Dialogue & Conversational Agent

Le Anh Cuong

Reading

- Chapter 22 [1]

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 (?).

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

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.

More on Turn-Taking

- Turn-taking behaviors in human-human conversation
  - Task/circumstance dependencies
  - Linguistic/cultural differences
  - How do we take and give up turns?

Expectations of What to Say May Depend on Task at Hand

- Telephone
  - Openings

    Pat: Hello?

    Chris: Hi, Pat. It's Chris.

    Pat: Hi!
  - Closings (6-turn)

    Chris: Well, I just wanted to see how you were doing

    Pat: Thanks for calling. We'll have to have lunch sometime

    Chris: I'd like to

    Pat: Okay

    Chris: Okay

    Pat: See you

    Chris: Yeah, see you

- Email
  - Pat: "Hi, can we switch lunch to 12:30? I'm running late."
  - Chris: "Sure. 12:30."
  - Pat: "Great. See you."
- Service encounters
  - Clerk: Good morning. Is there something I can help you with?
  - Pat: Hi. Yeah. I wonder if you could show me….
- Meetings
  - Boss: Today I want to focus on next year's goal statements. Chris, could you report please….
  - Chris: …
  - Boss: Pat, now let's hear from you…
  - Pat: …
- News broadcasts
  - Anchor: …Chris Smith reports from Rome now on the upcoming conclave. Chris?
  - Reporter: Thanks, Pat….. And now back to Pat Jones in New York.

Cultural Differences in Turn-Taking

- Chinese telephone conversations
  - Openings (Zhu '04)
    - Mandarin vs. British
    - Identification differences
      - British self-report
      - Chinese callees ask the caller
  - Closings (Sun '05)
    - 39 female-female telephone conversations
    - Closings initiated through matter-of-fact statement of intention to end conversation
    - Verbalized thanking occurs except in mother/daughter closings – not the standard English model
  - Finnish business calls (Halmari '93) vs. American
    - Americans get right to the point
    - Finns chat

Individual Differences: British Politicians (Beattie '82)

- Data: 25m televised interviews before 1979 British General election
  - Margaret Thatcher (Tory leader): the Iron Lady
  - Jim Callaghan (Prime Minister): Sunny Jim
- Who interrupts?
  - Less intelligent, highly neurotic, extroverted
  - Men interrupt women
  - Interruptions may indicate
    - Desire for dominance
    - Desire for social approval
    - Conveyance of 'joint enthusiasm', heightened involvement

- Method:
  - Identify spkr 2 attempts to take the turn
    - Smooth switches: no simultaneous speech, spkr 1's utterance complete, turn to spkr 2
    - Simple interruptions: simultaneous speech, spkr 1 doesn't complete utterance, turn to spkr 2
    - Overlap: simultaneous speech, spkr 1 completes utterance, turn to spkr 2
    - Butting-in: simultaneous speech but no change of turn, spkr 1 keeps the turn
    - Silent interruption: spkr 1's utterance incomplete, no simultaneous speech, turn to spkr 2

- Analyze acoustic/prosodic and gestural information
  - Turn-yielding behavior
    - Pauses
    - Speaking rate slows
    - Drawl at end of clause
    - Drop in pitch or loudness
    - Completion of syntactic clause
    - Gesture of termination
  - Attempt suppression signals
    - Filled pauses
    - Gestures

Results

- Mrs. Thatcher interrupted almost twice as often as she interrupts interviewer (19/10)– unlike Callaghan (14/23)
  - Thatcher: Starts slow and gets faster, few FPs (4)
  - Callaghan: starts fast and gets slower, many FPs (22)
- Public perception: Thatcher is domineering in interviews and Callaghan is a 'nice guy'
  - But Thatcher does not dominate
  - Why is Thatcher interrupted?
    - Interruptions come at end of syntactic clause when drawl on stressed syllable in clause and falling intonation

- No suppression signals
- Why does she do this?
  - Speech training before election?
- Why is she still perceived as domineering?
  - When interrupted she doesn't cede the floor despite lengthy stretches of simultaneous speech

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	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.

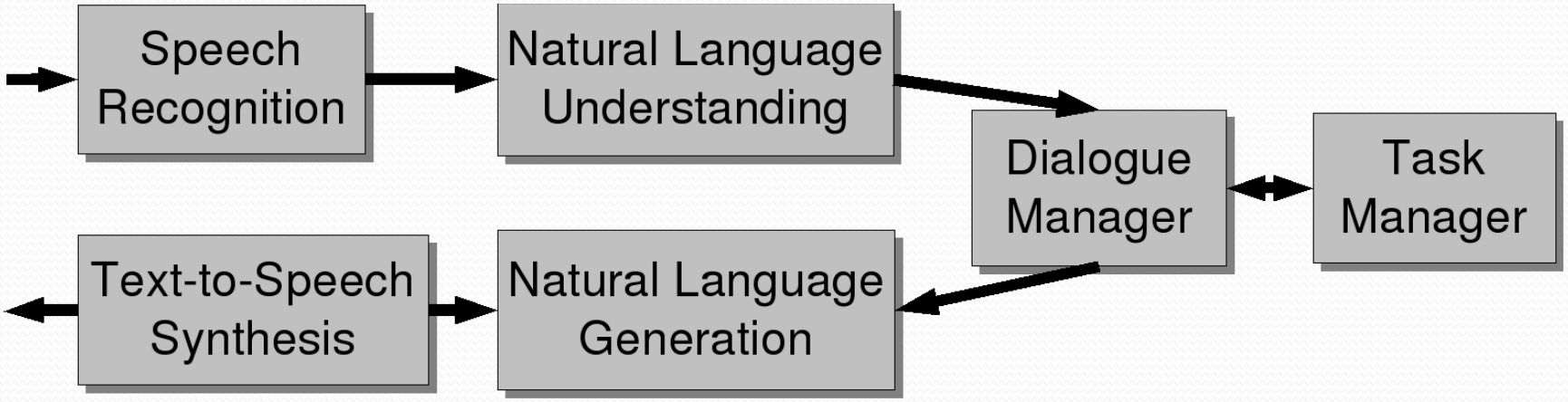
Quantity

- 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

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

Dialogue System Architecture



ASR engine

- Standard Automatic Speech Recognition (ASR) engine
  - 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

