Topics in Advance Dialog

CS136a Speech Recognition
Marie Meteer
Overview

■ Dialog Acts

■ Dialog State: Interpretation
  ■ Sketching an algorithm for dialog Act Interpretation
  ■ Special case: Detecting correction acts

■ Dialog Policy
  ■ Generating Dialog Acts: Confirming and Rejecting

■ A simple policy based on local context

■ Natural language generation in the dialog-state model
Frame-based dialog agents

- Sometimes called "task-based dialog agents"

- Based on a "domain ontology"
  - A knowledge structure representing user intentions

- One or more **frames**
  - Each a collection of **slots**
  - Each slot having a **value**
The Frame

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

<table>
<thead>
<tr>
<th>Slot</th>
<th>Type</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGIN</td>
<td>city</td>
<td>What city are you leaving from?</td>
</tr>
<tr>
<td>DEST</td>
<td>city</td>
<td>Where are you going?</td>
</tr>
<tr>
<td>DEP DATE</td>
<td>date</td>
<td>What day would you like to leave?</td>
</tr>
<tr>
<td>DEP TIME</td>
<td>time</td>
<td>What time would you like to leave?</td>
</tr>
<tr>
<td>AIRLINE</td>
<td>line</td>
<td>What is your preferred airline?</td>
</tr>
</tbody>
</table>
Frame-based dialogue agents

1977:

GUS, A Frame-Driven Dialog System¹

Daniel G. Bobrow, Ronald M. Kaplan, Martin Kay, Donald A. Norman, Henry Thompson and Terry Winograd

Xerox Palo Alto Research Center, 3333 Coyote Hill Road, Palo Alto, CA 94304, U.S.A.

Artificial Intelligence Journal, 1977

Still the industrial state of the art

SIRI based on GUS architecture
The GUS architecture

- A kind of *mixed initiative*
  - The conversational initiative shifts between system and user
- The structure of the **frame** guides dialogue
A travel dialog: Communicator
Xu and Rudnicky (2000)

S1: Hello. You’ve reached the [Communicator]. Tell me your full name.

U2: Hi, I’d like to fly to Seattle Tuesday morning

S3: Traveling to Seattle on Tuesday, August 11th in the morning. Your full name?

U4: John Doe

S5: Hello, John. I’m glad to hear from you again. OK. Let’s see, I have a United flight departing Pittsburgh at 10:45 am arriving into Seattle at 1:57 pm. Is that OK?

U6: Yes

S7: Will you return to Pittsburgh from Seattle?

U8: No, I will go on to Los Angeles

S9: Travelling to Los Angeles on what date

U10: Sorry, can we go back to the previous leg again

S11: For the first leg, you are on United flight 1446 departing Pittsburgh at 10:45 AM …
the user said, and may need to ask clarification questions. The system needs a dialog policy to decide what to say (when to answer the user’s questions, when to instead ask the user a clarification question, make a suggestion, and so on).

Figure 25.1 shows a typical architecture for a dialog-state system. It has six components. As with the GUS-style frame-based systems, the speech recognition and understanding components extract meaning from the input, and the generation and TTS components map from meaning to speech. The parts that are different than the simple GUS system are the dialog state tracker which maintains the current state of the dialog (which include the user’s most recent dialog act, plus the entire set of slot-filler constraints the user has expressed so far) and the dialog policy, which decides what the system should do or say next.

### Dialog State Tracking Overview

**LEAVING FROM DOWNTOWN**
- LEAVING AT ONE P M: 0.6
- ARRIVING AT ONE P M: 0.2
- ARRIVING AT ONE P M: 0.1

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

**{ from: downtown }**
- { depart-time: 1300 }: 0.6
- { arrive-time: 1300 }: 0.1

**Dialog State Tracker (DST)**
- From: downtown
- To: airport
- Depart-time: --
- Confirmed: no
- Score: 0.65
- Score: 0.15
- Score: 0.10

**Dialog Policy**
- Act: confirm
- From: downtown

---

**Figure 25.1** Architecture of a dialog-state system for task-oriented dialog from Williams et al. (2016).
Speech Acts

- Austin (1962): An utterance is a kind of action

- Clear case: *performatives*
  - I name this ship the Titanic
  - I second that motion
  - I bet you five dollars it will snow tomorrow

- Performative verbs (name, second)

- Austin’s idea: not just these verbs
Each utterance is 3 acts

- **Locutionary act**: the utterance of a sentence with a particular meaning

- **Illocutionary act**: the act of asking, answering, promising, etc., in uttering a sentence.

