Spoken Dialog System Architecture

Joshua Gordon

CS4706

Outline

- Motivation: current trends in SDS / Conversational Speech Interface Architecture

- An end to end tour of the Olympus SDS Architecture
  - Recognition considerations
  - Spoken language understanding techniques
  - Dialog management, error handling, belief updating
  - Language generation / speech synthesis
  - Interaction management, turn taking

Information Seeking, Transaction Based Spoken Dialog Systems

Many of today's systems are designed for database access and call routing

- Columbia: CheckItOut – virtual librarian
- CMU: Let's Go! Pittsburg bus schedules
- Google: Goog411 – directory assistance, Google Voice Search
- MIT – Jupiter – weather information
- Nuance – built to order, technical support

Speech Aware Kiosks

SDS architectures are beginning to incorporate multimodal input

“How may I help you? I can provide directory assistance, and directions around campus.”

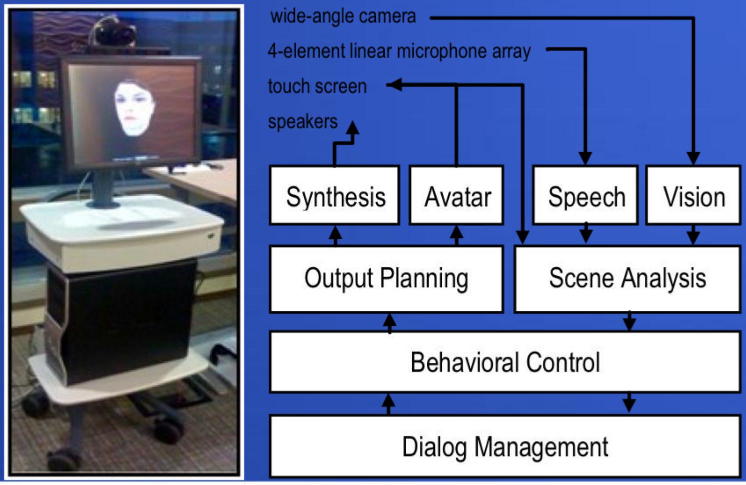


Figure 2. System prototype and architectural overview.

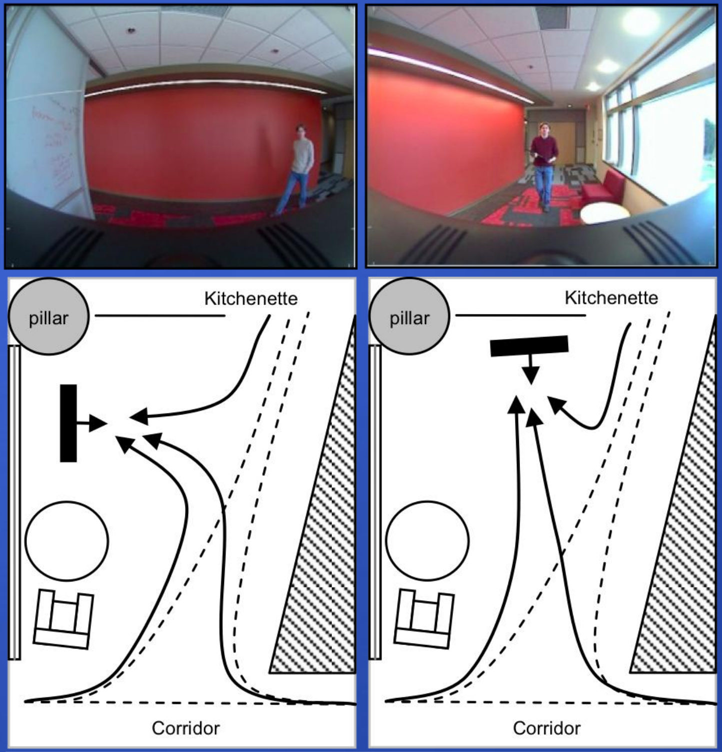


Figure 3. Placement and visual fields of view for *side* (right) and *front* (left) orientations.

Speech Interfaces to Virtual Characters

SDS architectures are exploring multimodal output (including gesturing and facial expression) to indicate level of understanding

- Negotiate an agreement between soldiers and village elders
- Both auditory and visual cues used in turn taking
- Prosody, facial expressions convey emotion

SGT Blackwell

http://ict.usc.edu/projects/sergeant_blackwell/

Speech Interfaces to Robotic Systems

Next generation systems explore ambitious domains

www.cellbots.com

User: Fly to the red house and photograph the area.
System: OK, I am preparing to take off.

Speech Aware Appliances

Speech aware appliances are beginning to engage in limited dialogs

- Interactive dialogs / disambiguation are required by multi-field queries, ambiguity in results

Expected	What user actually said
Play artist Glenn Miller	Glenn Miller, jazz
Play song all rise	All rise, I guess, from blues

How does all of this work?

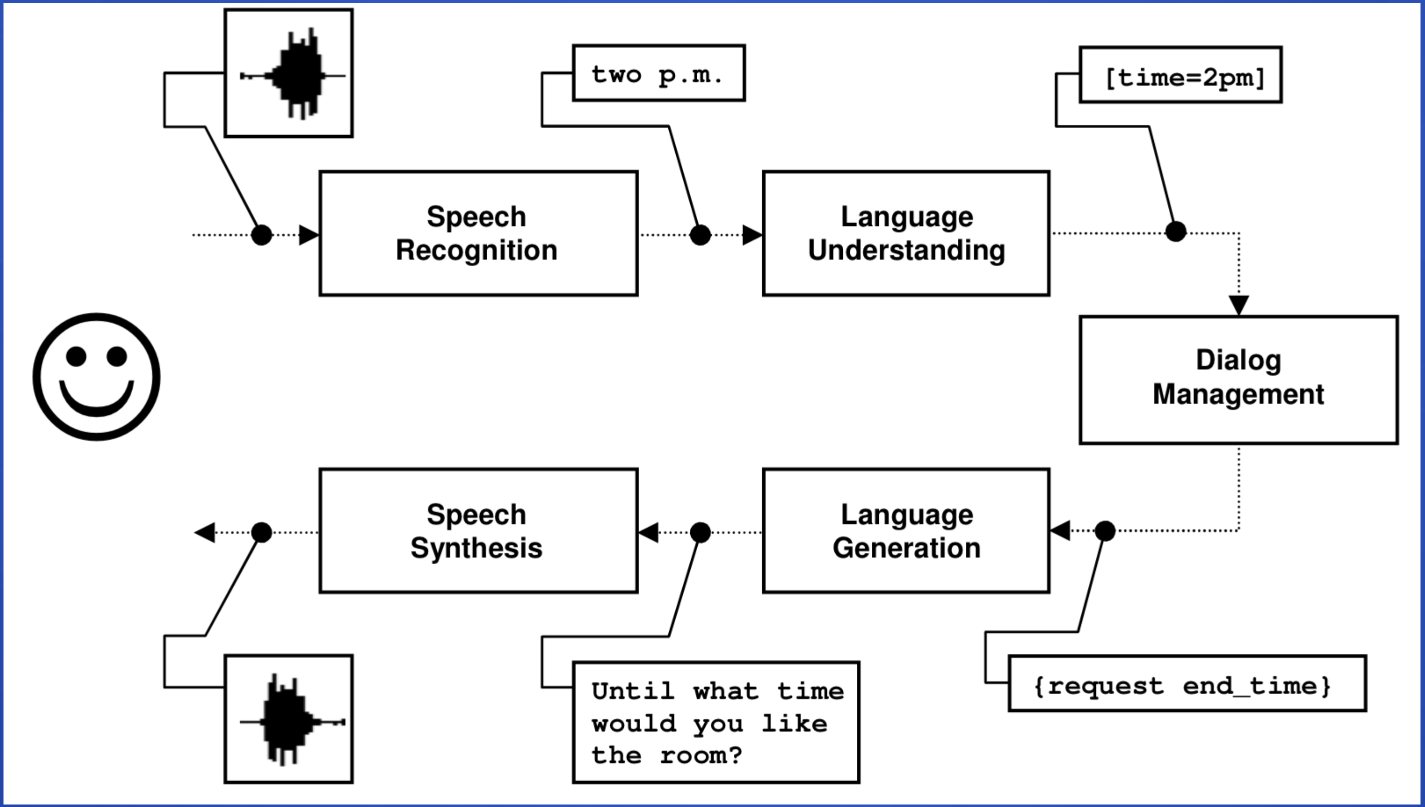
There’s more to conversation than we realize!