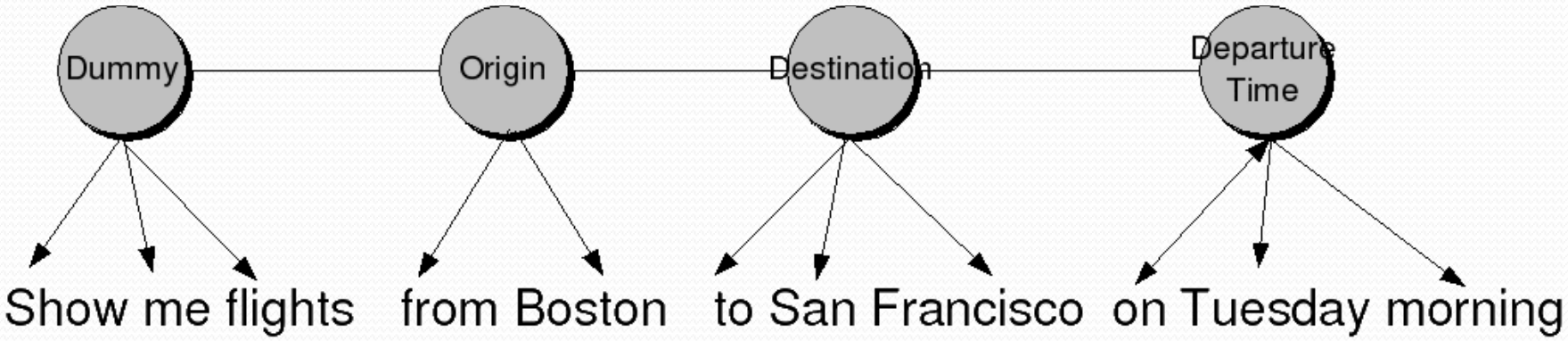
Problems with any of these semantic grammars

- 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

$$\underset{C}{\operatorname{argmax}} P(C \mid W) = \underset{C}{\operatorname{argmax}} \frac{P(W \mid C)P(C)}{P(W)}$$

$$= \underset{C}{\operatorname{argmax}} P(W \mid C)P(C)$$

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

Semantic HMM

- From previous slide: $= \underset{C}{\operatorname{argmax}} \prod_{i=2}^{N} P(w_i \mid w_{i-1} \ldots w_1, C) P(w_1 \mid C) \prod_{i=2}^{M} P(c_i \mid c_{i-1} \ldots c_1)$
- Assume simplification: $P(w_i \mid w_{i-1} \ldots w_1, C) = P(w_i \mid w_{i-1}, \ldots, w_{i-N+1}, c_i)$
- Final form: $P(c_i \mid c_{i-1} \ldots c_1, C) = P(c_i \mid c_{i-1}, \ldots, c_{i-M+1})$

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

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

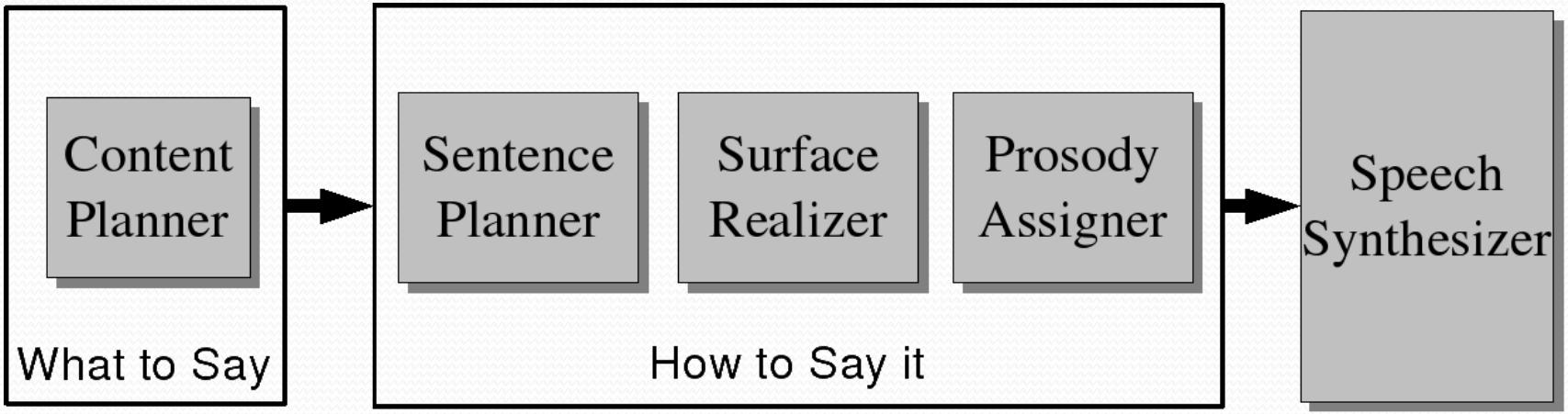
Generation Component

- 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 [illegible]