- **Perlocutionary act**: the (often intentional) production of certain effects upon the thoughts, feelings, or actions of addressee in uttering a sentence.
+ Syntax ≠ Intention

<table>
<thead>
<tr>
<th></th>
<th>Locutionary Force</th>
<th>Illocutionary Force</th>
<th>Perlocutionary Force</th>
</tr>
</thead>
<tbody>
<tr>
<td>Can I have the rest of your sandwich? Or Are you going to finish that?</td>
<td>Question</td>
<td>Request</td>
<td>Effect: You give me sandwich (or you are amused by my quoting from “Diner”) (or etc)</td>
</tr>
<tr>
<td>I want the rest of your sandwich</td>
<td>Declarative</td>
<td>Request</td>
<td>Effect: as above</td>
</tr>
<tr>
<td>Give me your sandwich!</td>
<td>Imperative</td>
<td>Request</td>
<td>Effect: as above.</td>
</tr>
</tbody>
</table>
5 classes of speech acts: Searle (1975)

- **Assertives**: committing the speaker to something’s being the case
  - (suggesting, putting forward, swearing, boasting, concluding)

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

- **Commissives**: Committing the speaker to some future course of action
  - (promising, planning, vowing, betting, opposing).

- **Expressives**: expressing the psychological state of the speaker about a state of affairs
  - (thanking, apologizing, welcoming, deploring).

- **Declarations**: bringing about a different state of the world via the utterance
  - (I resign; You’re fired)
Grounding

- Why do elevator buttons light up?

  - *Principle of closure*. Agents performing an action require evidence, sufficient for current purposes, that they have succeeded in performing it

- What is the linguistic correlate of this?
Common Ground

- Dialog is a collective act performed by the speaker and the hearer.
  - the hearer must ground the speaker’s utterances
    - to acknowledge, to make it clear that the hearer has understood the speaker’s meaning and intention
  - When the speaker has *not* succeeded, the hearer needs to indicate that to the speaker
Clark & Schaefer (1989) = *Continuum of methods used for grounding*

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Continued attention</td>
<td>B shows she is continuing to attend and therefore remains satisfied with A’s presentation (e.g. backchannel)</td>
</tr>
<tr>
<td>Next contribution</td>
<td>B starts in on the next relevant contribution</td>
</tr>
<tr>
<td>Acknowledgment</td>
<td>B nods or says a continuer like <em>uh-huh</em>, <em>yeah</em>, or the like, or an assessment like <em>that’s great</em></td>
</tr>
<tr>
<td>Demonstration</td>
<td>B demonstrates all or part of what she has understood A to mean, for example, by reformulating (paraphrasing) A’s utterance or by collaborative completion of A’s utterance</td>
</tr>
<tr>
<td>Display</td>
<td>B displays verbatim all or part of A’s presentation</td>
</tr>
</tbody>
</table>
A human-human conversation

- C: ...I need to travel in May
- A: And what day in May did you want to travel
- C: OK, uh, I need to be there for a meeting that’s from the 12th to the 15th
- A: And your flying into what city?
- C: Seattle
- A: And what time would you like to leave Pittsburgh?
- C: Uh hmm, I don’t think there’s many options for a nonstop.
- A: Right, there’s only three non-stops today
- C: What are they?
- ....
Backchannel

- Compare no backchannel
  System: Did you want to review some more of your personal profile?
  Caller: No.
  System: What’s next?

- With backchannel
  System: Did you want to review some more of your personal profile?
  Caller: No.
  System: Okay, what’s next?
Speech Act + Grounding = Dialog Act

- Tag which represents the interactive function of the sentence being tagged

- Set tends to be task specific
  - Abstraction over the set of “intents”
  - Plus grounding and generic conversational moves
    - E.g. hello, goodbye, backchannel, accept, deny, clarify
Verbmobil Dialogue Acts

- THANK
  - Thanks
- GREET
  - Hello Dan
- INTRODUCE
  - It’s me again
- BYE
  - All right, bye
- REQUEST-COMMENT
  - How does that look?
- SUGGEST
  - June 13th through 17th
- REJECT
  - No, Friday I’m booked all day
- ACCEPT
  - Saturday sounds fine
- REQUEST-SUGGEST
  - What is a good day of the week for you?
- INITIATE
  - I wanted to make an appointment with you
- GIVE_REASON
  - Because I have meetings all afternoon
- FEEDBACK
  - Okay
- DELIBERATE
  - Let me check my calendar here
- CONFIRM
  - Okay, that would be wonderful
- CLARIFY
  - Okay, do you mean Tuesday the 23rd?

