CS 224S LING 281
Speech Recognition and Synthesis

Lecture 13: Dialogue and Conversational Agents (I)

Dan Jurafsky

Outline

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

Conversational Agents

- AKA:
  - Spoken Language Systems
  - Dialogue Systems
  - Speech Dialogue Systems
- Applications:
  - Travel arrangements (Amtrak, United airlines)
  - Telephone call routing
  - Tutoring
  - Communicating with robots
  - Anything with limited screen/keyboard

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 what?
U:	*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.

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

	Locutionary Force	Illocutionary Force	Perlocutionary Force
Can I have the rest of your sandwich?	Question	Request	Intent: You give me sandwich
I want the rest of your sandwich	Declarative	Request	Intent: You give me sandwich
Give me your sandwich!	Imperative	Request	Intent: You give me sandwich

Illocutionary Acts

- What are they?

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

- 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
- (Interestingly, Clark points out that this idea draws from Norman (1988) work on non-linguistic acts)
- 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

A human-human conversation

C_1:	. . . I need to travel in May.
A_1:	And, what day in May did you want to travel?
C_2:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2:	And you're flying into what city?
C_3:	Seattle.
A_3:	And what time would you like to leave Pittsburgh?
C_4:	Uh hmm I don't think there's many options for non-stop.
A_4:	Right. There's three non-stops today.
C_5:	What are they?
A_5:	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_6:	OK I'll take the 5ish flight on the night before on the 11th.
A_6:	On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7:	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 negative responses

From Cohen et al. (2004)

- System: Did you want to review some more of your personal profile?
- Caller: No.
- System: Okay, what's next?

Good!

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

Bad!

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
  - Stifelman et al. (1993), Yankelovich et al. (1995)

Conversational Structure

- Telephone conversations
  - Stage 1: Enter a conversation
  - Stage 2: Identification
  - Stage 3: Establish joint willingness to converse
  - Stage 4: First topic is raised, usually by caller

Stage	Speaker & Utterance
1	A_1: (rings B's telephone)
1,2	B_1: Benjamin Holloway
2	A_1: this is Professor Dwight's secretary, from Polymania College
2,3	B_1: ooh yes –
4	A_1: uh:m . about the: lexicology *seminar*
4	B_1: *yes*

Why is this customer confused?

- Customer: (rings)
- Operator: Directory Enquiries, for which town please?
- Customer: Could you give me the phone number of um: Mrs. um: Smithson?
- Operator: Yes, which town is this at please?
- Customer: Huddleston.
- Operator: Yes. And the name again?
- Customer: Mrs. Smithson

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

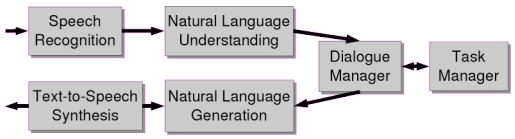
Relevance

- 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



ASR engine

- Standard ASR engine that we've seen
  - Speech to words
- But specific characteristics for dialogue
  - Language models could depend on where we are in the dialogue
  - Could make use of the fact that we are talking to the same human over time.
  - Barge-in (human will talk over the computer)
  - Confidence values
    - (As we will see), we want to know if we misunderstood the human.

Language Model

- Language models for dialogue are often based on hand-written Context-Free or finite-state grammars rather than N-grams
- Why? Because of need for understanding; we need to constrain user to say things that we know what to do with.

Language Models for Dialogue (2)

- We can have LM specific to a dialogue state
- If system just asked "What city are you departing from?"
- LM can be
  - City names only
  - FSA: (I want to (leave|depart)) (from) [CITYNAME]
  - N-grams trained on answers to "Cityname" questions from labeled data
- A LM that is constrained in this way is technically called a "restricted grammar" or "restricted LM"

Talking to the same human over the whole conversation.

- Same speaker
- So can adapt to speaker
  - Acoustic Adaptation
    - Vocal Tract Length Normalization (VTLN)
    - Maximum Likelihood Linear Regression (MLLR)
  - Language Model adaptation
  - Pronunciation adaptation

Barge-in

- Speakers barge-in
- Need to deal properly with this via speech-detection, etc.

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

Frame-filling

- We use a parser to take these rules and apply them to the sentence.
- Resulting in a semantics for the sentence
- We can then write some simple code
- That takes the semantically labeled sentence
- And fills in the frame.

Other NLU Approaches

- Syntactic rules with semantic attachments
  - This latter is what is done in VoiceXML
- Cascade of Finite-State-Transducers
  - In practice, many rules have no recursion
  - So don't need CFG
  - Can use finite automata instead

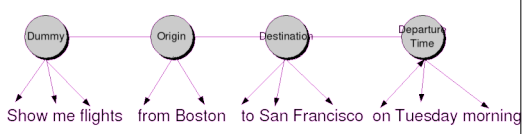
Problems with any of these semantic grammars

- Relies on hand-written grammar
  - Expensive
  - May miss possible ways of saying something if the grammar-writer just doesn't think about them
- Not probabilistic
  - In practice, every sentence is ambiguous
  - Probabilities are best way to resolve ambiguities
  - We know a lot about how to learn and build good statistical models!