…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

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

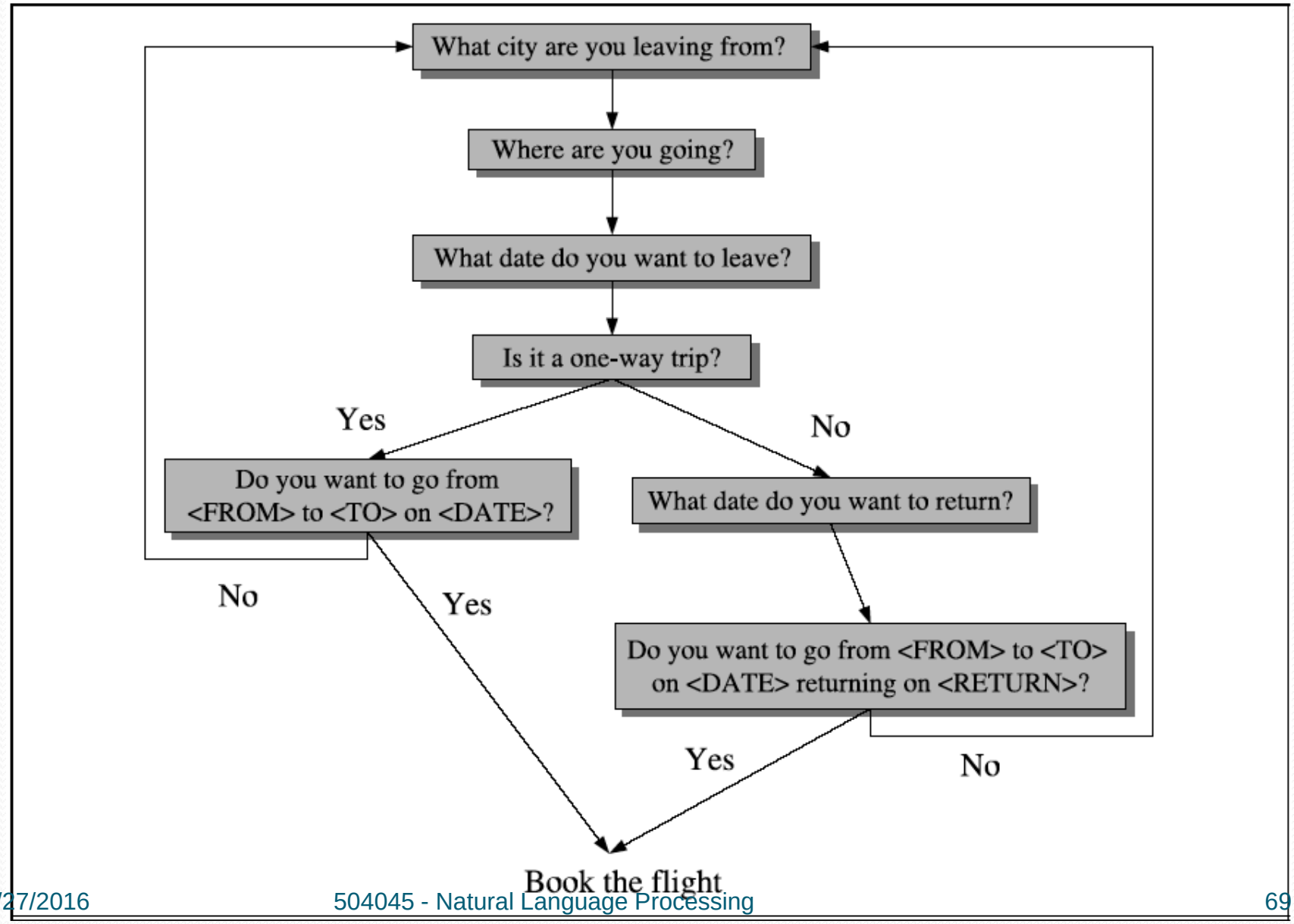
Four architectures for dialogue management

- 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

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

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

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>. (Form is the VoiceXML word for frame)
- Each <form> generally consists of a sequence of <field>s, with other commands

Sample vxml doc

```
<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)

<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

Another vxml doc (2)

```
<form>
 <block>    Welcome to the air travel consultant.  </block>
 <field name="origin">
    <prompt>     Which city do you want to leave from?  </prompt>
    <grammar type="application/x=nuance-gsl">
     [(san francisco) denver (new york)  barcelona]
    </grammar>
    <filled>
      <prompt>  OK, from <value expr="origin">  </prompt>
    </filled>
 </field>
```

- “filled” tag is executed by interpreter as soon as field filled by user

Another vxml doc (3)

```
<field name="destination">
    <prompt>  And which city do you want to go to?   </prompt>
    <grammar type="application/x=nuance-gsl">
     [(san francisco) denver (new york)  barcelona]
    </grammar>
    <filled>
      <prompt>   OK, to <value expr="destination">   </prompt>
    </filled>
 </field>
 <field name="departdate" type="date">
    <prompt>  And what date do you want to leave?  </prompt>
    <filled>
      <prompt>   OK, on <value expr="departdate">   </prompt>
    </filled>
 </field>
```

Another vxml doc (4)

