Speech and Language Processing

Chapter 24: part 1
Dialog(ue) and Conversational Agents

Outline

- The Linguistics of Conversation
- Basic Conversational Agents
  - ASR
  - NLU
  - Generation
  - Dialogue Manager
- Dialogue Manager Design
  - Finite State vs Frame-based
  - Initiative: User, System, Mixed
- VoiceXML
Conversational Agents

- AKA:
  - Spoken Language Systems
  - Dialogue Systems
  - Speech Dialogue Systems

Applications

- Information providing
  - Amtrak’s Julie, CMU Let’s Go
- Customer Care
  - T-Mobile’s Call Center, AT&T Call Routing
- Training
  - Research platforms: e.g. Pitt ItSpoke
- Siri
- Communicating with robots
- Anything with limited screen/keyboard (mobile apps)
- Fun and games....

- Goal: Emulate Human-Human Behavior?
A travel dialog: Communicator

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 eleventh in the morning. Your full name?
U4: [John Doe].
S5: Hello, [John]. I’m glad to hear from you again. You can now ask me for help.
S6: OK. Let’s see I have a United flight departing Pittsburgh at ten forty five a.m. arriving into Seattle at one fifty seven p.m. Is that OK?
U7: yes.
S8: Will you return to Pittsburgh from Seattle?
U9: no i will go to Los Angeles.
S10: Traveling to Los Angeles. On what date?
U11: Sorry can we go back to talk about the previous leg again.
S12: For the first leg, you are on a United flight 1446 departing Pittsburgh at ten forty five A.M. tomorrow, arriving in Seattle at one fifty seven P.M. Do you want to take that?
U13: actually i want to leave on wednesday.
S14: a flight on wednesday.
S15: I have a flight departing Pittsburgh at ten forty five a.m., arrives Seattle at one fifty seven p.m. Is that OK?
U16: yes.

Figure 19.1: The travel domain: a fragment from a successful conversation between a user (U) and the Communicator system (S) of Xu and Rudnicky (2000).

Call routing: ATT HMIHY

S: How may I help you?
U: can you tell me how much it is to tokyo?
S: You want to know the cost of a call?
U: yes that’s right
S: Please hold on for rate information.

S: How may I help you?
U: what’s the area code for chicago?
S: You want an area code.
U: (silence)
S: The area code for Chicago is 312.

Figure 19.2: The call routing/telephone information domain: two actual conversations between a user (U) and the AT&T How May I Help You (HMIHY) system (S) (Gorin et al., 1997).
A tutorial dialogue: ITSPoke

S: I didn’t understand. What force acts on all objects within close proximity to earth?
U: *uh the force of gravity*
S: Fine. Besides the Earth’s gravitational force, are there any other forces acting on the person, keys, and elevator after the release?
U: *no*
S: Yep. Thus, keep in mind that after the release, the only force on the person, keys, and elevator is the force of gravity. If gravity is the only force acting on an object, will it be moving or staying still?
U: *it will be moving*
S: Fine. Not only are the person, keys, and elevator moving, they have only gravitational forces on them. When an object is falling and has only gravitational force on it, it is said to be in *freefall*.

Figure 19.3 The tutoring domain: part of an actual conversation between a student user (U) and the ITSPoke system (S) of (Litman and Silliman, 2004), based on the Why2-Atlas text-based tutoring system (?).

Linguistics of Human Conversation

- Turn-taking
- Speech Acts
- Grounding
- Conversational Structure
- Implicature
Turn-taking

- Dialogue is characterized by turn-taking.
  - A:
  - B:
  - A:
  - B:
  - ...

- Resource allocation problem:
  - How do speakers know when to take the floor?
    - Total amount of overlap relatively small (5% - Levinson 1983)
    - Don't pause either
    - Must be a way to know who should talk and when.

Turn-taking rules

- At each transition-relevance place of each turn:
  - a. If during this turn the current speaker has selected B as the next speaker then B must speak next.
  - b. If the current speaker does not select the next speaker, any other speaker may take the next turn.
  - c. If no one else takes the next turn, the current speaker may take the next turn.
Implications of subrule a

- For some utterances the current speaker selects the next speaker
  - Adjacency pairs
    - Question/answer
    - Greeting/greeting
    - Compliment/downplayer
    - Request/grant
  - Silence between 2 parts of adjacency pair is different than silence after
    - A: Is there something bothering you or not?
      - (1.0)
    - A: Yes or no?
      - (1.5)
    - A: Eh
    - B: No.

