MIT COMPUTER SCIENCE AND ARTIFICIAL INTELLIGENCE LABORATORY

Partially Observable Markov Decision Processes (POMDPs) for Spoken Dialog Systems

William Li

Presentation Goals

-
- POMDP theory and application in dialog systems
  - Mini-tutorial
- System implementation and experiments
  - Formulation of POMDP
  - Confidence scoring with boosting
- Results and further work

POMDP Theory and Implementation

Why POMDPs for Dialog?

CSAIL

- Handle uncertainty about the **user's intention (due to noisy data, ambiguity, or inherent uncertainty in the world)** in a principled, “data-driven” way

- Incorporate **information-gathering actions** about the user's intent (clarification/confirmation questions)

- Guarantee **optimal behavior** based on a reward function that encodes:
  - Exploration/exploitation trade-offs (asking questions vs. taking an action)
  - Asking the “right” clarification/confirmation question
  - Other dialog desiderata (e.g. shorter dialogs, preferred dialog paths)

- Other: Good off-the-shelf solvers for large POMDPs, some work done on actively learning/adapting POMDP parameters (Doshi-Velez 2009)

What is a POMDP?

- **Partially observable:** state is hidden, as opposed to a fully observable Markov decision process (MDP)
  - POMDPs explicitly model the **user's intent as a latent variable** (state-based model)
- **Markov:** transition functions depend only on entities (states, system actions, and observations) in time *t-1*
- **Decision process:** The system infers the state to choose **actions**

What is a POMDP?

- Hidden Markov Model (HMM) + Markov Decision Process (MDP)

		Are there system actions?	
		NO	YES
Are states known (fully observable)?	YES	Markov Chain	Markov Decision Process (MDP)
	NO	Hidden Markov Model (HMM)	Partially Observable Markov Decision Process (POMDP)

Spoken Dialog Management

Intuition: Use dialog to help determine the user's intent

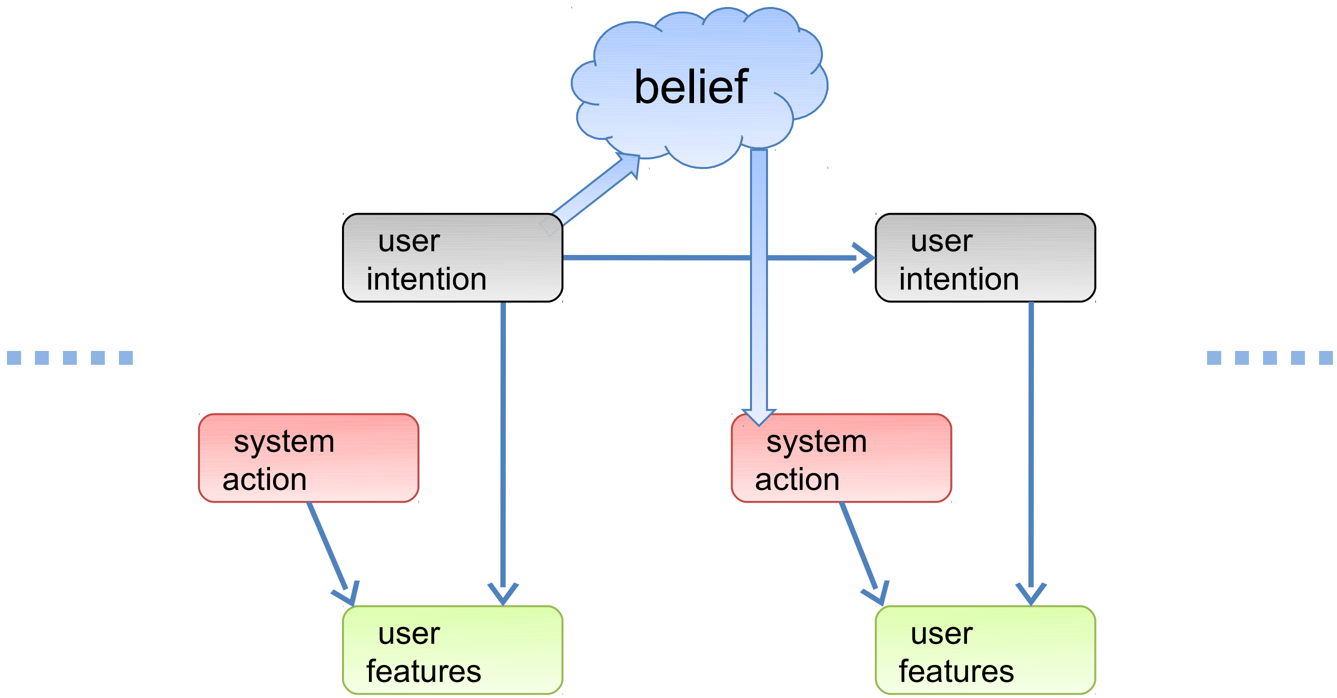
- Information-gathering system actions (clarifying/confirming questions) in addition to "terminal" actions
- User has a **state (goal/intent)** that is not directly observable
- Spoken dialog system (SDS) receives noisy **sensor observations (speech recognition results)**
- SDS decides, based on observation and dialog history, what **action (response)** to take