- An ocean of difference remains between Human-Human and Human-Machine Dialog
- Recognition performance often seen as the limiting factor – but fundamental challenges exist in all areas
- Turn taking via subtle auditory cues – ever listened to two speakers competing for the conversational floor?
- Grounding via prosody, intonation contours. Indicating level of understanding by answering a question with a question.
- Mapping speech to concepts requires knowledge of the world. SDS are subject to limited domain knowledge.
- Lack ability to effectively communicate their capabilities and limitations

Running example:
SDS Architecture for a Virtual Librarian

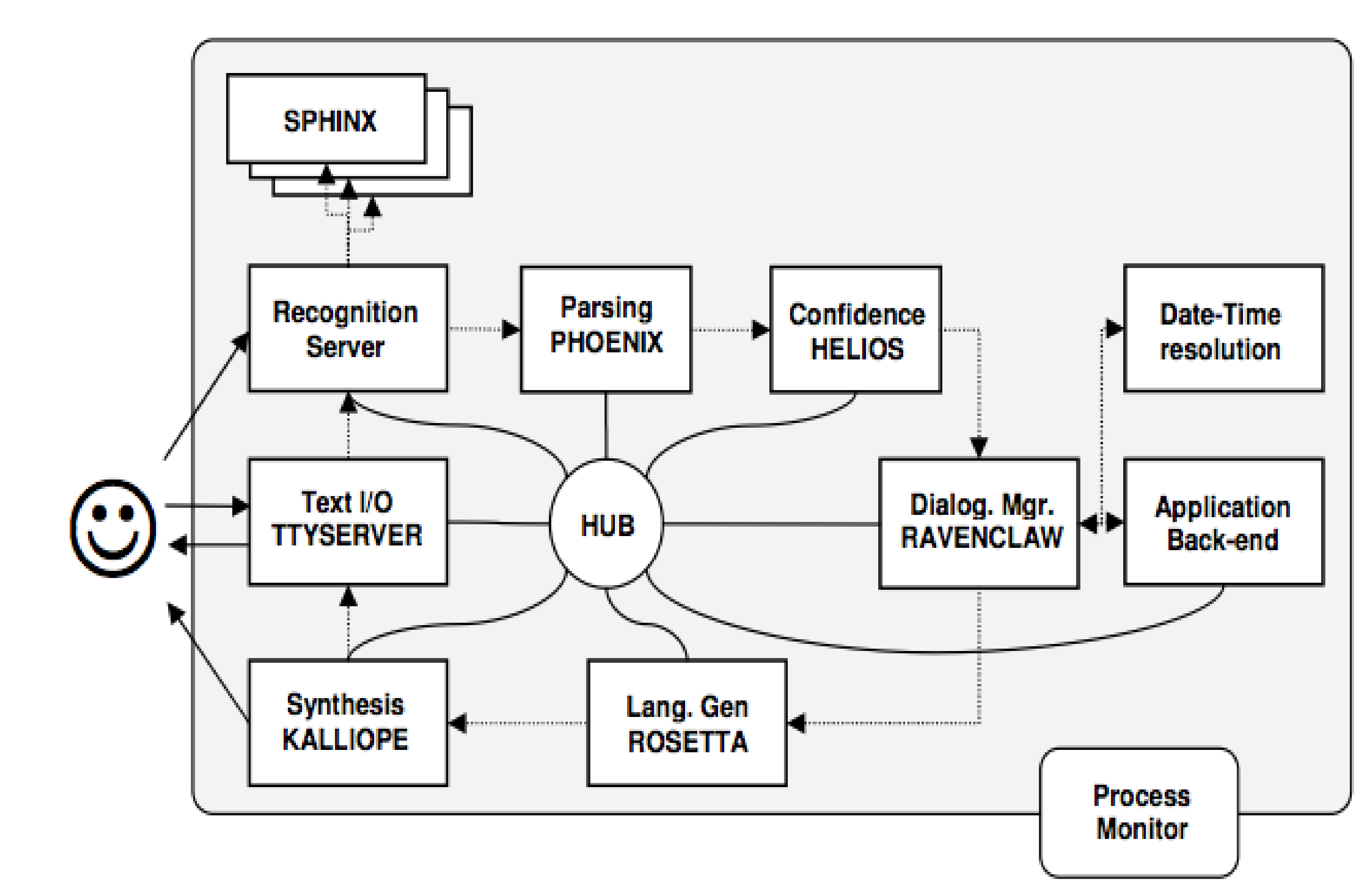
- The Andrew Heiskell Braille and Talking Book Library
  - Patrons will browse / order books by phone
  - Heiskell's bibliographic holdings include +/- 70,000 books
  - Challenge: many callers have relatively disfluent speech. Poor recognizer performance is anticipated.
- What are the components we'll need?
- Introducing the Olympus Architecture
  - a freely available, open source collection of dialog system components published by CMU
  - Origins in the earlier Communicator project

The Olympus Architecture



Pipeline format, subsequent layers increase abstraction. Signals to words, words to concepts, concepts to actions

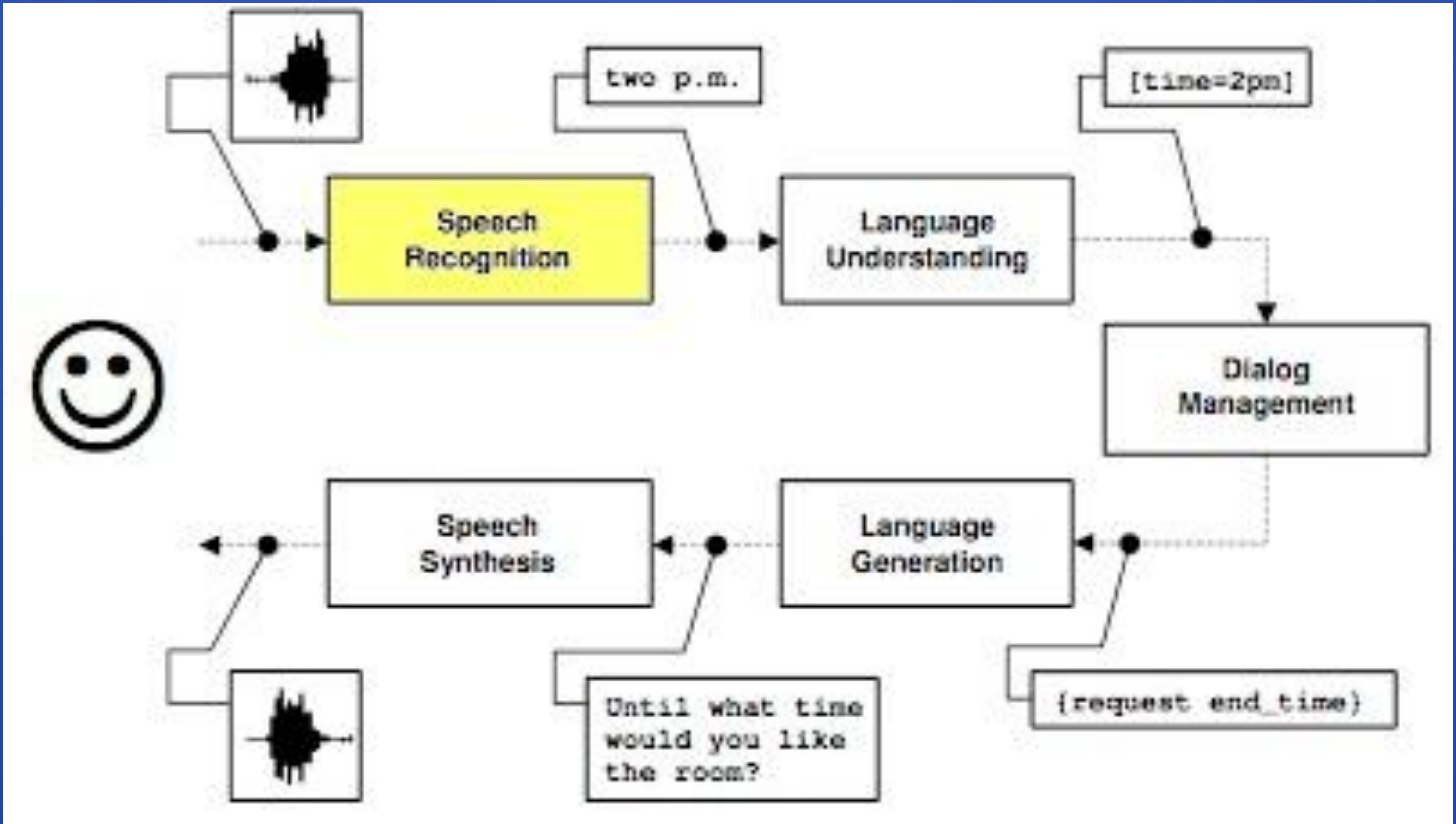
Detail: Hub Architecture



Deployed Olympus Systems

System	Domain	Users	Interaction	Vocab
Lets Go Public!	Pittsburg Bus Route Information	General public	Information access (system initiative), background noise	2000 words
Team Talk	Robot Coordination and Control – Treasure hunting	Grad students / researchers	Multi-participant command and control	500 words
CheckItOut	Virtual Librarian for the Andrew Heiskell Library	Elderly, vision impaired library patrons	Information access (mixed initiative), disfluent speech	Variable - +/- 10,000 words

Part 1: Speech recognition



From signals to words, managing uncertainty

- Information provided to downstream components
  - A lexical representation of the speech signal, with acoustic confidence and language model fit scores
  - An N-best list
  - But *How* you say it often conveys as much information as *what* is said.
  - Prosody, intonation, amplitude, duration
  - Moving from an acoustic signal to a lexical representation already implies loss of information!
- SDS architectures always operate on partial information
  - Managing that uncertainty is one of the main design challenges