Turn-taking Behaviors
Important for SDS

- System understanding:
  - Is the user backchanneling or is she taking the turn (does ‘ok’ mean ‘I agree’ or ‘I’m listening’)?
  - Is this a good place for a system backchannel?

- System generation:
  - How to signal to the user that the system system’s turn is over?
  - How to signal to the user that a backchannel might be appropriate?
Types of Behavior

- **Smooth Switch:** S1 is speaking and S2 speaks and takes and holds the floor
- **Hold:** S1 is speaking, pauses, and continues to speak
- **Backchannel:** S1 is speaking and S2 speaks -- to indicate continued attention -- not to take the floor (e.g. mhmm, ok, yeah)
- How do people coordinate these behaviors with their interlocutor?
- Acoustic-prosodic and lexical cues....

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.

Locutionary and illocutionary

- “You can’t do that!”
- Illocutionary force:
  - Protesting
- Perlocutionary force:
  - Intent to annoy addressee
  - Intent to stop addressee from doing something
The 3 levels of act revisited

<table>
<thead>
<tr>
<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?</td>
<td>Question</td>
<td>Request</td>
</tr>
<tr>
<td>I want the rest of your sandwich</td>
<td>Declarative</td>
<td>Request</td>
</tr>
<tr>
<td>Give me your sandwich!</td>
<td>Imperative</td>
<td>Request</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)
**Dialogue Acts in SDS**

- Roughly correspond to Illocutionary acts
  - Motivation: Improving Spoken Dialogue Systems
  - Many coding schemes (e.g. DAMSL)
  - Many-to-many mapping between DAs and words
    - Agreement DA can be realized by *Okay, Um, Right, Yeah, ...
    - But each of these can express multiple DAs, e.g.

  S: You should take the 10pm flight.
  U: Okay

  ...*that sounds perfect.*
  ...*but I’d prefer an earlier flight.*
  ...*(I’m listening)*

**DA recognition important for**

- Turn recognition (which grammar to use when)
- Turn disambiguation, e.g.

  S: What city do you want to go to?
  U1: Boston. (reply)
  U2: Boston? (request for information)
  S: Do you want to go to Boston?
  U1: Boston. (confirmation)
  U2: Boston? (question)
Automatic DA Detection

- Can we detect DAs automatically?
  - Lexicons are domain-dependent
  - ASR output is errorful
- Typical methods
  - Supervised learning
  - Sequence learning / tagging

Grounding

- Dialogue is a collective act performed by speaker and hearer
- Common ground: set of things mutually believed by both speaker and hearer
- Need to achieve common ground, so hearer must ground or acknowledge speakers utterance.
- Clark (1996):
  - Principle of closure. Agents performing an action require evidence, sufficient for current purposes, that they have succeeded in performing it
  - Need to know whether an action succeeded or failed
Clark and Schaefer: Grounding

- Continued attention: B continues attending to A
- Relevant next contribution: B starts in on next relevant contribution
- Acknowledgement: B nods or says continuer like *uh-huh, yeah, assessment (great!)*
- Demonstration: B demonstrates understanding A by paraphrasing or reformulating A’s contribution, or by collaboratively completing A’s utterance
- Display: B displays verbatim all or part of A’s presentation

Grounding Strategies from Weak to Strong

I need to get your homework by tonight.

- Continued attention
  ...
- Next contribution
  I should be finished after class.
- Acknowledgment
  Mhmm...
- Demonstration
  You need this soon.
- Display
  You need to get my homework by tonight.
A human-human conversation

C1: ...I need to travel in May.
A1: And, what day in May did you want to travel?
C2: OK uh I need to be there for a meeting that’s from the 12th to the 15th.
A2: And you’re flying into what city?
C3: Seattle.
A3: And what time would you like to leave Pittsburgh?
C4: Uh mm I don’t think there’s many options for non-stop.
A4: Right. There’s three non-stops today.
C5: What are they?
A5: 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.
C6: OK I’ll take the 5ish flight on the night before on the 11th.
A6: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air
   flight 115.
C7: OK.

Grounding examples

- **Display:**
  - C: I need to travel in May
  - A: And, what day in May did you want to travel?
- **Acknowledgement**
  - C: He wants to fly from Boston
  - A: mm-hmm
  - C: to Baltimore Washington International
  - [Mm-hmm (usually transcribed “uh-huh”) is a backchannel, continuer, or acknowledgement token]
Grounding Examples (2)

- Acknowledgement + next relevant contribution
  - And, what day in May did you want to travel?
  - And you’re flying into what city?
  - And what time would you like to leave?
- The and indicates to the client that agent has successfully understood answer to the last question.