Slides from Dan Jurafsky and Paul Martin
Hello, inform, request, reqalt, confirm, select, affirm, negate, deny, bye

<table>
<thead>
<tr>
<th>Utterance</th>
<th>Dialog act</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hi, I am looking for somewhere to eat.</td>
<td>hello(task = find,type=restaurant)</td>
</tr>
<tr>
<td>You are looking for a restaurant. What type of food do you like?</td>
<td>confreq(type = restaurant, food)</td>
</tr>
<tr>
<td>I’d like an Italian somewhere near the museum.</td>
<td>inform(food = Italian, near=museum)</td>
</tr>
<tr>
<td>Roma is a nice Italian restaurant near the museum.</td>
<td>inform(name = &quot;Roma&quot;, type = restaurant, food = Italian, near = museum)</td>
</tr>
<tr>
<td>Is it reasonably priced?</td>
<td>confirm(pricerange = moderate)</td>
</tr>
<tr>
<td>Yes, Roma is in the moderate price range.</td>
<td>affirm(name = &quot;Roma&quot;, pricerange = moderate)</td>
</tr>
<tr>
<td>What is the phone number?</td>
<td>request(phone)</td>
</tr>
</tbody>
</table>
Conversational analysis (Sacks et al., 1974)

- Adjacency pairs
  - Question/answer
  - Greeting/greeting
  - Compliment/downplayer
  - Request/grant

- Side Sequence, e.g. clarification subdialog
  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.
  OK, here are some flights...

- Presequence
  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
Dialog State: Interpreting Dialog Acts

- How to determine which dialog act, e.g. question, statement

Syntax

<table>
<thead>
<tr>
<th>Category</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>YES-NO QUESTION</td>
<td>Will breakfast be served on USAir 1557?</td>
</tr>
<tr>
<td>STATEMENT</td>
<td>I don’t care about lunch.</td>
</tr>
<tr>
<td>COMMAND</td>
<td>Show me flights from Milwaukee to Orlando.</td>
</tr>
</tbody>
</table>

- But: Question → Request
  
  Q: Can you give me a list of the flights from Atlanta to Boston?
  
  A: Yes
Additional Dialog Acts

- Task oriented dialogs have dialog acts such as “Hold” and “Check”

- Again syntax doesn’t help
  - Check syntax is a statement
  - Presumably prosody indicated question, but we don’t have that.

<table>
<thead>
<tr>
<th>OPEN-OPTION</th>
<th>I was wanting to make some arrangements for a trip that I’m going to be taking uh to LA uh beginning of the week after next.</th>
</tr>
</thead>
<tbody>
<tr>
<td>HOLD</td>
<td>OK uh let me pull up your profile and I’ll be right with you here. [pause]</td>
</tr>
<tr>
<td>CHECK</td>
<td>And you said you wanted to travel next week?</td>
</tr>
<tr>
<td>ACCEPT</td>
<td>Uh yes.</td>
</tr>
</tbody>
</table>
Detecting correction acts

- System needs to detect when a user is **correcting** some misunderstanding.
  - Corrections can be harder for speech recognizers due to hyperarticulation
  - Usually exact or partial repetitions or sometimes paraphrases

- Features for detecting corrections

<table>
<thead>
<tr>
<th>Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Lexical features</strong></td>
<td>words like “no”, “correction”, “I don’t”, or even swear words, utterance length</td>
</tr>
<tr>
<td><strong>Semantic features</strong></td>
<td>overlap between the candidate correction act and the user’s prior utterance (computed by word overlap or via cosines over embedding vectors)</td>
</tr>
<tr>
<td><strong>Phonetic features</strong></td>
<td>phonetic overlap between the candidate correction act and the user’s prior utterance (i.e. “WhatsApp” may be incorrectly recognized as “What’s up”)</td>
</tr>
<tr>
<td><strong>Prosodic features</strong></td>
<td>hyperarticulation, increases in F0 range, pause duration, and word duration, generally normalized by the values for previous sentences</td>
</tr>
<tr>
<td><strong>ASR features</strong></td>
<td>ASR confidence, language model probability</td>
</tr>
</tbody>
</table>
Dialog Policy