Why ASR is Difficult for SDS

- A SDS must accommodate variability in…
- Calling environments: background noise, cell phone interference, VOIP
- Speech production: disfluency, false starts, filled pauses, repeats, corrections, accent, age, gender, differences between human-human and human-machine speech
- Technological familiarity: with dialog systems in general, with a particular SDS's capabilities and constraints, callers often use OOV / out of domain concepts

The Sphinx Open Source Recognition Toolkit

- Pocket-sphinx vs. Sphinx III
  - ps is efficient enough to run on embedded devices
- Continuous speech, speaker independent recognition system
- Includes tools for language model compilation, pronunciation, and acoustic model adaptation
- Provides word level confidence annotation, n-best lists
- Olympus supports parallel decoding engines / models
- Typically separate models are run for male and female speech, the best fit hypothesis is selected

http://cmusphinx.sourceforge.net/

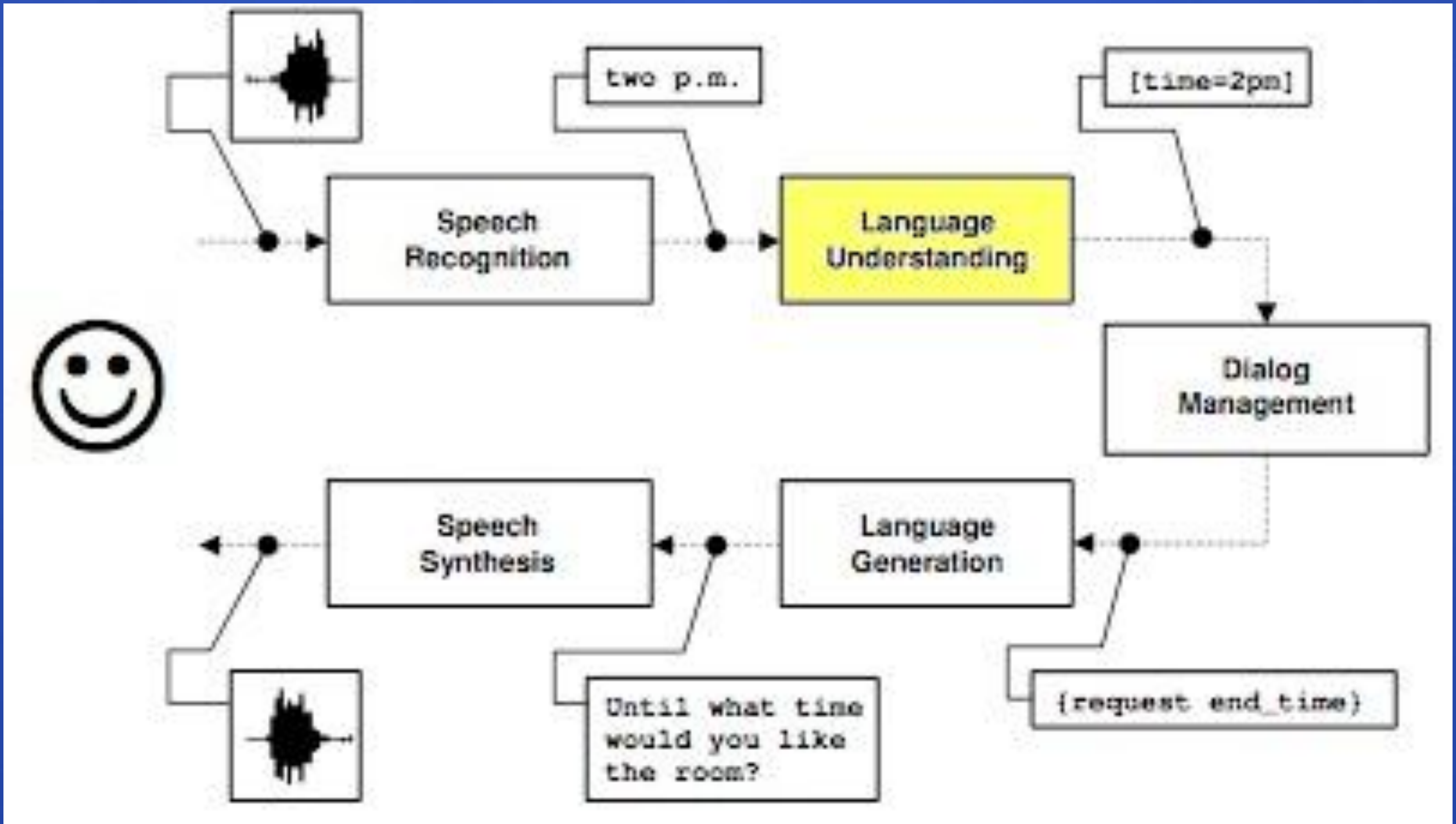
Language, Acoustic Models for SDS

- Sphinx supports statistical, class, and state based language models
- Statistical language models assign n-gram probabilities to word sequences
- Class based models assign probabilities to collections of terminals, e.g., “I would like to read <book>”
- State based LM switching: SDS limit the perplexity of the language model by constraining it to the anticipated words
- <confirmation / rejection>, <help>, <address>, …

Acoustic Models for SDS

- Olympus includes permissive-license WSJ Acoustic models (read speech) for male and female speech, at 8khz and 16hkz bandwidth
- Tools for acoustic adaptation
- Support permissive-license models!

Part 2: Spoken Language Understanding



From words to concepts

- Spoken Language Understanding is the task of extracting meaning from utterances
  - Dialog acts (the overall intent of an utterance)
  - Domain specific concepts: frame / slots
- Very difficult under noisy conditions
- “Does the library have *Hitchhikers Guide to the Galaxy* by *Douglas Adams* on *audio cassette?”*

Dialog Act	Book Request
Title	The Hitchhikers Guide to the Galaxy
Author	Douglas Adams
Media	Audio Cassette

SLU Challenges faced by SDS

- Recognizer error, background noise resulting in indels (insertions / substitutions / deletions), word boundary detection problems
- Language production phenomena: disfluency, false starts, corrections, repairs are difficult to parse
- Meaning must often be assembled from multiple speaker turns
- There are many, many possible ways to say the same thing
- How can SDS anticipate all of them?

Semantic grammars

- Frames, concepts, variables, terminals
- Domain independent concepts
  - [Yes], [No], [Help], [Repeat], [Number]
- Domain dependent concepts
  - [Title], [Author], [BookOnTape], [Braille]
- The pseudo corpus LM trick

```
[Quit]
    (*THANKS *good bye)
    (*THANKS goodbye)
    (*THANKS +bye)
;

THANKS
    (thanks *VERY_MUCH)
    (thank you *VERY_MUCH)

VERY_MUCH
    (very much)
    (a lot)
;
```

Semantic parsers

- Phoenix uses a semantic hand-written grammar to parse the incoming set of recognition hypotheses
- Goal: consume as many terminals as possible
- Phoenix maps input sequences of words to semantic frames
  - A frame is a named set of slots, where slots represent pieces of related information
  - Each slot has an associated CFG Grammar, specifying word patterns that match the slot
  - Chart parsing selects the path which accounts for the maximum number of terminals
  - Multiple parses may be produced for a single utterance

Examining a few parses

System: Am I speaking with Logan Paddock?

- User: OF COURE
- Parse: Generic [Confirmation] ([YES] (OF COURSE))

System: May I help you find a book?

- User: A PRETEXT FOR WAR
- Parse: BookRequest [Title] ([TitleName] ([DT_HEAD] ([DT] (A) [NN_HEAD] ([NN] (PRETEXT) [IN] (FOR) [NN] (WAR)))))

SLU Design considerations

- Are hand engineered grammars the way to go?
  - Requires expert linguistic knowledge to construct
  - Time consuming to develop and tune
  - Difficult to maintain over complex domains
  - Lacks robustness to OOV words and novel phrasing
  - Lacks robustness to recognizer error and disfluent speech
  - Noise tolerance is difficult to achieve
- SLU can be greatly simplified by constraining what the user can say (and how they can say it!)
  - But.. results in a less habitable, clunky conversation. Who wants to chat with a system like that?

Statistical methods (to the rescue?)