```
<block>
    <prompt> OK, I have you are departing from
        <value expr="origin”>  to <value
  expr="destination”> on <value expr="departdate">
    </prompt>
    send the info to book a flight…
  </block>
</form>
```

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.

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

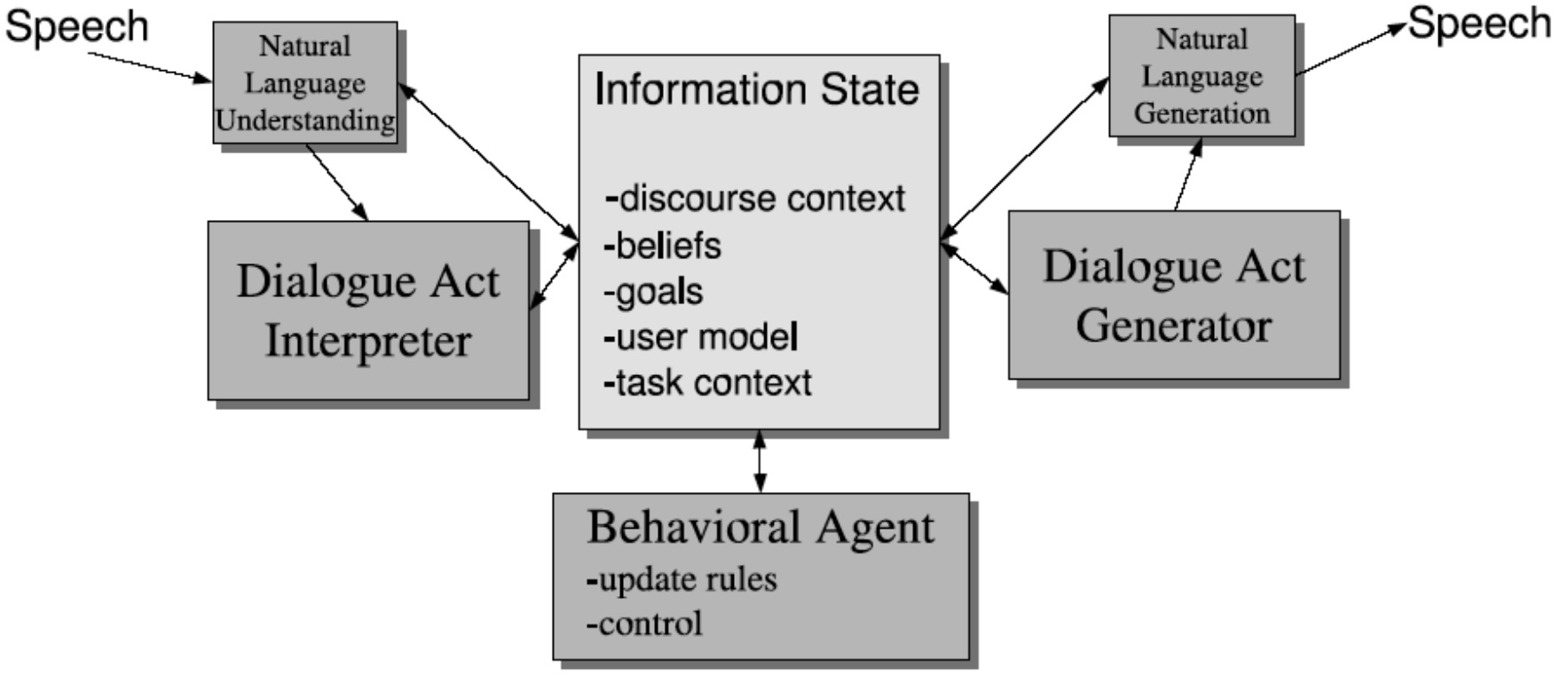
Information-State and Dialogue Acts

- If we want a dialogue system to be more than just form-filling
- Needs to:
  - Decide when the user has asked a question, made a proposal, rejected a suggestion
  - Ground a user's utterance, ask clarification questions, suggestion plans
- Suggests:
  - Conversational agent needs sophisticated models of interpretation and generation
    - In terms of speech acts and grounding
    - Needs more sophisticated representation of dialogue context than just a list of slots

Information-state architecture

- Information state
- Dialogue act interpreter
- Dialogue act generator
- Set of update rules
  - Update dialogue state as acts are interpreted
  - Generate dialogue acts
- Control structure to select which update rules to apply

Information-state



Dialogue acts

- Also called “conversational moves”
- An act with (internal) structure related specifically to its dialogue function
- Incorporates ideas of grounding
- Incorporates other dialogue and conversational functions that Austin and Searle didn’t seem interested in

Verbmobil task

- Two-party scheduling dialogues
- Speakers were asked to plan a meeting at some future date
- Data used to design conversational agents which would help with this task
- (cross-language, translating, scheduling assistant)

Verbmobil Dialogue Acts

THANK	thanks
GREET	Hello Dan
INTRODUCE	It's me again
BYE	Allright, 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?
INIT	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?

DAMSL: forward looking func.

STATEMENT	a claim made by the speaker
INFO-REQUEST	a question by the speaker
CHECK	a question for confirming information
INFLUENCE-ON-ADDRESSEE	(=Searle's directives)
OPEN-OPTION	a weak suggestion or listing of options
ACTION-DIRECTIVE	an actual command
INFLUENCE-ON-SPEAKER	(=Austin's commissives)
OFFER	speaker offers to do something
COMMIT	speaker is committed to doing something
CONVENTIONAL	other
OPENING	greetings
CLOSING	farewells
THANKING	thanking and responding to thanks

DAMSL: backward looking func.

AGREEMENT speaker's response to previous proposal
- ACCEPT accepting the proposal
- ACCEPT-PART accepting some part of the proposal
- MAYBE neither accepting nor rejecting the proposal
- REJECT-PART rejecting some part of the proposal
- REJECT rejecting the proposal
- HOLD putting off response, usually via subdialogue

ANSWER answering a question

UNDERSTANDING whether speaker understood previous
- SIGNAL-NON-UNDER. speaker didn't understand
- SIGNAL-UNDER. speaker did understand
  - ACK demonstrated via continuer or assessment
  - REPEAT-REPHRASE demonstrated via repetition or reformulation
  - COMPLETION demonstrated via collaborative completion

[assert]	C_1:	. . . I need to travel in May.
[info-req,ack]	A_1:	And, what day in May did you want to travel?
[assert, answer]	C_2:	OK uh I need to be there for a meeting that's from the 12th to the 15th.
[info-req,ack]	A_2:	And you're flying into what city?
[assert,answer]	C_3:	Seattle.
[info-req,ack]	A_3:	And what time would you like to leave Pittsburgh?
[check,hold]	C_4:	Uh hmm I don't think there's many options for non-stop.
[accept,ack]	A_4:	Right.
[assert]		There's three non-stops today.
[info-req]	C_5:	What are they?
[assert, open-option]	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.
[accept,ack]	C_6:	OK I'll take the 5ish flight on the night before on the 11th.
[check,ack]	A_6:	On the 11th?
