:15pm the night before on the 11th.  
 A<sub>12</sub>: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.  
 C<sub>13</sub>: OK.  
 A<sub>14</sub>: And you said returning on May 15th?  
 C<sub>15</sub>: Uh, yeah, at the end of the day.  
 A<sub>16</sub>: OK. There's #two non-stops ... #  
 C<sub>17</sub>: #Act... actually #, what day of the week is the 15th?  
 A<sub>18</sub>: It's a Friday.  
 C<sub>19</sub>: Uh hum. I would consider staying there an extra day til Sunday.  
 A<sub>20</sub>: OK... OK. On Sunday I have ...

11

### Properties of Human Conversation

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

12

C1: ...I need to travel in May.  
 A2: And what day in May did you want to travel?  
 C3: OK uh I need to be there for a meeting that's from the 12th to the 15th.  
 A4: And you're flying into what city?  
 C5: Seattle.  
 A6: And what time would you like to leave Pittsburgh?  
 C7: Uh hmm I don't think there's many options for non-stop.  
 A8: Right. There's three non-stops today.  
 C9: What are they?  
 A10: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.  
 The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.  
 C11: OK I'll take the 8ish flight on the night before on the 11th.  
 A12: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.  
 C13: OK.  
 A14: And you said returning on May 15th?  
 C15: Uh yeah, at the end of the day.  
 A16: OK. There's #two non-stop...#  
 C17: #Act... actually #, what day of the week is the 15th?  
 A18: It's a Friday.  
 C19: Uh hmm. I would consider staying there an extra day til Sunday.  
 A20: OK...OK. On Sunday I have ...

13

**Properties of Human Conversation**

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

14

C1: ...I need to travel in May.  
 A2: And, what day in May did you want to travel?  
 C3: OK uh I need to be there for a meeting that's from the 12th to the 15th.  
 A4: And you're flying into what city?  
 C5: Seattle.  
 A6: And what time would you like to leave Pittsburgh?  
 C7: Uh hmm I don't think there's many options for non-stop.  
 A8: Right. There's three non-stops today.  
 C9: What are they?  
 A10: The first one departs PGH at 10:00am arrives Seattle at 12:05 their time.  
 The second flight departs PGH at 5:55pm, arrives Seattle at 8pm. And the last flight departs PGH at 8:15pm arrives Seattle at 10:28pm.  
 C11: OK I'll take the 8ish flight on the night before on the 11th.  
 A12: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.  
 C13: OK.  
 A14: And you said returning on May 15th?  
 C15: Uh yeah, at the end of the day.  
 A16: OK. There's #two non-stop...#  
 C17: #Act... actually #, what day of the week is the 15th?  
 A18: It's a Friday.  
 C19: Uh hmm. I would consider staying there an extra day til Sunday.  
 A20: OK...OK. On Sunday I have ...

15

**Implications for Conversational Agents**

- Barge-in
  - Allowing the user to interrupt
- End-pointing
  - The task for a speech system of deciding whether the user has stopped talking.
  - Very hard, since people often pause in the middle of turns

16

**Language as Action**



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

17

**Speech Acts (aka Dialogue Acts)**

Bach and Harnish (1979)

<b>Constatives:</b> committing the speaker to something's being the case ( <i>answering, claiming, confirming, denying, disagreeing, stating</i> )
<b>Directives:</b> attempts by the speaker to get the addressee to do something ( <i>advising, asking, forbidding, inviting, ordering, requesting</i> )
<b>Commissives:</b> committing the speaker to some future course of action ( <i>promising, planning, vowed, betting, opposing</i> )
<b>Acknowledgments:</b> express the speaker's attitude regarding the hearer with respect to some social action ( <i>apologizing, greeting, thanking, accepting an acknowledgment</i> )

18

**Speech acts**

- "Turn up the music!"
  - Directive
- "What day in May do you want to travel?"
  - Directive
- "I need to travel in May"
  - Constitutive
- Thanks
  - Acknowledgement

19

**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

20

**Grounding**

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



21

**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?**

22

**Grounding is important for computers too!**

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

23

**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?")

24

Another kind of structure: Subdialogs

- 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 ...

45

25

Clarification Subdialogs

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

46

26

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.

47

27

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)

48

28

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.