POMDPs in Other Domains

- Any agent that "has" a "state", receives noisy "observations" about the state, and takes actions with some goal could be modeled as a POMDP

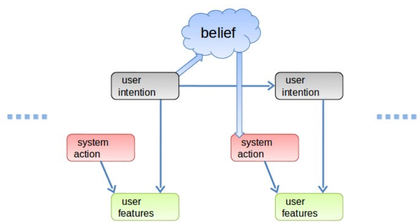
	States, S	Observations, Z	Actions, A
Spoken Dialog System	User's goal	ASR outputs (word hypotheses, confidence scores)	Dialog response
Robot planning/navigation	Position	Sensor readings (camera, wheel encoders)	Movements (stop, forward, left, right)
Intelligent handwashing prompt system (Mihailidis et al)	State of handwashing task	Video inputs	Prompts/reminders to user
Search engine???	User's goal	User text input	Search results

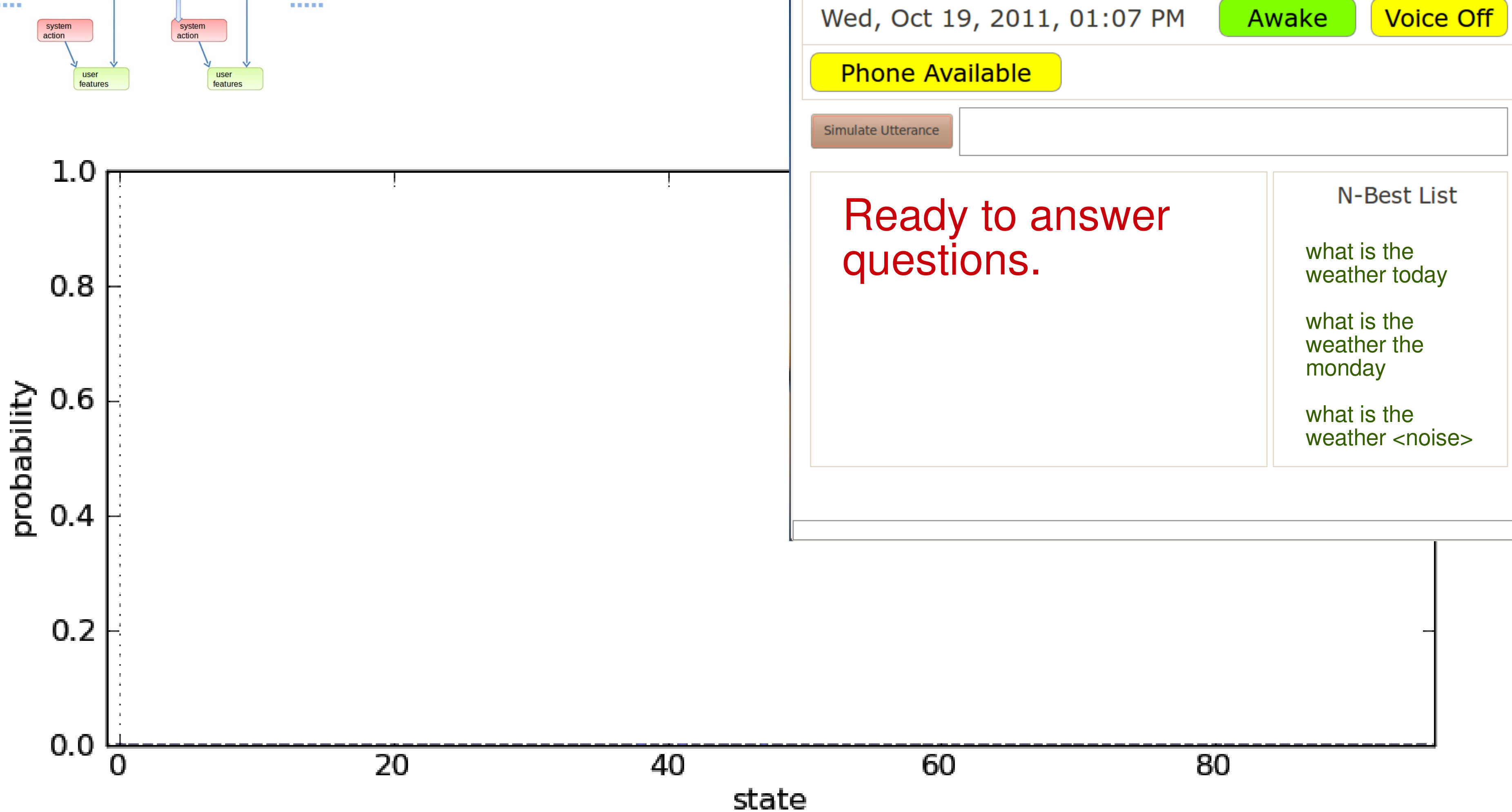
What is a POMDP?