[assert,ack]		OK. Departing at 5:55pm arrives Seattle at 8pm, U.S. Air flight 115.
[ack]	C_7:	OK.

Automatic Interpretation of Dialogue Acts

- How do we automatically identify dialogue acts?
- Given an utterance:
  - Decide whether it is a QUESTION, STATEMENT, SUGGEST, or ACK
- Recognizing illocutionary force will be crucial to building a dialogue agent
- Perhaps we can just look at the form of the utterance to decide?

Can we just use the surface syntactic form?

- YES-NO-Q's have auxiliary-before-subject syntax:
  - Will breakfast be served on USAir 1557?
- STATEMENTs have declarative syntax:
  - I don't care about lunch
- COMMAND's have imperative syntax:
  - Show me flights from Milwaukee to Orlando on Thursday night

Surface form != speech act type

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

Dialogue act disambiguation is hard! Who's on First?

Abbott: Well, Costello, I'm going to New York with you. Bucky Harris the Yankee's manager gave me a job as coach for as long as you're on the team.
Costello: Look Abbott, if you're the coach, you must know all the players.
Abbott: I certainly do.
Costello: Well you know I've never met the guys. So you'll have to tell me their names, and then I'll know who's playing on the team.
Abbott: Oh, I'll tell you their names, but you know it seems to me they give these ball players now-a-days very peculiar names.
Costello: You mean funny names?
Abbott: Strange names, pet names...like Dizzy Dean...
Costello: His brother Daffy Abbott: Daffy Dean...
Costello: And their French cousin.
Abbott: French?
Costello: Goofe'
Abbott: Goofe' Dean. Well, let's see, we have on the bags, Who's on first, What's on second, I Don't Know is on third...
Costello: That's what I want to find out.
Abbott: I say Who's on first, What's on second, I Don't Know's on third.

Dialogue act ambiguity

- Who's on first?
  - INFO-REQUEST
  - or
  - STATEMENT

Dialogue Act ambiguity

- Can you give me a list of the flights from Atlanta to Boston?
  - This looks like an INFO-REQUEST.
  - If so, the answer is:
    - YES.
  - But really it's a DIRECTIVE or REQUEST, a polite form of:
  - Please give me a list of the flights…
- What looks like a QUESTION can be a REQUEST

Dialogue Act ambiguity

- Similarly, what looks like a STATEMENT can be a QUESTION:

Us	OPEN-OPTION	**I was wanting to make some arrangements for a trip that I'm going to be taking uh to LA uh beginnning of the week after next**
Ag	HOLD	**OK uh let me pull up your profile and I'll be right with you here. [pause]**
Ag	CHECK	And you said you wanted to travel next week?
Us	ACCEPT	**Uh yes.**

Indirect speech acts

- Utterances which use a surface statement to ask a question
- Utterances which use a surface question to issue a request

DA interpretation as statistical classification

- Lots of clues in each sentence that can tell us which DA it is:
- Words and Collocations:
  - *Please* or *would you*: good cue for REQUEST
  - *Are you*: good cue for INFO-REQUEST
- Prosody:
  - Rising pitch is a good cue for INFO-REQUEST
  - Loudness/stress can help distinguish *yeah*/AGREEMENT from *yeah*/BACKCHANNEL
- Conversational Structure
  - *Yeah* following a proposal is probably AGREEMENT; *yeah* following an INFORM probably a BACKCHANNEL

HMM model of dialogue act interpretation

- A dialogue is an HMM
- The hidden states are the dialogue acts
- The observation sequences are sentences
  - Each observation is one sentence
    - Including words and acoustics
- The observation likelihood model includes
  - N-grams for words
  - Another classifier for prosodic cues
- Summary: 3 probabilistic models:
  - A: Conversational Structure: Probability of one dialogue act following another P(Answer|Question)
  - B: Words and Syntax: Probability of a sequence of words given a dialogue act: P("do you" | Question)