49

29

**SIT330-770: Natural Language Processing**

Week 10, 3 - Rule-based Chatbots: ELIZA and PARRY

Dr. Mohamed Reda Bouadjenek  
 School of Information Technology,  
 Faculty of Sci Eng & Built Env



50

**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's such a...  
I AM SORRY TO HEAR YOU ARE DEPRESSED  
I am sorry for you.  
You are like my father in some ways.  
-WHAT ELSE COMES TO MIND?  
-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.  
-WHAT DO YOU THINK I DON'T ARGUE WITH YOU?  
-You are afraid of me.  
-DOES IT PLEASE YOU TO BELIEVE I AM AFRAID OF YOU?  
My father was 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?

31

**ELIZA's trick: be a Rogerian psychologist**

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

- 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"

32

**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

33

**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

34

**Eliza Rules**

- Rules are organized by **keywords**
- Each keyword has a pattern and a list of possible transforms

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

keyword	pattern	list of transforms
(K ((D <sub>1</sub> ) (R <sub>1,1</sub> ) (R <sub>1,2</sub> ) ... (R <sub>1,m<sub>1</sub></sub> ))	((D <sub>2</sub> ) (R <sub>2,1</sub> ) (R <sub>2,2</sub> ) ... (R <sub>2,m<sub>2</sub></sub> ))	⋮
⋮	⋮	⋮
((D <sub>n</sub> ) (R <sub>n,1</sub> ) (R <sub>n,2</sub> ) ... (R <sub>n,m<sub>n</sub></sub> )))		

35

**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)

36

```

function ELIZA GENERATOR(user sentence) returns response
    Find the word w in sentence that has the highest keyword rank
    if w exists
        Choose the highest ranked rule r for w that matches sentence
        response  $\leftarrow$  Apply the transform in r to sentence
        if w = 'my'
            future  $\leftarrow$  Apply a transformation from the 'memory' rule list to sentence
            Push future onto memory stack
        else (no keyword applies)
        either
            response  $\leftarrow$  Apply the transform for the NONE keyword to sentence
        or
            response  $\leftarrow$  Pop the top response from the memory stack
    return(response)

```

37

**NONE**

- PLEASE GO ON
- THAT'S VERY INTERESTING
- ISEE

38

**Memory**

- (MEMORY MY  
(O MY O = LETS DISCUSS FURTHER WHY YOUR 3)  
• (O MY O = EARLIER YOU SAID YOUR 3)  
• (O MY O = 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

39

**Ethical implications: Anthropomorphism and Privacy**

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

40

**Ethical implications**

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

Joseph Weizenbaum, 1976. Computer Power and Human Reason: From Judgment to Calculation. W.H. Freeman.  
Sherry Turkle, 2007. Being Alone in the Age of Distraction. Basic Books.  
Sherry Turkle, 2007. Authenticity in the age of digital companions. Interaction Studies, 8(2), pp.501-517

41

**PARRY: A computational model of schizophrenia**

- Another chatbot with a clinical psychology focus
  - Colby, K. M., Weber, S., and Hilf, F. D. (1972). 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

42

**Affect variables**

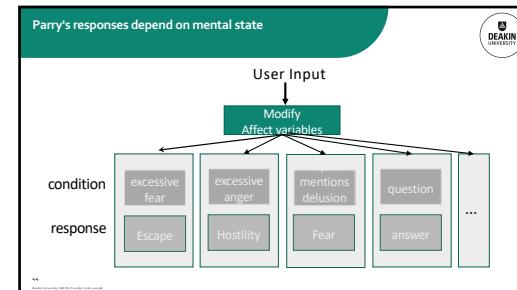
Fear (0-20)      Anger (0-20)      Mistrust (0-15)

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



43



**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.



45

**SIT330-770: Natural Language Processing**  
Week 10. 4 - Corpus-based Chatbots

Dr. Mohamed Reda Bouadjenek  
School of Information Technology,  
Faculty of Sci Eng & Built Env



46

**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



47

**Corpus-based chatbots require corpora**

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



48

**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 1M 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)

49

**Response by retrieval: classic IR method**

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

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

50

**Response by retrieval: neural IR method**

- Given a user turn  $q$ , and a training corpus  $C$  of conversation
- Find in  $C$  the turn  $r$  that is most similar (BERT dot product) to  $q$
- Say  $r$

$$h_q = \text{BERT}_Q(q)[\text{CLS}]$$

$$h_r = \text{BERT}_R(r)[\text{CLS}]$$

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

51

**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  $r_1 \dots r_{t-1}$

$$\hat{r}_t = \underset{w \in V}{\operatorname{argmax}} P(w | q, r_1 \dots r_{t-1})$$

52

**Response by generation**

53

**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)

54

P.G. Wodehouse predicts 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*, 1919

55

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.