Grounding and Dialogue Systems

- Grounding is not just a tidbit about humans
- Is key to design of conversational agent
- Why?
Grounding and Dialogue Systems

- Grounding is not just a tidbit about humans
- Is key to design of conversational agent
- Why?
  - HCI researchers find users of speech-based interfaces are confused when system doesn't give them an explicit acknowledgement signal

Discourse Structure

- Dialogue variants of discourse chapter
Pragmatics

- Context-dependent meaning, invited inference, intended meaning – vs. “propositional content”
- Indirect Speech Acts
- Implicature

Conversational Implicature

- 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.**

  - Note that client did not answer question.
  - Meaning of client’s sentence:
    - Meeting
      - Start-of-meeting: 12th
      - End-of-meeting: 15th
      - Doesn't say anything about flying!!!!!
  - What is it that licenses agent to infer that client is mentioning this meeting so as to inform the agent of the travel dates?
Conversational Implicature (2)

- A: ... there’s 3 non-stops today.
- This would still be true if 7 non-stops today.
- But no, the agent means: 3 and only 3.
- How can client infer that agent means:
  - only 3

Grice: conversational implicature

- Implicature means a particular class of licensed inferences.
- Grice (1975) proposed that what enables hearers to draw correct inferences is:
- Cooperative Principle
  - This is a tacit agreement by speakers and listeners to cooperate in communication
4 Gricean Maxims

- Relevance: Be relevant
- Quantity: Do not make your contribution more or less informative than required
- Quality: try to make your contribution one that is true (don’t say things that are false or for which you lack adequate evidence)
- Manner: Avoid ambiguity and obscurity; be brief and orderly

Relevance

- A: *Is Regina here?*
- B: *Her car is outside.*
- Implication: yes
  - Hearer thinks: *why would he mention the car? It must be relevant. How could it be relevant? It could since if her car is here she is probably here.*
- Client: *I need to be there for a meeting that’s from the 12th to the 15th*
  - Hearer thinks: Speaker is following maxims, would only have mentioned meeting if it was relevant. How could meeting be relevant? If client meant me to understand that he had to depart in time for the mtg.
Quantity

- A: How much money do you have on you?
- B: I have 5 dollars
  - Implication: not 6 dollars
- Similarly, 3 non stops can’t mean 7 non-stops (hearer thinks:
  - if speaker meant 7 non-stops she would have said 7 non-stops
- A: Did you do the reading for today’s class?
- B: I intended to
  - Implication: No
  - B’s answer would be true if B intended to do the reading AND did the reading, but would then violate maxim

Dialogue System Architecture

[Diagram of dialogue system architecture]
Speech recognition

- Or ASR (Automatic Speech Recognition)
  - Speech to words
- Input: acoustic waveform
- Output: string of words
- Basic components:
  - a recognizer for phones, small sound units like [k] or [ae].
  - a pronunciation dictionary like cat = [k ae t]
  - a grammar telling us what words are likely to follow what words
  - A search algorithm to find the best string of words

Natural Language Understanding

- Or “NLU”
- Or “Computational semantics”
- There are many ways to represent the meaning of sentences
- For speech dialogue systems, most common is “Frame and slot semantics”. 
An example of a frame

- Show me morning flights from Boston to SF on Tuesday.

SHOW:
FLIGHTS:
ORIGIN:
  CITY: Boston
  DATE: Tuesday
  TIME: morning
DEST:
  CITY: San Francisco

How to generate this semantics?

- Many methods,
- Simplest: "semantic grammars"
- CFG in which the LHS of rules is a semantic category:
  - LIST -> show me | I want | can I see |...
  - DEPARTTIME -> (after|around|before) HOUR  | morning | afternoon | evening
  - HOUR -> one|two|three...|twelve (am|pm)
  - FLIGHTS -> (a) flight|flights
  - ORIGIN -> from CITY
  - DESTINATION -> to CITY
  - CITY -> Boston | San Francisco | Denver | Washington
Semantics for a sentence

LIST   FLIGHTS   ORIGIN
Show me flights from Boston

DESTINATION   DEPARTDATE
to San Francisco on Tuesday

DEPARTTIME
morning

Generation and TTS

- Generation component
  - Chooses concepts to express to user
  - Plans out how to express these concepts in words
  - Assigns any necessary prosody to the words
- TTS component
  - Takes words and prosodic annotations
  - Synthesizes a waveform
Generation Component

- Content Planner
  - Decides what content to express to user
    - (ask a question, present an answer, etc)
  - Often merged with dialogue manager

- Language Generation
  - Chooses syntactic structures and words to express meaning.
  - Simplest method
    - All words in sentence are prespecified!
    - "Template-based generation"
    - Can have variables:
      - What time do you want to leave CITY-ORIG?
      - Will you return to CITY-ORIG from CITY-DEST?