  - B: Prosody: probability of prosodic features given a dialogue act : P("rise at end of sentence" | Question)

HMMs for dialogue act interpretation

- Goal of HMM model:
  - to compute labeling of dialogue acts $D = d_1, d_2, \ldots, d_n$
  - that is most probable given evidence E

$$D^* = \underset{D}{\operatorname{argmax}} P(D \mid E) = \underset{D}{\operatorname{argmax}} \frac{P(W \mid D)P(E)}{P(E)}$$

$$= \underset{D}{\operatorname{argmax}} P(E \mid D)P(E)$$

HMMs for dialogue act interpretation

$$D^* = \underset{D}{\operatorname{argmax}}\, P(E \mid D)P(D)$$

- Let W be word sequence in sentence and F be prosodic feature sequence
- Simplifying (wrong) independence assumption
- (What are implications of this?) $P(E \mid D) = P(F \mid D)P(W \mid D)$

$$D^* = \underset{D}{\operatorname{argmax}}\, P(D)P(F \mid D)P(W \mid D)$$

HMM model for dialogue

$$D^* = \underset{D}{\operatorname{argmax}}\, P(D)P(F \mid D)P(W \mid D)$$

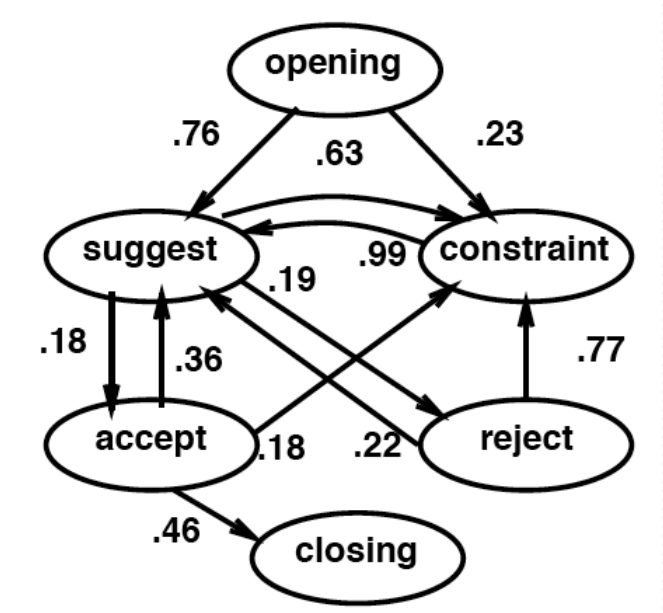
- Three components
  - P(D): probability of sequence of dialogue acts
  - P(F|D): probability of prosodic sequence given one dialogue act
  - P(W|D): probability of word string in a sentence given dialogue act

P(D)

- Markov assumption
- Each dialogue act depends only on previous N. (In practice, N of 3 is enough).

- Woszczyna and Waibel (1994):

$$P(D) = \prod_{i=2}^{M} P(d_i \mid d_{i-1}, \ldots, d_{i-M+1})$$



P(W|D)

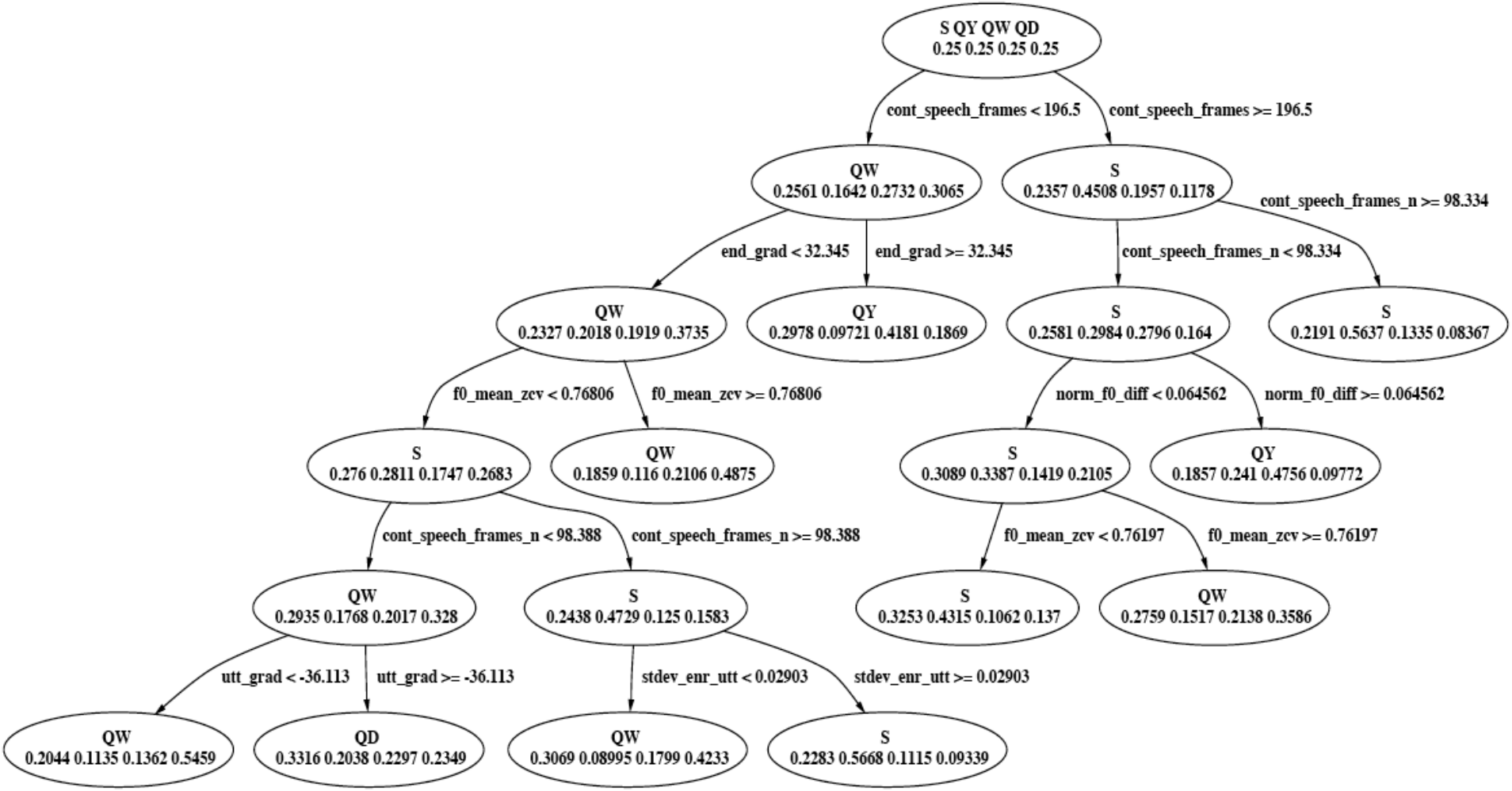
- Each dialogue act has different words
- Questions have “are you…”, “do you…”, etc

$$P(W \mid D) = \prod_{i=2}^{N} P(w_i \mid w_{i-1}, \ldots, w_{i-N+1}, d_i)$$

P(F|D)

- Shriberg et al. (1998)
- Decision tree trained on simple acoustically-based prosodic features
  - Slope of Fo at the end of the utterance
  - Average energy at different places in utterance
  - Various duration measures
  - All normalized in various ways
- These helped distinguish
  - Statement (S)
  - Yes-No-Question (QY)
  - Declarative-Question (QD)
  - Wh-Question (QW)

Prosodic Decision Tree for making S/QY/QW/QD decision



Getting likelihoods from decision tree

- Decision trees give posterior p(d|F) [discriminative, good]
- But we need p(F|d) to fit into HMM

$$p(d \mid F) = \frac{P(F \mid d)p(d)}{p(F)}$$

- Rearranging terms to get a likelihood:

$$\frac{p(F \mid d)}{p(F)} = \frac{P(d \mid F)}{p(d)}$$

- **scaled likelihood** is ok since p(F) is constant

Final HMM equation for dialogue act tagging

$$D^* = \arg\max_D P(D)P(F \mid D)P(W \mid D)$$

$$= \prod_{i=2}^{M} P(d_i \mid d_{i-1} \ldots d_{i-M+1}) \prod_{i=1}^{N} \frac{P(d_i \mid F)}{P(d_i)} \prod_{i=2}^{N} P(w_i \mid w_{i-1} \ldots w_{i-N+1}, d_i)$$

- Then can use Viterbi decoding to find D*
- In real dialogue systems, obviously can't use FUTURE dialogue acts, so predict up to current act
- In rescoring passes (for example for labeling human-human dialogues for meeting summarization), can use future info.
- Most other supervised ML classifiers have been applied to DA tagging task

An example of dialogue act detection: Correction Detection

- Despite all these clever confirmation/rejection strategies, dialogue systems still make mistakes (Surprise!)
- If system misrecognizes an utterance, and either
  - Rejects
  - Via confirmation, displays its misunderstanding
- Then user has a chance to make a correction
  - Repeat themselves
  - Rephrasing
  - Saying “no” to the confirmation question.

Corrections

- Unfortunately, corrections are harder to recognize than normal sentences!
  - Swerts et al (2000): corrections misrecognized twice as often (in terms of WER) as non-corrections!!!
  - Why?
    - Prosody seems to be largest factor: hyperarticulation
    - English Example from Liz Shriberg
      - "NO, I am DE-PAR-TING from Jacksonville)