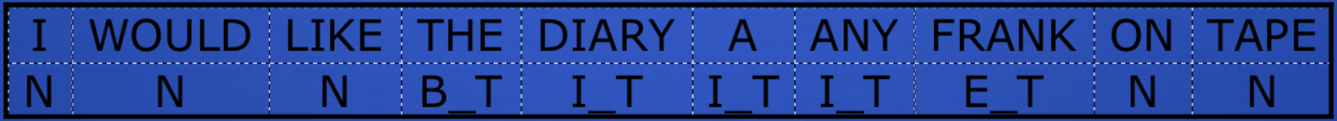
- Language understanding as pattern recognition
- Given word sequence *W*, find the semantic representation of meaning M that has maximum a posteriori probability P(M|W)
- P(M): prior meaning probability, based on dialogue state
- P(W|M): assigns probability to word sequence W given the semantic structure

[illegible]

Supervised Methods

- Dialog act classification
  - Request for book by author, by title, by ISBN
  - Useful for grounding, error handling, maintaining the situational frame

- Named entity recognition via statistical tagging – as a preprocessor for voice search



I	WOULD	LIKE	THE	DIARY	A	ANY	FRANK	ON	TAPE
N	N	N	B_T	I_T	I_T	I_T	E_T	N	N

Voice search

Database search with noisy ASR queries

- Phonetic, partial matching database queries
- Frequently used in information retrieval domains where Spoken Dialog Systems must access a database
- Challenges
  - Multiple database fields
  - Confusability of concepts

"The Language of Issa Come Wars"

Return	Confidence
The language of sycamores	.8
the language of clothes	.65
the language of threads	.51
The language of love	.40

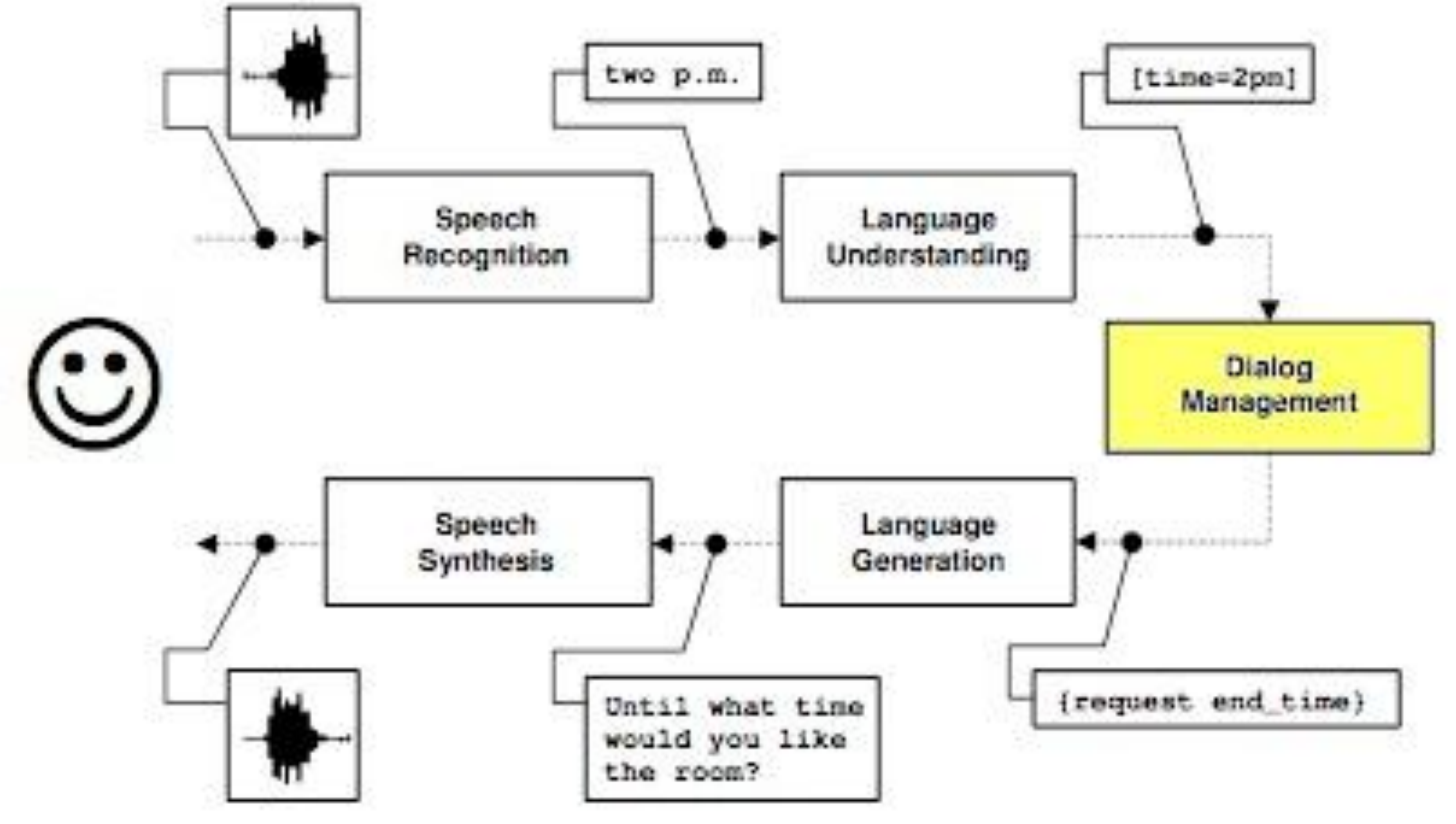
Relative merits: Statistical vs. Knowledge based SLU

- Statistical methods
  - Provide more robust coverage, especially for naïve users who respond frequently with OOV (out of vocabulary) words
  - Require labeled training data (some efforts to produce via simulation studies)
  - Better for shallow understanding
  - Excellent for call routing, question answering (assuming the question is drawn from a predefined set!)
- Semantic parsers
  - Provide a richer representation of meaning
  - Require substantially more effort to develop
  - Assist in the develop of state based language models

In Practice techniques are combined

- Institute for Creative Technologies: Virtual Humans
  - Question answering: maps user utterances to a small set of predefined answers
  - Robust to high word error rate (WER) up to 50%
- The AT&T Spoken Language Understanding System
  - Couples statistical methods for call-routing with semantic grammars for named-entity extraction

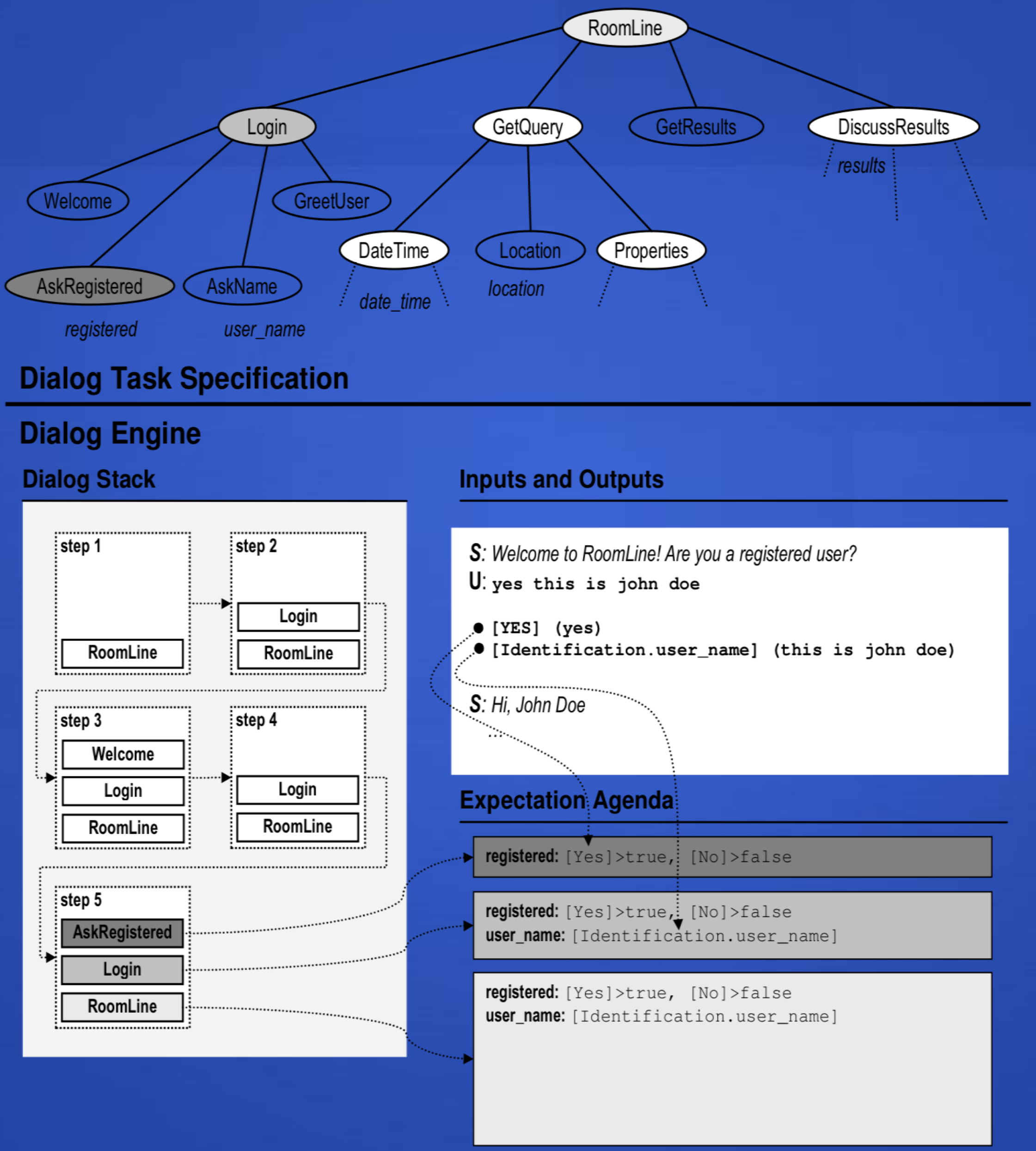
Part 3: Dialogue Management



From concepts to actions

- Raven Claw: a two tier dialog management architecture which decouples the domain specific aspects of dialog control from belief updating and error handling
- Represents dialog as set of hierarchal plans
- Domain independent error handling strategies
- The idea is to generalize dialog management framework across domains
- Architectural tradeoffs between system and mixed initiative dialog management
- A system initiative design has no uncertainty about the dialog state… but is inherently clunky and rigid