More sophisticated language generation component

- Natural Language Generation
- This is a field, like Parsing, or Natural Language Understanding, or Speech Synthesis, with its own (small) conference

- Approach:
  - Dialogue manager builds representation of meaning of utterance to be expressed
  - Passes this to a “generator”
  - Generators have three components
    - Sentence planner
    - Surface realizer
    - Prosody assigner
Architecture of a generator for a dialogue system
(after Walker and Rambow 2002)

HCI constraints on generation for dialogue: "Coherence"

- Discourse markers and pronouns ("Coherence"):
  1. Please say the date.
  1. Please say the start time.
  1. Please say the duration...
  1. Please say the subject...
  2. First, tell me the date.
  2. Next, I'll need the time it starts.
  2. Thanks. <pause> Now, how long is it supposed to last?
  2. Last of all, I just need a brief description
HCI constraints on generation for dialogue: coherence (II): tapered prompts

- Prompts which get incrementally shorter:
  - System: Now, what's the first company to add to your watch list?
  - Caller: Cisco
  - System: What's the next company name? (Or, you can say, “Finished”)
  - Caller: IBM
  - System: Tell me the next company name, or say, “Finished.”
  - Caller: Intel
  - System: Next one?
  - Caller: America Online.
  - System: Next?
  - Caller: ...

Dialogue Manager

- Controls the architecture and structure of dialogue
  - Takes input from ASR/NLU components
  - Maintains some sort of state
  - Interfaces with Task Manager
  - Passes output to NLG/TTS modules
Four architectures for dialogue management

- Finite State
- Frame-based
- Information State
  - Markov Decision Processes
  - AI Planning

Finite-State Dialogue Mgmt

- Consider a trivial airline travel system
  - Ask the user for a departure city
  - For a destination city
  - For a time
  - Whether the trip is round-trip or not
**Finite State Dialogue Manager**

- What city are you leaving from?
- Where are you going?
- What date do you want to leave?
- Is it a one-way trip?
  - Yes
    - Do you want to go from <FROM> to <TO> on <DATE>?
    - No
    - What do you want to return?
  - No
    - Do you want to go from <FROM> to <TO> on <DATE> returning on <RETURN>?
    - Yes
    - Book the flight
    - No

**Finite-state dialogue managers**

- System completely controls the conversation with the user.
- It asks the user a series of questions
- Ignoring (or misinterpreting) anything the user says that is not a direct answer to the system’s questions
Dialogue Initiative

- Systems that control conversation like this are system initiative or single initiative.
- "Initiative": who has control of conversation
- In normal human-human dialogue, initiative shifts back and forth between participants.

System Initiative

- Systems which completely control the conversation at all times are called system initiative.
- Advantages:
  - Simple to build
  - User always knows what they can say next
  - System always knows what user can say next
    - Known words: Better performance from ASR
    - Known topic: Better performance from NLU
  - Ok for VERY simple tasks (entering a credit card, or login name and password)
- Disadvantage:
  - Too limited
User Initiative

- User directs the system
- Generally, user asks a single question, system answers
- System can’t ask questions back, engage in clarification dialogue, confirmation dialogue
- Used for simple database queries
- User asks question, system gives answer
- Web search is user initiative dialogue.

Problems with System Initiative

- Real dialogue involves give and take!
- In travel planning, users might want to say something that is not the direct answer to the question.
- For example answering more than one question in a sentence:
  - Hi, I’d like to fly from Seattle Tuesday morning
  - I want a flight from Milwaukee to Orlando one way leaving after 5 p.m. on Wednesday.
Single initiative + universals

- We can give users a little more flexibility by adding universal commands
- Universals: commands you can say anywhere
- As if we augmented every state of FSA with these
  - Help
  - Start over
  - Correct
- This describes many implemented systems
- But still doesn’t allow user to say what they want to say