- Sets of random variables: user intentions S, actions A, and observations Z
- System models: transition function T = P(S'|S,A), observation function Ω = P(Z'|S,A), reward function R(S,A), discount factor γ
- Probability distribution over states: belief b = P(Z)
- Policy of beliefs to actions: Π(b) → A

Spoken Dialog System POMDPs





Spoken Dialog System POMDPs

Legend: A: ['weather', ['today']]; B: ['weather', ['monday']]; C: ['weather', ['sunday']]

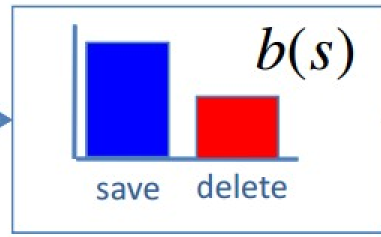
From Jason Williams (AT&T/MSR)

CSAIL

A Simple Two State Example

"Save"
"Delete"
mumble

o



a

ask
doSave
doDelete

Observation Probability

$$P(o' \mid s', a)$$

eg

"Save"	0.7
"Delete"	0.1
mumble	0.2

$$P(o' \mid save, ask)$$

Transition Probability

$$P(s' \mid s, a)$$

	save	delete
save	1.0	0.0
delete	0.0	1.0

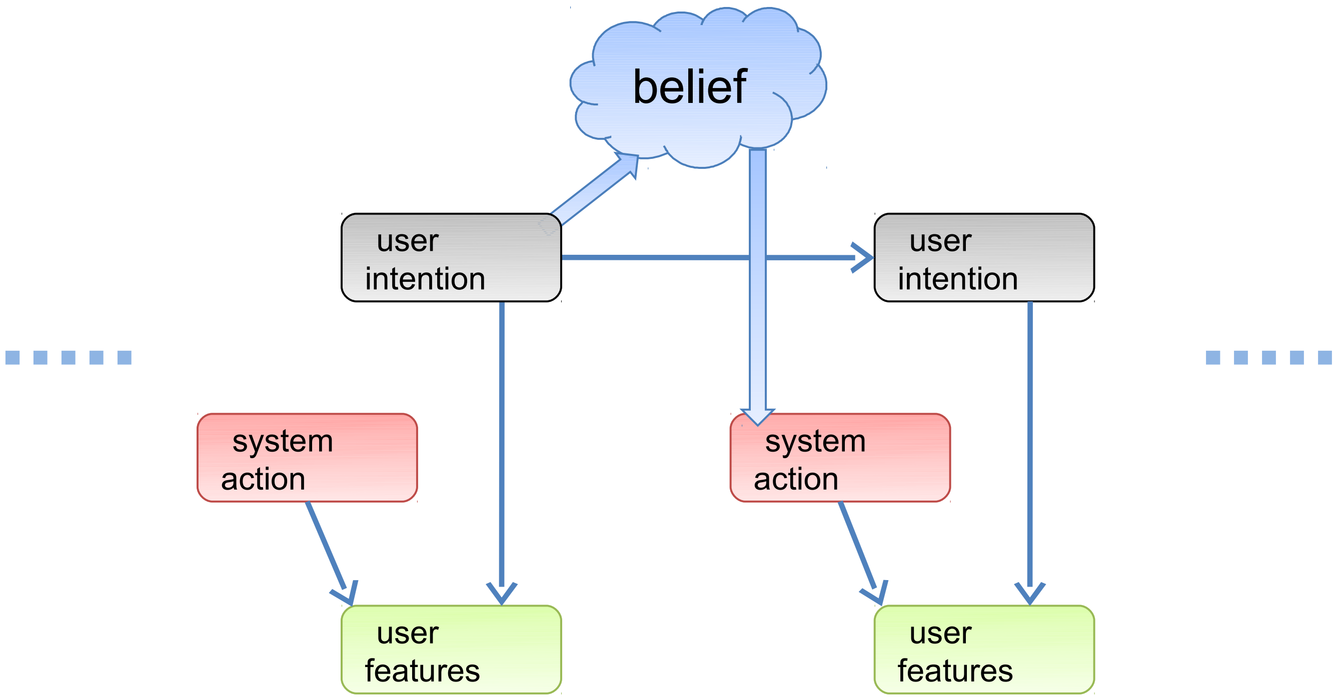
$$P(s' \mid s, ask)$$

Reward Function

$$R(s, a)$$

	save	delete
ask	-1	-1
doSave	+5	-10
doDelete	-20	+5

What is a POMDP?