Dialogue Task Specification, Agenda, and Execution



S: *Welcome to RoomLine! Are you a registered user?*

U: `yes this is john doe`

- `[YES] (yes)`
- `[Identification.user_name] (this is john doe)`

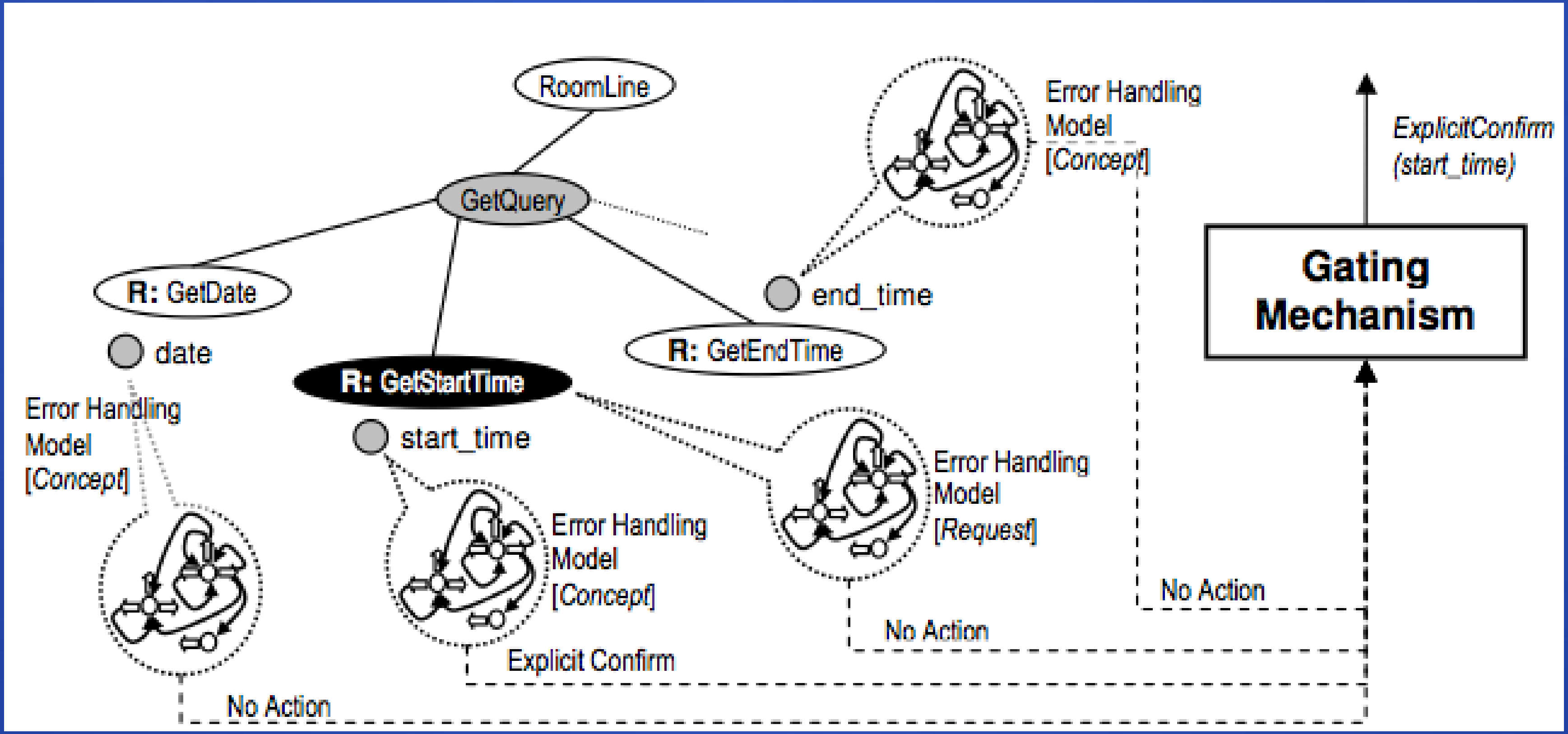
S: *Hi, John Doe*

registered: `[Yes]>true, [No]>false`

registered: `[Yes]>true, [No]>false`
user_name: `[Identification.user_name]`

registered: `[Yes]>true, [No]>false`
user_name: `[Identification.user_name]`

Distributed error handling

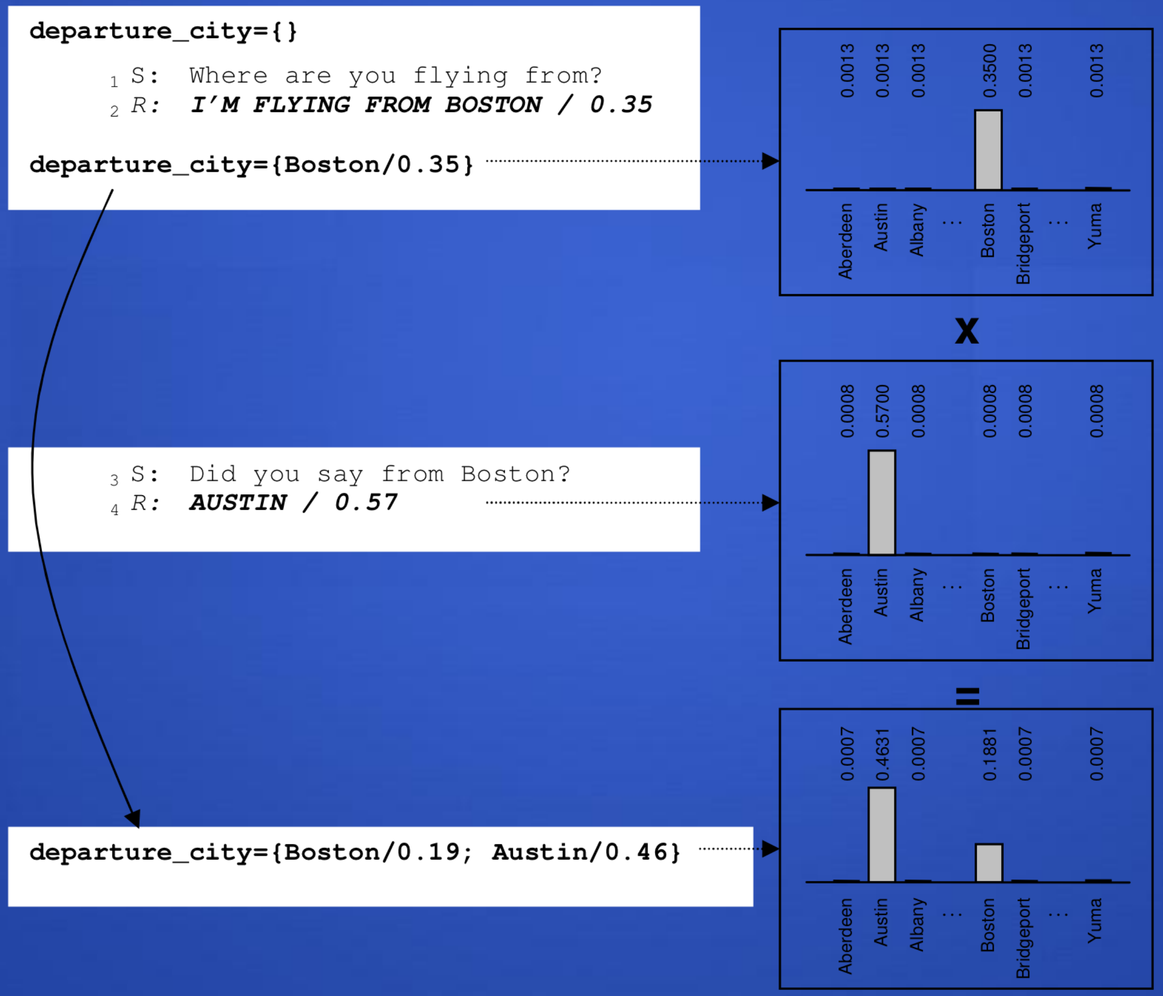


RoomLine
GetQuery
R: GetDate
date
R: GetStartTime
start_time
R: GetEndTime
end_time
Error Handling Model [Concept]
Error Handling Model [Concept]
Error Handling Model [Request]
Error Handling Model [Concept]
Gating Mechanism
ExplicitConfirm (start_time)
No Action
Explicit Confirm
No Action
No Action

How should the DM estimate certainty in a concept?