A Labeled dialogue (Swerts et al)

	Turn	Corr Type
S	On which day of the week do you want to leave?	
S	Sorry, I didn't hear anything from you. Please tell me your departure date	
U	*Um tomorrow*	NA
S	You can say the departure date. For example, you can say 'today' or 'tomorrow'. Or you can say 'Monday', 'Tuesday', 'Wednesday'...	
U	*Tomorrow*	REP
S	You can say the departure date. For example, you can say 'today' or 'tomorrow'. Or you can say 'Monday', 'Tuesday', 'Wednesday'...	
U	*Saturday*	PAR
S	Do you want to leave on Saturday?	
U	Yes	
S	What time do you want to leave?	
U	Eight o'clock A M	
S	Do you want to leave around ten oclock p-m	
U	*Eight o'clock*	OMIT
S	Do you want to leave around ten oclock p-m	
U	*No*	OMIT
S	What time do you want to leave?	
U	*Eight o'clock A M*	ADD

Machine Learning and Classifiers

- Given a labeled training set
- We can build a classifier to label observations into classes
  - Decision Tree
  - Regression
  - SVM
- I won't introduce the algorithms here.
- But these are at the core of NLP/computational linguistics/Speech/Dialogue
- You can learn them in:
  - AI
  - Machine Learning

Machine learning to detect user corrections

- Build classifiers using features like
  - Lexical information (words “no”, “correction”, “I don’t”, swear words)
  - Prosodic features (various increases in Fo range, pause duration, and word duration that correlation with hyperarticulation)
  - Length
  - ASR confidence
  - LM probability
  - Various dialogue features (repetition)

Disambiguating Ambiguous DAs Intonationally

- Nickerson & Chu-Carroll '99: Can info-requests be disambiguated reliably from action-requests?
- Modal (Can/would/would..willing) questions
  - Can you move the piano?
  - Would you move the piano?
  - Would you be willing to move the piano?

Experiments

- Production studies:
  - Subjects read ambiguous questions in disambiguating contexts
  - Control for given/new and contrastiveness
  - Polite/neutral/impolite
- Problems:
  - Cells imbalanced
  - No pretesting
  - No distractors
  - Same speaker reads both contexts

Results

- Indirect requests (e.g. for action)
  - If L%, more likely (73%) to be indirect
  - If H%,46% were indirect: differences in height of boundary tone?
  - Politeness: can differs in impolite (higher rise) vs. neutral
  - Speaker variability

Corpus Studies: Jurafsky et al '98

- Lexical, acoustic/prosodic/syntactic differentiators for yeah, ok, uhuh, mhmm, um…
- Labeling
  - Continuers: Mhmm (not taking floor)
  - Assessments: Mhmm (tasty)
  - Agreements: Mhmm (I agree)
  - Yes answers: Mhmm (That's right)
  - Incipient speakership: Mhmm (taking floor)

Corpus

- Switchboard telephone conversation corpus
  - Hand segmented and labeled with DA information (initially from text)
  - Relabeled for this study
  - Analyzed for
    - Lexical realization
    - Fo and rms features
    - Syntactic patterns

Results: Lexical Differences

- Agreements
  - yeah (36%), right (11%),…
- Continuer
  - uhuh (45%), yeah (27%),…
- Incipient speaker
  - yeah (59%), uhuh (17%), right (7%),…
- Yes-answer
  - yeah (56%), yes (17%), uhuh (14%),…

Results: Prosodic and Syntactic Cues

- Relabeling from speech produces only 2% changed labels over all (114/5757)
  - 43/987 continuers --> agreements
  - Why?
    - Shorter duration, lower Fo, lower energy, longer preceding pause
- Over all DA's, duration best differentiator but…
  - Highly correlated with DA length in words
- Assessments: That's X (good, great, fine,…)

More Automatic DA Detection

- Rosset & Lamel '04: Can we detect DAs automatically w/ minimal reliance on lexical content?
  - Lexicons are domain-dependent
  - ASR output is errorful
- Corpora (3912 utts total)
  - Agent/client dialogues in a French bank call center, in a French web-based stock exchange customer service center, in an English bank call center