HMMs for semantics

- Idea: use an HMM for semantics, just as we did for part-of-speech tagging and for speech recognition
- Hidden units:
  - Semantic slot names
    - Origin
    - Destination
    - Departure time
- Observations:
  - Word sequences

HMM model of semantics - Pieraccini et al (1991)



Semantic HMM

- Goal of HMM model:
  - to compute labeling of semantic roles C = $c_1, c_2, \ldots, c_n$ (C for 'cases' or 'concepts')
  - that is most probable given words W

$$\begin{aligned}\operatorname*{argmax}_C P(C \mid W) &= \operatorname*{argmax}_C \frac{P(W \mid C)P(C)}{P(W)} \\ &= \operatorname*{argmax}_C P(W \mid C)P(C) \\ &= \operatorname*{argmax}_C \prod_{i=2}^{N} P(w_i \mid w_{i-1} \ldots w_1, C) P(w_1 \mid C) \prod_{i=2}^{M} P(c_i \mid c_{i-1} \ldots c_1)\end{aligned}$$

Semantic HMM

- From previous slide:

$$= \underset{C}{\operatorname{argmax}} \prod_{i=2}^{N} P(w_i \mid w_{i-1}...w_1, C) P(w_1 \mid C) \prod_{i=2}^{M} P(c_i \mid c_{i-1}...c_1)$$

- Assume simplification:

$$P(w_i \mid w_{i-1}...w_1, C) = P(w_i \mid w_{i-1}, ..., w_{i-N+1}, c_i)$$

- Final form:

$$P(c_i \mid c_{i-1}...c_1, C) = P(c_i \mid c_{i-1}, ..., c_{i-M+1})$$

$$= \underset{C}{\operatorname{argmax}} \prod_{i=2}^{N} P(w_i \mid w_{i-1}...w_{i-N+1}, c_i) \prod_{i=2}^{M} P(c_i \mid c_{i-1}...c_{i-M+1})$$

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

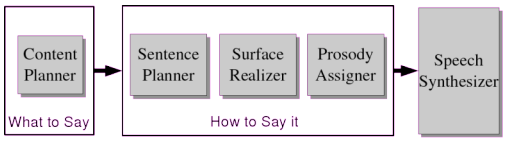
Generation Component

- 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.
Please say the start time.
Please say the duration…
Please say the subject…
(2) First, tell me the date.
Next, I'll need the time it starts.
Thanks. <pause> Now, how long is it supposed to last?
Last of all, I just need a brief description

Bad!

Good!

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

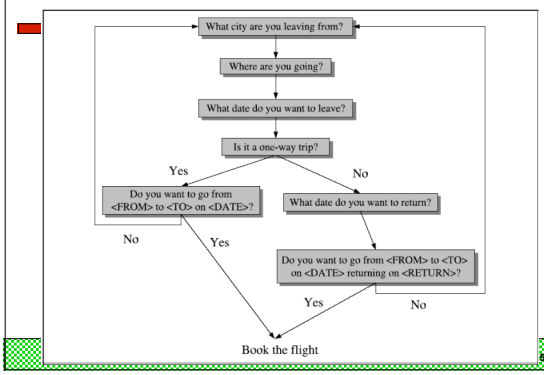
Four architectures for dialogue management

- 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



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 the 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

Slot	Question
ORIGIN	What city are you leaving from?
DEST	Where are you going?
DEPT DATE	What day would you like to leave?
DEPT TIME	What time would you like to leave?
AIRLINE	What is your preferred airline?

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.

Multiple frames

- flights, hotels, rental cars
- Flight legs: Each flight can have multiple legs, which might need to be discussed separately
- Presenting the flights (If there are multiple flights meeting users constraints)
  - It has slots like 1ST_FLIGHT or 2ND_FLIGHT so user can ask "how much is the second one"
- General route information:
  - Which airlines fly from Boston to San Francisco
- Airfare practices:
  - Do I have to stay over Saturday to get a decent airfare?

Multiple Frames

- Need to be able to switch from frame to frame
- Based on what user says.
- Disambiguate which slot of which frame an input is supposed to fill, then switch dialogue control to that frame.
- Main implementation: production rules
  - Different types of inputs cause different productions to fire
  - Each of which can flexibly fill in different frames
  - Can also switch control to different frame

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

C_1: . . . I need to travel in May.
A_1: And, what day in May did you want to travel?
C_2: OK uh I need to be there for a meeting that's from the 12th to the 15th.
A_2: And you're flying into what city?
C_3: Seattle.
A_3: And what time would you like to leave Pittsburgh?
C_4: Uh hmm I don't think there's many options for non-stop.
A_4: Right. There's three non-stops today.
C_5: What are they?
A_5: 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_6: OK I'll take the 5ish flight on the night before on the 11th.
A_6: On the 11th? OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
C_7: 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 gramamrs

- 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

Grammar	Open Prompt	Directive Prompt
Restrictive	*Doesn't make sense*	System Initiative
Non-restrictive	User Initiative	Mixed Initiative