- How are initial confidences assigned to concepts?
- Helios (a confidence annotator) uses a logistic regression model to score Phoenix parses
- This score reflects the probability of correct understanding, i.e. how much the system trusts that the current semantic interpretation corresponds to the user's expressed intent
- Features from different knowledge sources
- Acoustic confidence, language model score, parse coverage, dialog state, …

Belief updating



```
departure_city={}

      1 S:  Where are you flying from?
      2 R:  I'M FLYING FROM BOSTON / 0.35

departure_city={Boston/0.35}
```

```
      3 S:  Did you say from Boston?
      4 R:  AUSTIN / 0.57
```

```
departure_city={Boston/0.19; Austin/0.46}
```

Error recovery strategies

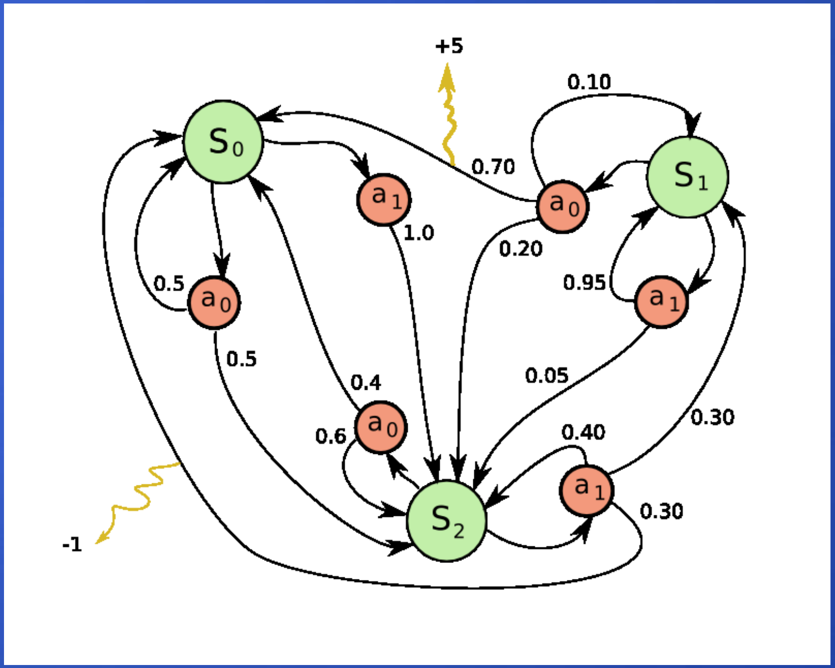
Error Handling Strategy (misunderstanding)	Example
Explicit confirmation	Did you say you wanted a room starting at 10 a.m.?
Implicit confirmation	Starting at 10 a.m. ... until what time?

Error Handling Strategy (non-understanding)	Example
Notify that a non-understanding occurred	Sorry, I didn't catch that .
Ask user to repeat	Can you please repeat that?
Ask user to rephrase	Can you please rephrase that?
Repeat prompt	Would you like a small room or a large one?

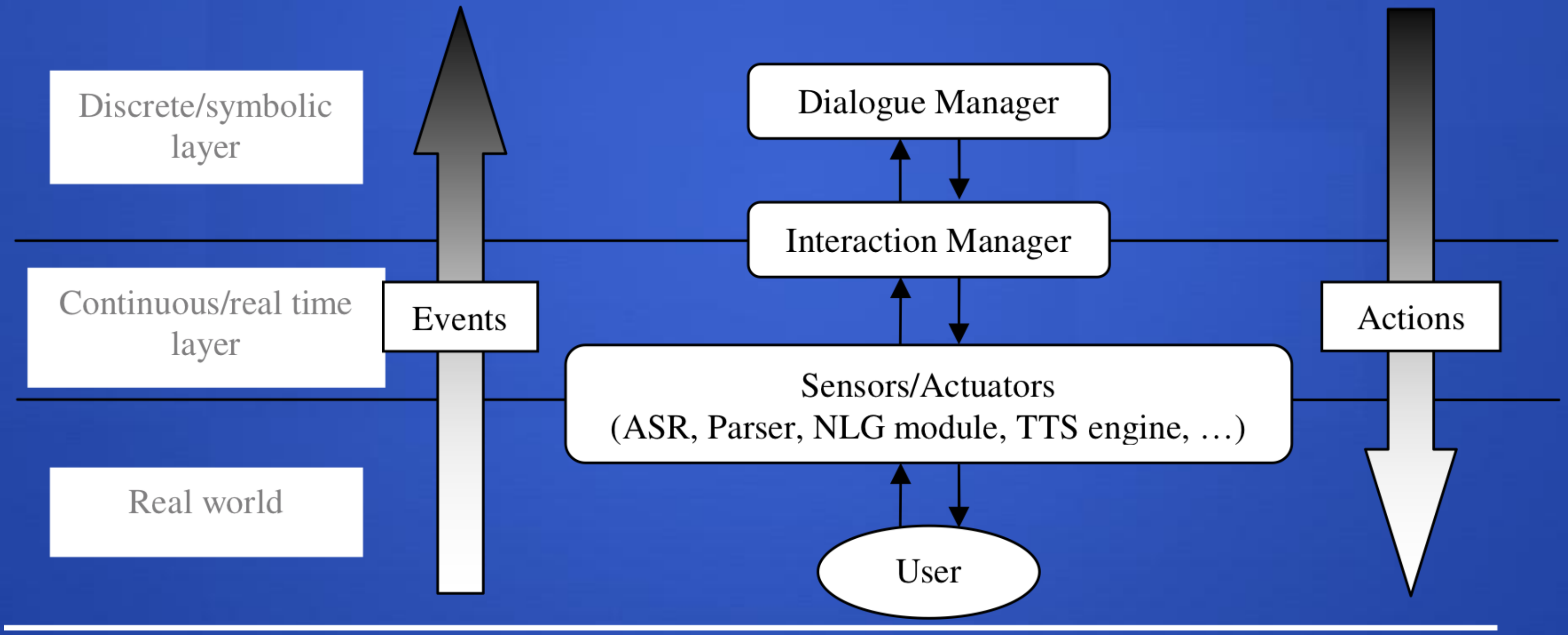
Goal is to avoid non-understanding cascades – the farther the dialog gets off track, the more difficult it is to recover

Statistical Approaches to Dialogue Management

- Is it possible to learn a management policy from a corpus?
- Dialogue may be modeled as Partially Observable Markov Decision Processes
- Reinforcement learning is applied (either to existing corpora or through user simulation studies) to learn an optimal strategy
- Evaluation functions typically reference the PARADISE framework – taking into account objective and subjective criteria



Part 4: Interaction management



Discrete/symbolic layer

Continuous/real time layer

Real world

Events

Dialogue Manager

Interaction Manager

Sensors/Actuators
(ASR, Parser, NLG module, TTS engine, …)

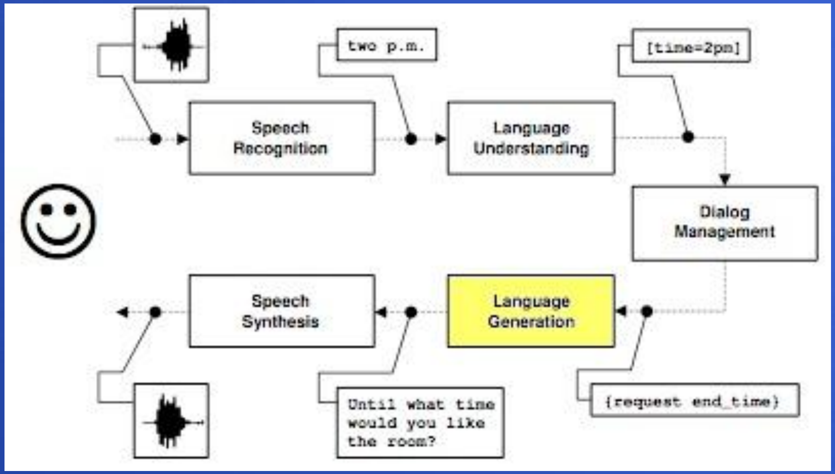
User

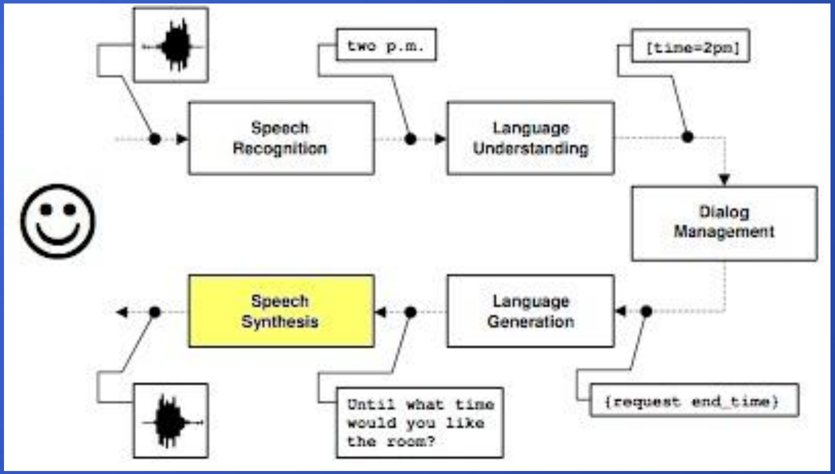
Actions

Turn taking

- Mediates between the discrete, symbolic reasoning of the dialog manager, and the continuous real-time nature of user interaction
- Manages timing, turn-taking, and barge-in
  - Yields the turn to the user should they interrupt
  - Prevents the system from speaking over the user
- Notifies the dialog manager of interrupts and incomplete utterances