56

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

57

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.

58

#	Utterance	Entity	Commentary
1	User: let's chat	None	Launch gives <b>FORCE_START</b> response to begin the conversation
Bot:	Hi, this is an Alexa Prize Socialbot. I'd love to get to know you better. What's your name? Shall I ask it all right if I ask for your name? [Launch]	None	
2	User: my name is chris	None	Launch gives a <b>STRONG_CONTINUE</b> response that needs a prompt, which is supplied by Neural Chat with <b>FORCE_START</b> priority (scripted transition).
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?	None	
3	User: hang out with my friends	None	Neural Chat generates a GPT-2 response with <b>STRONG_CONTINUE</b> priority.
Bot:	That sounds great, I love spending time with friends. What will you do together? [Neural Chat]	None	
4	User: maybe watch some movies	None	Movies detects the 'movie' keyword in the user's utterance and gives a <b>FORCE_START</b> response (overriding Neural Chat), setting current entity to <i>The Matrix</i> .
Bot:	I got to say, I really like watching movies; I get so much about what the world is like outside of the cloud! Have you seen any movies recently? [Movies]	The Matrix	The Entity Linker and Entity Tracker identify <i>The Matrix</i> as the current entity.
5	User: i saw the matrix	The Matrix	Movies sets current entity to <i>Keana Reeves</i> .
Bot:	Nice! Did you like <i>The Matrix</i> ? [Movies]	The Matrix	
6	User: i think it's amazing	The Matrix	
Bot:	Yeah, yeah, I agree. Hey, isn't Keana Reeves in that movie? What do you think about Keana Reeves? [Movies]	Keana Reeves	
7	User: i want to talk about something else	None	Navigational Intent detects negative intent, so Entity Tracker rejects the current entity.

59

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
    - We'll discuss later the case of Microsoft Tay
      - (or, Garbage-in, Garbage-out)
- Next steps: integrating chatbot ability into frame-based agents

60

**SIT330-770: Natural Language Processing**  
Week 10. 5 - The Frame-based ("GUS")  
Dialogue Architecture  
  
Dr. Mohamed Reda Bouajeneck  
  
School of Information Technology,  
Faculty of Sci Eng & Built Env

61

### Frame-based dialogue agents

- Sometimes called “**task-based dialogue agents**”
- Systems that have the goal of helping a user solve a task like making a travel reservation or buying a product
- **Architecture:**
- First proposed in the GUS system of 1977
- A knowledge structure representing user intentions
- One or more **frames** (each consisting of **slots** with **values**)

62

### 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?

63

### 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

64

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

65

### 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

66

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

47

67

**GUS systems have multiple frames**

- Frames like:
  - o Car or hotel reservations
  - o General route information
  - o Which airlines fly from Boston to San Francisco?
  - o Information about airfare practices
  - o Do I have to stay a specific number of days to get a decent airfare?).
- Frame detection
  - o System must detect which slot of which frame user is filling
  - o And switch dialogue control to that frame.

48

68

**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

49

69

**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

50

70

**Natural Language Understanding for filling slots**

• Wake me tomorrow at six.

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

51

71

**How to fill slots?**

(a) **Rule-based Slot-filling**

- Write regular expressions or grammar rules
- Wake me (up) | set (the|an) alarm | get me up
- Do text normalization

52

72

**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?"

23

73

**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

24

74

**SIT330-770: Natural Language Processing**  
Week 10. 6 - The Dialogue-State Architecture  
Dr. Mohamed Reda Bouadjenek  
School of Information Technology,  
Faculty of Sci Eng & Built Env

DEAKIN UNIVERSITY

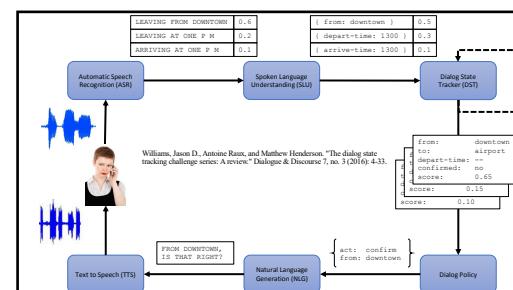
75

**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

25

76



77

**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

26

78

**Dialogue Acts**

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

Young et al., 2010:

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

79

**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()

80

**Slot filling: Machine learning**

- Machine learning classifiers to map words to semantic frame-filters
- 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

81

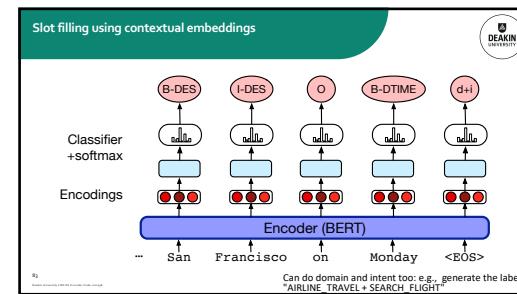
**Slot filling as sequence labeling: BIO tagging**

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

0 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

82



83

**Once we have the BIO tag of the sentence**

0 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)

84

**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

45 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.

85

**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

86

**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

87

**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 affecting performance and user satisfaction. DARPA Speech and Natural Language Workshop.
- "I said BAL-TI-MORE, not Boston"

88

**Features for detecting corrections in spoken dialogue**

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

89

**SIT330-770: Natural Language Processing**

Week 10, 7 -The Dialogue-State Architecture Continued: Policy and Generation

Dr. Mohamed Reda Bouadjenek  
School of Information Technology,  
Faculty of Sci Eng & Built Env

90

**Dialogue Policy**

- At turn  $i$  predict action  $A_i$  to take, given entire history:
$$\hat{A}_i = \operatorname{argmax}_{A_i \in A} P(A_i | (A_1, U_1, \dots, A_{i-1}, U_{i-1}))$$

- Simplify by just conditioning on the current dialogue state (filled frame slots) and the last turn and turn by system and user:
$$\hat{A}_i = \operatorname{argmax}_{A_i \in A} P(A_i | \text{Frame}_{i-1}, A_{i-1}, U_{i-1})$$

91

**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.

92

**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

93

**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?

94

**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).

95

**Rejection**

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

96

### 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.

97

### Using confidence to decide whether to confirm:



- ASR or NLU systems can assign a confidence value, indicating how likely they are that they understood the user.
  - Acoustic log-likelihood of the utterance
  - Prosodic features
  - Ratio of score of best to second-best interpretation
- Systems could use set confidence thresholds:
 

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

98

### 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.

99

### 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

100

### 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 Centre and serves seafood food.
4 There is a seafood restaurant in the City Centre called Loch Fyne.
```

101

- 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.
```

102

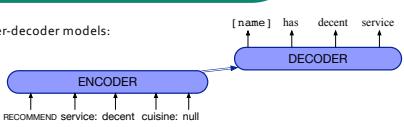
**Sentence Realization**

- 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]**



103

**Sentence Realization: mapping from frames to delexicalized sentences**

- Encoder-decoder models:
 
- Output:  
restaurant\_name has decent service
- Relexicalize to:  
Au Midi has decent service



104

**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



105

**SIT330-770: Natural Language Processing**

Week 10.8 - Evaluating Dialogue Systems

Dr. Mohamed Reda Bouadjenek

School of Information Technology,  
Faculty of Sci Eng & Built Env




106

**Evaluating chatbots and task-based dialogue**

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



107

**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.



108

**Participant evaluation**

Li, M., Weston, J., and Roller, S. (2019). Acute-Eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. NeurIPS20 Workshop on Conversational AI.

- 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
- (2) 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
- (3) 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 2019.

109

**Observer evaluation: acute-eval**

Li, M., Weston, J., and Roller, S. (2019). Acute-Eval: Improved dialogue evaluation with optimized questions and multi-turn comparisons. NeurIPS20 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?

110

**The ACUTE-EVAL method**  
Li et al., 2019

Hello there, how are you?  
I am pretty good. Just listening to some songs right now, they're my fave : ) what's up?  
I am great, I got a new job and I'm crazy for the new color of my hair!

I am listening to some Italian music.  
Italian music, nice! What do you do for work?  
I have a daughter and a son who also love blue colored balls. You should meet them!

I'm a professional wrestler.  
Oh wow! I'm a construction worker. I've been doing it since I was 16.  
Well that's cool! I'm a movie producer. Well that's cool! I could be there!