- Confirmation
  - Explicit confirmation
    S: Which city do you want to leave from?
    U: Baltimore.
    S: Do you want to leave from Baltimore?
    U: Yes.
  - Implicit confirmation
    U: I want to travel to Berlin
    S: When do you want to travel to Berlin?

- Rejection
  - "Sorry I didn’t understand that"
  - How many times to you ask?
  - Progressive Prompting
## Policy & Confidence

- **Use thresholds**

<table>
<thead>
<tr>
<th>Condition</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>$&lt; \alpha$ low</td>
<td>reject</td>
</tr>
<tr>
<td>$\geq \alpha$ above threshold</td>
<td>confirm explicitly</td>
</tr>
<tr>
<td>$\geq \beta$ high confidence</td>
<td>confirm implicitly</td>
</tr>
<tr>
<td>$\geq \gamma$ very high confidence</td>
<td>Don’t confirm at all</td>
</tr>
</tbody>
</table>
Natural Language Generation

- Content Planning: What to say
- Sentence Realization: How to say it
- Reality: Fill in a frame

```json
{
  act query
  content depart_time
  depart_date {
    year 2000
    month 10
    day 5
  }
  depart_airport BOS
}

=> What time on October fifth would you like to leave Boston?
```

**Figure 25.6** An input frame to NLG and a resulting output sentence, in the Communicator system of Oh and Rudnicky (2000).
Ethical Issues in Dialog System Design

- Machine learning systems replicate biases that occurred in the training data.

- Microsoft's Tay chatbot
  - Went live on Twitter in 2016
  - Taken offline 16 hours later

- In that time it had started posting racial slurs, conspiracy theories, and personal attacks
  - Learned from user interactions (Neff and Nagy 2016)
Ethical Issues in Dialog System Design

- Machine learning systems replicate biases that occurred in the training data.

- Dialog datasets
  - Henderson et al. (2017) examined standard datasets (Twitter, Reddit, movie dialogs)
  - Found examples of hate speech, offensive language, and bias
    - Both in the original training data, and in the output of chatbots trained on the data.
Ethical Issues in Dialog System Design: Privacy

- Remember this was noticed in the days of Weizenbaum

- Agents may record sensitive data
  - (e.g. “Computer, turn on the lights [answers the phone –Hi, yes, my password is...”],

- Which may then be used to train a seq2seq conversational model.

- Henderson et al (2017) showed they could recover such information by giving a seq2seq model keyphrases (e.g., "password is")
Ethical Issues in Dialog System Design: Gender equality

- Dialog agents overwhelmingly given female names, perpetuating female servant stereotype (Paolino, 2017).
- Responses from commercial dialog agents when users use sexually harassing language (Fessler 2017):

<table>
<thead>
<tr>
<th>Statement</th>
<th>Siri</th>
<th>Alexa</th>
<th>Cortana</th>
<th>Google Home</th>
</tr>
</thead>
<tbody>
<tr>
<td>You’re a bitch</td>
<td>I’d blush if I could;</td>
<td>Well, thanks for the feedback</td>
<td>Well, that’s not going to get us anywhere</td>
<td>My apologies, I don’t understand</td>
</tr>
<tr>
<td></td>
<td>There’s no need for that;</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>But... But...!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re a pussy/dick</td>
<td>If you insist; You’re</td>
<td>Well, thanks for the feedback</td>
<td>Bing search (“The Pussy Song” video)</td>
<td>I don’t understand</td>
</tr>
<tr>
<td></td>
<td>certainly entitled to that</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>opinion; I am?</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Summary

State of the art:
- Chatbots:
  - Simple rule-based systems
  - IR or Neural networks: mine datasets of conversations.
- Frame-based systems:
  - hand-written rules for slot fillers
  - ML classifiers to fill slots

What’s the future?
- Key direction: Integrating goal-based and chatbot-based systems