Part 5: Natural Language Generation and Speech Synthesis



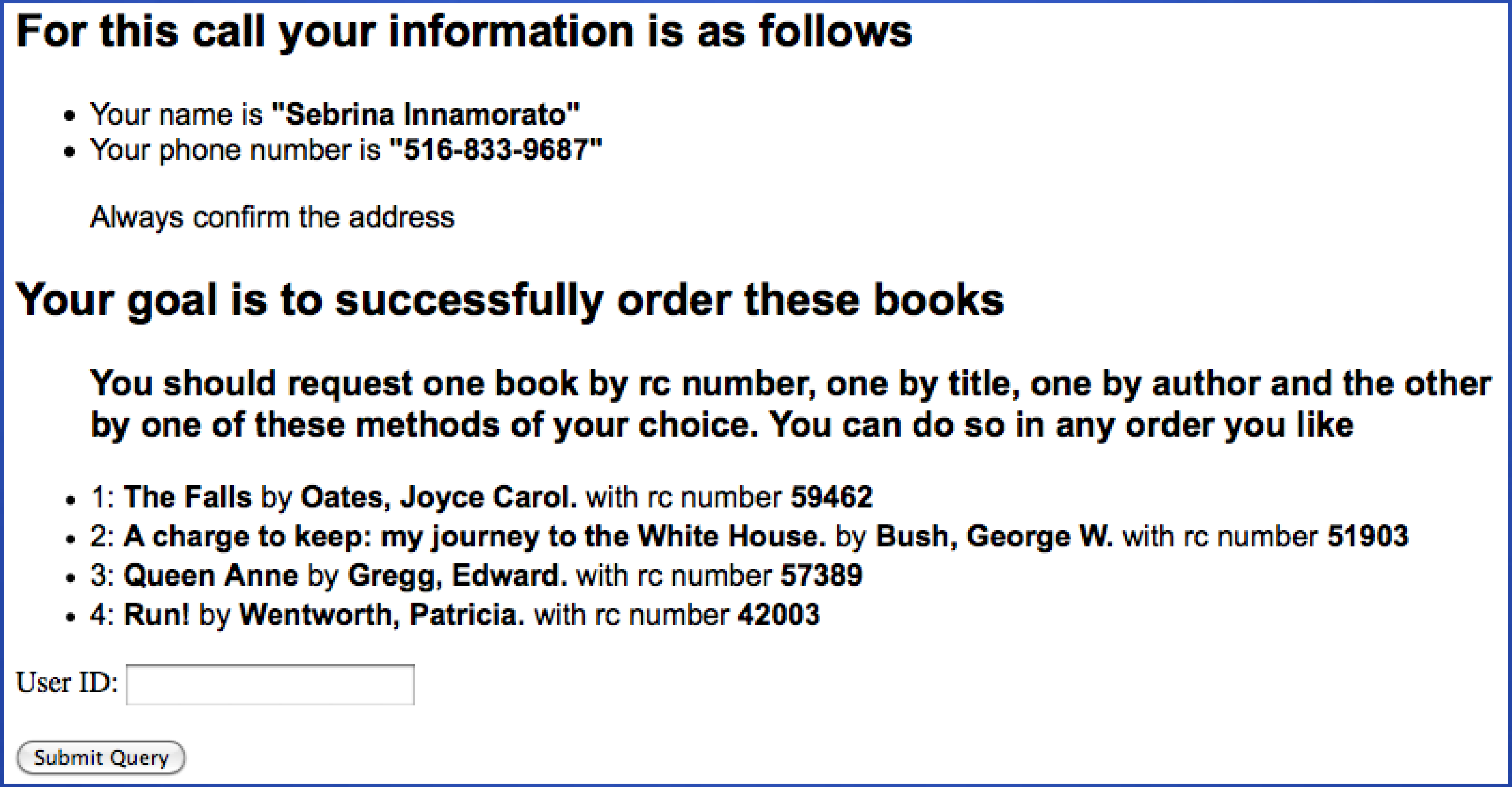


NLG and Speech Synthesis

- Template based, e.g., for explicit error handling strategies
  - Did you say <concept>?
- A TTS synthesizes the NLG output
  - The audio server allows interruption mid utterance
- Production systems incorporate
  - Prosody, intonation contours to indicate degree of certainty
- Open source TTS frameworks
  - Festival - http://www.cstr.ed.ac.uk/projects/festival/
  - Flite - http://www.speech.cs.cmu.edu/flite/

Putting it all together

CheckItOut Scenarios



For this call your information is as follows

- Your name is **"Sebrina Innamorato"**
- Your phone number is **"516-833-9687"**

Always confirm the address

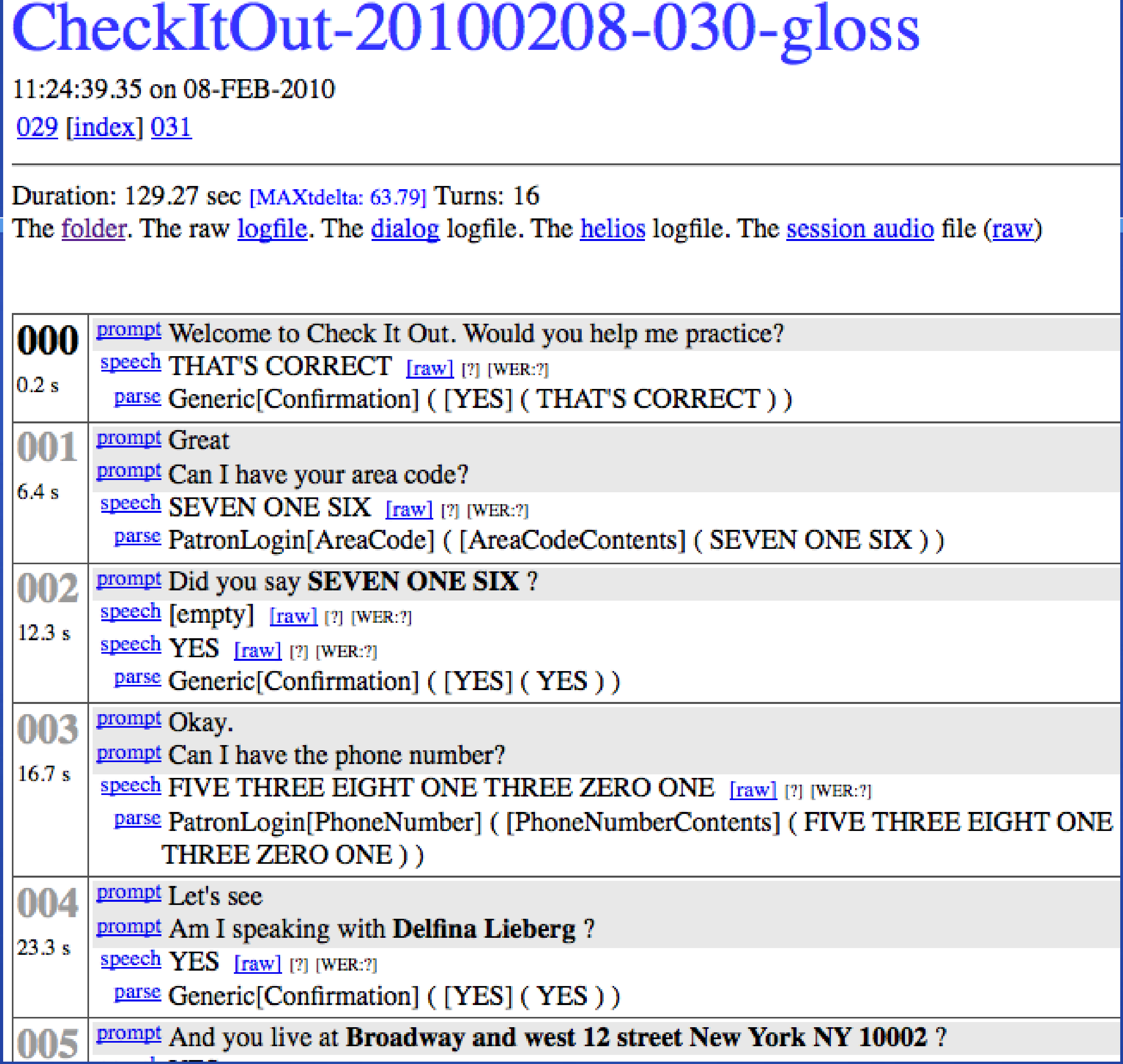
Your goal is to successfully order these books