- DA tags new again (44)
  - Conventional (openings, closings)
  - Information level (items related to the semantic content of the task)
  - Forward Looking Function:
    - statement (e.g. assert, commit, explanation)
    - infl on Hearer (e.g. confirmation, offer, request)
  - Backward Looking Function:
    - Agreement (e.g. accept, reject)
    - Understanding (e.g. backchannel, correction)
  - Communicative Status (e.g. self-talk, change-mind)
  - NB: each utt could receive a tag for each class, so utts represented as vectors
    - But…only 197 combinations observed

- Method: Memory-based learning (TIMBL)
  - Uses all examples for classification
  - Useful for sparse data
- Features
  - Speaker identity
  - First 2 words of each turn
  - # utts in turn
  - Previously proposed DA tags for utts in turn
- Results
  - With true utt boundaries:
    - ~83% accuracy on test data from same domain
    - ~75% accuracy on test data from different domain

- On automatically identified utt units: 3.3% ins, 6.6% del, 13.5% sub

- Which DAs are easiest/hardest to detect?

DA	GE.fr	CAP.fr	GE.eng
Resp-to	52.0%	33.0%	55.7%
Backch	75.0%	72.0%	89.2%
Accept	41.7%	26.0%	30.3%
Assert	66.0%	56.3%	50.5%
Expression	89.0%	69.3%	56.2%
Comm-mgt	86.8%	70.7%	59.2%
Task	85.4%	81.4%	78.8%

Practical Goals

- In Spoken Dialogue Systems
  - Disambiguate current DA
    - Represent user input correctly
    - Responding appropriately
  - Predict next DA
    - Switch Language Models for ASR
    - Switch states in semantic processing

Generating Dialogue Acts

- Confirmation
- Rejection

Confirmation

- Another reason for grounding
- Errors: Speech is a pretty errorful channel
  - Even for humans; so they use grounding to confirm that they heard correctly
- ASR is way worse than humans!
- So dialogue systems need to do even more grounding and confirmation than humans

Explicit confirmation

- S: Which city do you want to leave from?
- U: Baltimore
- S: Do you want to leave from Baltimore?
- U: Yes

Explicit confirmation

- U: *I'd like to fly from Denver Colorado to New York City on September 21st in the morning on United Airlines*
- S: Let's see then. I have you going from Denver Colorado to New York on September 21st. Is that correct?
- U: *Yes*

Implicit confirmation: display

- U: *I'd like to travel to Berlin*
- S: When do you want to travel to Berlin?

- U: *Hi I'd like to fly to Seattle Tuesday morning*
- S: Traveling to Seattle on Tuesday, August eleventh in the morning. Your name?

Implicit vs. Explicit

- Complementary strengths
- Explicit: easier for users to correct systems's mistakes (can just say "no")
- But explicit is cumbersome and long
- Implicit: much more natural, quicker, simpler (if system guesses right).

Implicit and Explicit

- Early systems: all-implicit or all-explicit
- Modern systems: adaptive
- How to decide?
  - ASR system can give confidence metric.
  - This expresses how convinced system is of its transcription of the speech
  - If high confidence, use implicit confirmation
  - If low confidence, use explicit confirmation

Computing confidence

- Simplest: use acoustic log-likelihood of user's utterance
- More features
  - Prosodic: utterances with longer pauses, Fo excursions, longer durations
  - Backoff: did we have to backoff in the LM?
  - Cost of an error: Explicit confirmation before moving money or booking flights

Rejection

- e.g., VoiceXML “nomatch”
- “I’m sorry, I didn’t understand that.”
- Reject when:
  - ASR confidence is low
  - Best interpretation is semantically ill-formed
- Might have four-tiered level of confidence:
  - Below confidence threshhold, reject
  - Above threshold, explicit confirmation
  - If even higher, implicit confirmation
  - Even higher, no confirmation