Mixed Initiative

- Conversational initiative can shift between system and user
- Simplest kind of mixed initiative: use the structure of the frame itself to guide dialogue

<table>
<thead>
<tr>
<th>Slot</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>ORIGIN</td>
<td>What city are you leaving from?</td>
</tr>
<tr>
<td>DEST</td>
<td>Where are you going?</td>
</tr>
<tr>
<td>DEPT DATE</td>
<td>What day would you like to leave?</td>
</tr>
<tr>
<td>DEPT TIME</td>
<td>What time would you like to leave?</td>
</tr>
<tr>
<td>AIRLINE</td>
<td>What is your preferred airline?</td>
</tr>
</tbody>
</table>
Frames are mixed-initiative

- User can answer multiple questions at once.
- System asks questions of user, filling any slots that user specifies.
- When frame is filled, do database query.
- If user answers 3 questions at once, system has to fill slots and not ask these questions again!
- Anyhow, we avoid the strict constraints on order of the finite-state architecture.

Defining Mixed Initiative

- Mixed Initiative could mean
  - User can arbitrarily take or give up initiative in various ways
    - This is really only possible in very complex plan-based dialogue systems
    - No commercial implementations
    - Important research area
  - Something simpler and quite specific which we will define in the next few slides
True Mixed Initiative

C₁: ...I need to travel in May.
A₁: And, what day in May did you want to travel?
C₂: OK uh I need to be there for a meeting that’s from the 12th to the 15th.
A₂: And you’re flying into what city?
C₃: Seattle.
A₃: And what time would you like to leave Pittsburgh?
C₄: Uh hmm I don’t think there’s many options for non-stop.
A₄: Right. There’s three non-stops today.
C₅: What are they?
A₅: 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.
C₆: OK I’ll take the 5ish flight on the night before on the 11th.
C₇: OK.

How mixed initiative is usually defined

- First we need to define two other factors
- Open prompts vs. directive prompts
- Restrictive versus non-restrictive grammar
Open vs. Directive Prompts

- **Open prompt**
  - System gives user very few constraints
  - User can respond how they please:
    - “How may I help you?” “How may I direct your call?”

- **Directive prompt**
  - Explicit instructs user how to respond
  - “Say yes if you accept the call; otherwise, say no”

Restrictive vs. Non-restrictive grammars

- **Restrictive grammar**
  - Language model which strongly constrains the ASR system, based on dialogue state

- **Non-restrictive grammar**
  - Open language model which is not restricted to a particular dialogue state
### Definition of Mixed Initiative

<table>
<thead>
<tr>
<th>Grammar</th>
<th>Open Prompt</th>
<th>Directive Prompt</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restrictive</td>
<td><em>Doesn't make sense</em></td>
<td><em>System Initiative</em></td>
</tr>
<tr>
<td>Non-restrictive</td>
<td><em>User Initiative</em></td>
<td><em>Mixed Initiative</em></td>
</tr>
</tbody>
</table>

### VoiceXML

- Voice eXtensible Markup Language
- An XML-based dialogue design language
- Makes use of ASR and TTS
- Deals well with simple, frame-based mixed initiative dialogue.
- Most common in commercial world (too limited for research systems)
- But useful to get a handle on the concepts.
Voice XML

- Each dialogue is a `<form>`.
- Each `<form>` generally consists of a sequence of `<field>`s, with other commands

Sample vxml doc

```xml
<form>
  <field name="transporttype">
    <prompt>
      Please choose airline, hotel, or rental car.  </prompt>
    <grammar type="application/x-nuance-gsl">
      [airline hotel "rental car"]
    </grammar>
  </field>
  <block>
    <prompt>
      You have chosen <value expr="transporttype">. </prompt>
  </block>
</form>
```
VoiceXML interpreter

- Walks through a VXML form in document order
- Iteratively selecting each item
- If multiple fields, visit each one in order.
- Special commands for events

Another vxml doc (1)

```xml
<noinput>
I'm sorry, I didn't hear you. <reprompt/>
</noinput>
- "noinput" means silence exceeds a timeout threshold

<nomatch>
I'm sorry, I didn't understand that. <reprompt/>
</nomatch>
- "nomatch" means confidence value for utterance is too low
- notice "reprompt" command```
**Summary: VoiceXML**

- Voice eXtensible Markup Language
- An XML-based dialogue design language
- Makes use of ASR and TTS
- Deals well with simple, frame-based mixed initiative dialogue.
- Most common in commercial world (too limited for research systems)
- But useful to get a handle on the concepts.