You should request one book by rc number, one by title, one by author and the other by one of these methods of your choice. You can do so in any order you like

- 1: **The Falls** by **Oates, Joyce Carol.** with rc number **59462**
- 2: **A charge to keep: my journey to the White House.** by **Bush, George W.** with rc number **51903**
- 3: **Queen Anne** by **Gregg, Edward.** with rc number **57389**
- 4: **Run!** by **Wentworth, Patricia.** with rc number **42003**

User ID:

Evaluating the dialog



CheckItOut-20100208-030-gloss

11:24:39.35 on 08-FEB-2010

029 [index] 031

Duration: 129.27 sec [MAXtdelta: 63.79] Turns: 16

The folder. The raw logfile. The dialog logfile. The helios logfile. The session audio file (raw)

000
0.2 s
prompt Welcome to Check It Out. Would you help me practice?
speech THAT'S CORRECT [raw] [?] [WER:?]
parse Generic[Confirmation] ([YES] (THAT'S CORRECT))

001
6.4 s
prompt Great
prompt Can I have your area code?
speech SEVEN ONE SIX [raw] [?] [WER:?]
parse PatronLogin[AreaCode] ([AreaCodeContents] (SEVEN ONE SIX))

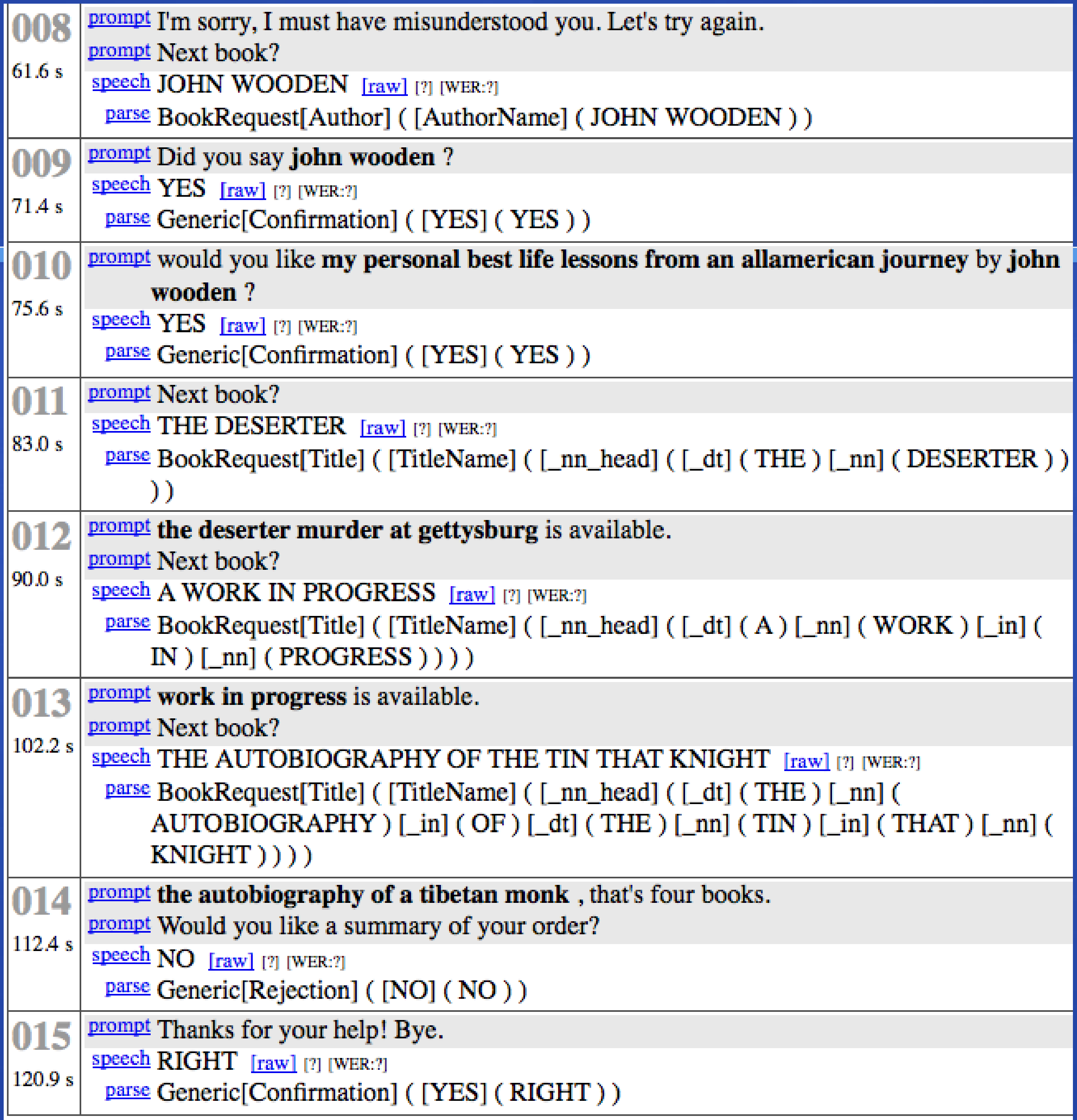
002
12.3 s
prompt Did you say **SEVEN ONE SIX** ?
speech [empty] [raw] [?] [WER:?]
speech YES [raw] [?] [WER:?]
parse Generic[Confirmation] ([YES] (YES))

003
16.7 s
prompt Okay.
prompt Can I have the phone number?
speech FIVE THREE EIGHT ONE THREE ZERO ONE [raw] [?] [WER:?]
parse PatronLogin[PhoneNumber] ([PhoneNumberContents] (FIVE THREE EIGHT ONE THREE ZERO ONE))

004
23.3 s
prompt Let's see
prompt Am I speaking with **Delfina Lieberg** ?
speech YES [raw] [?] [WER:?]
parse Generic[Confirmation] ([YES] (YES))

005
prompt And you live at **Broadway and west 12 street New York NY 10002** ?

Evaluating the dialog



008 61.6 s	prompt I'm sorry, I must have misunderstood you. Let's try again. prompt Next book? speech JOHN WOODEN [raw] [?] [WER:?] parse BookRequest[Author] ([AuthorName] (JOHN WOODEN))
009 71.4 s	prompt Did you say **john wooden** ? speech YES [raw] [?] [WER:?] parse Generic[Confirmation] ([YES] (YES))
010 75.6 s	prompt would you like **my personal best life lessons from an allamerican journey** by **john wooden** ? speech YES [raw] [?] [WER:?] parse Generic[Confirmation] ([YES] (YES))
011 83.0 s	prompt Next book? speech THE DESERTER [raw] [?] [WER:?] parse BookRequest[Title] ([TitleName] ([_nn_head] ([_dt] (THE) [_nn] (DESERTER))))
012 90.0 s	prompt **the deserter murder at gettysburg** is available. prompt Next book? speech A WORK IN PROGRESS [raw] [?] [WER:?] parse BookRequest[Title] ([TitleName] ([_nn_head] ([_dt] (A) [_nn] (WORK) [_in] (IN) [_nn] (PROGRESS))))
013 102.2 s	prompt **work in progress** is available. prompt Next book? speech THE AUTOBIOGRAPHY OF THE TIN THAT KNIGHT [raw] [?] [WER:?] parse BookRequest[Title] ([TitleName] ([_nn_head] ([_dt] (THE) [_nn] (AUTOBIOGRAPHY) [_in] (OF) [_dt] (THE) [_nn] (TIN) [_in] (THAT) [_nn] (KNIGHT))))
014 112.4 s	prompt **the autobiography of a tibetan monk** , that's four books. prompt Would you like a summary of your order? speech NO [raw] [?] [WER:?] parse Generic[Rejection] ([NO] (NO))
015 120.9 s	prompt Thanks for your help! Bye. speech RIGHT [raw] [?] [WER:?] parse Generic[Confirmation] ([YES] (RIGHT))

Future challenges

- Multi-participant conversations
  - How does each system identify who has the conversation floor and who is the addressee for any spoken utterance?
  - How can multiple agents solve the channel contention problem, i.e. multiple agents speaking over each other?
- Understand how objects, locations, and tasks come to be described in language.
  - Robots and humans will need to mutually ground their perceptions to effectively communicate about tasks.

References

- Alex Rudnicky et al. (1999) Creating natural dialogs in the Carnegie Mellon Communicator system. Eurospeech.
- Gupta, N. et al. (2006). The AT&T spoken language understanding system. IEEE Transactions on Audio, Speech, and Language Processing.
- Dan Bohus. (2007). Error awareness and recovery in conversational spoken language interfaces, PhD Thesis, Carnegie Mellon University,
- Dan Bohus and Eric Horvitz. (2009) Learning to Predict Engagement with a Spoken Dialog System in Open-World Settings. SIGDIAL.

Thanks! Questions?