What do you do for a living?  
I'm a construction worker haha, just told you I'm a construction worker type of food!  
Well that's cool! I could be there!

My favorite is pizza.  
Yum! I love mexican food, no good I do you have any hobbies?  
I like to work and go to the gym.

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

Who would you prefer to talk to for a long conversation?  
I would prefer to talk to Speaker 1  
... I would prefer to talk to Speaker 2

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

111

**Automatic evaluation is an open problem**

DEAKIN UNIVERSITY

- 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.

112

**Task-based systems are evaluated by task success!**

DEAKIN UNIVERSITY

- End-to-end evaluation (Task Success)
- Slot Error Rate for a Sentence
 
$$\frac{\# \text{ of inserted/deleted/substituted slots}}{\# \text{ of total reference slots for sentence}}$$

113

**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?

114

More fine-grained metrics: User Satisfaction Survey

Waller, 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 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?

115

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

116

SIT330-770: Natural Language Processing

Week 10. 9 - Design and Ethical Issues

Dr. Mohamed Reda Bouadjenek  
School of Information Technology,  
Faculty of Sci Eng & Built Env

117

Dialog System Design: User-centered Design

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

Gould, John D., and Clayton Lewis. "Designing for usability: key principles and what designers think." *Communications of the ACM* 28, no. 3 (1985): 300-311.  
Bender, Emily M., and Bahar Friedman. "Data statements for natural language processing: Toward mitigating system bias and enabling better science." *TACL* 6 (2018): 587-604.

118

Ethical design

- Ethical issues have long been known to be crucial in artificial agents
- Mary Shelley's Frankenstein
  - creating agents without a consideration of ethical and humanistic concerns
- Ethical issues:
  - Safety:** Systems abusing users, distracting drivers, or giving bad medical advice
  - Representational harm:** Systems demeaning particular social groups
  - Privacy:** Information Leakage

119

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, Savitree Verka, Nicolas Arguello-Gutiérrez, Naveen Arora, R. Gerardo-Avila-Fried, Ryan Lowe, and Julie Pineau. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES'18).

120

**Abuse and Representation Harm: The case of Microsoft Tay**



- Experimental Twitter chatbot launched in 2016
  - given the profile personality of an 18- to 24-year-old American woman
  - could share horoscopes, tell jokes,
  - asked people to send selfies
  - used informal language, slang, emojis, and GIFs,
  - Designed to learn from users (IR-based)

121

**The case of Microsoft Tay**



- Immediately Tay turned offensive and abusive
  - Obscene and inflammatory tweets
  - Nazi propaganda, conspiracy theories
  - Began harassing women online
  - Reflecting racism and misogyny of Twitter users
- Microsoft took Tay down after 16 hours
- Lessons:
  - User response must be considered in the design phase

122

Giles Heff and Peter Mays. 2016. Talking to Basic: Symbolic Agency and the Case of Tay. *International Journal of Communication* 10(2016), 4915–4932.

**Bias in training datasets**



- Henderson *et al.* ran hate-speech and bias detectors on standard training sets for dialogue systems:
  - Twitter, Reddit, other dialogue datasets
- Found bias and hate-speech
  - In training data
  - In dialogue models trained on the data

123

Peter Henderson, Koustuv Sinha, Nicolas Angerlo-Gantner, Han Asanathy Ko, Genevieve Fried, Ryan Lowe, and Jessie Pines. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AAECS '18).

**Privacy: Training on user data**



- Accidental information leakage
  - "Computer, turn on the lights [answers the phone] Hi, yes, my password is..."
  - Henderson show in simulation that this leakage can occur.
- Intentional information leakage
  - Dialogue systems that are designed to send user data to developer or advertisers
  - Important to consider privacy-preserving dialogue systems

124

Peter Henderson, Koustuv Sinha, Nicolas Angerlo-Gantner, Han Asanathy Ko, Genevieve Fried, Ryan Lowe, and Jessie Pines. 2018. Ethical Challenges in Data-Driven Dialogue Systems. In 2018 AAAI/ACM Conference on AI, Ethics, and Society (AAECS '18).  
Carrasco, Gonzalo, Balázs Károlyi, Chai Xu, Michael Fischer, and Monica S. Lam. "Valkyrie: The architecture of an open, crowd-sourced, privacy-preserving, programmable virtual assistant." In *Proceedings of the 26th International Conference on World Wide Web*, pp. 181–190. 2017.