- Belief update: b(s) → b'(s)

From Jason Williams (AT&T/MSR)

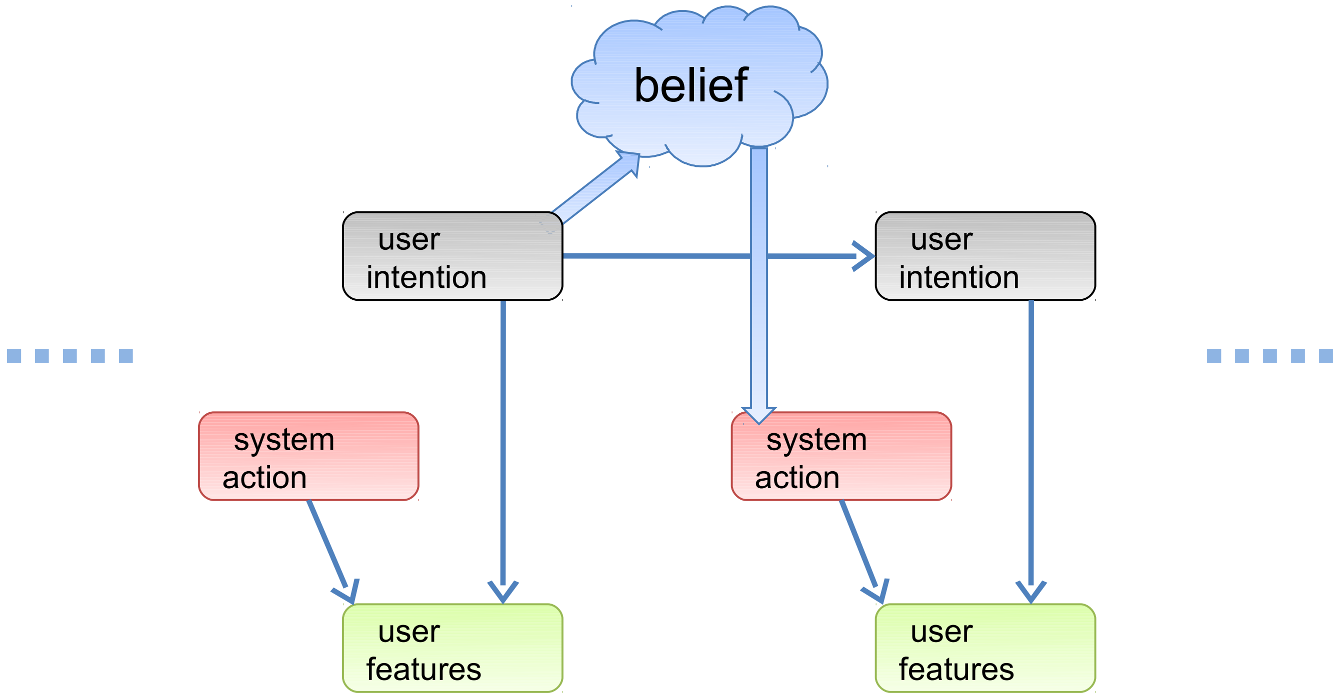
Derivation of the core update equation

$$
\begin{aligned}
b'(s') &= P(s' \mid b,a,o') \\
&= \frac{P(o' \mid s',a,b)P(s' \mid a,b)}{P(o' \mid a,b)} \\
&= \frac{P(o' \mid s',a)\sum_s P(s' \mid a,b,s)P(s \mid a,b)}{P(o' \mid a,b)} \\
&= \frac{P(o' \mid s',a)\sum_s P(s' \mid s,a)b(s)}{P(o' \mid a,b)} \\
&= \eta \cdot P(o' \mid s',a)\sum_s P(s' \mid s,a)b(s)
\end{aligned}
$$

Leslie Kaelbling, Michael Littman and Anthony Cassandra. Planning and Acting in Partially Observable Stochastic Domains. Artificial Intelligence, Vol. 101, 1998.

$$
\overbrace{b'(s')}^{\text{new belief state}} = \overbrace{\eta}^{\text{normalizing constant}} \cdot \overbrace{P(o' \mid s',a)}^{\text{observation function}} \sum_s \overbrace{P(s' \mid s,a)}^{\text{transition function}} \overbrace{b(s)}^{\text{old belief state}}
$$

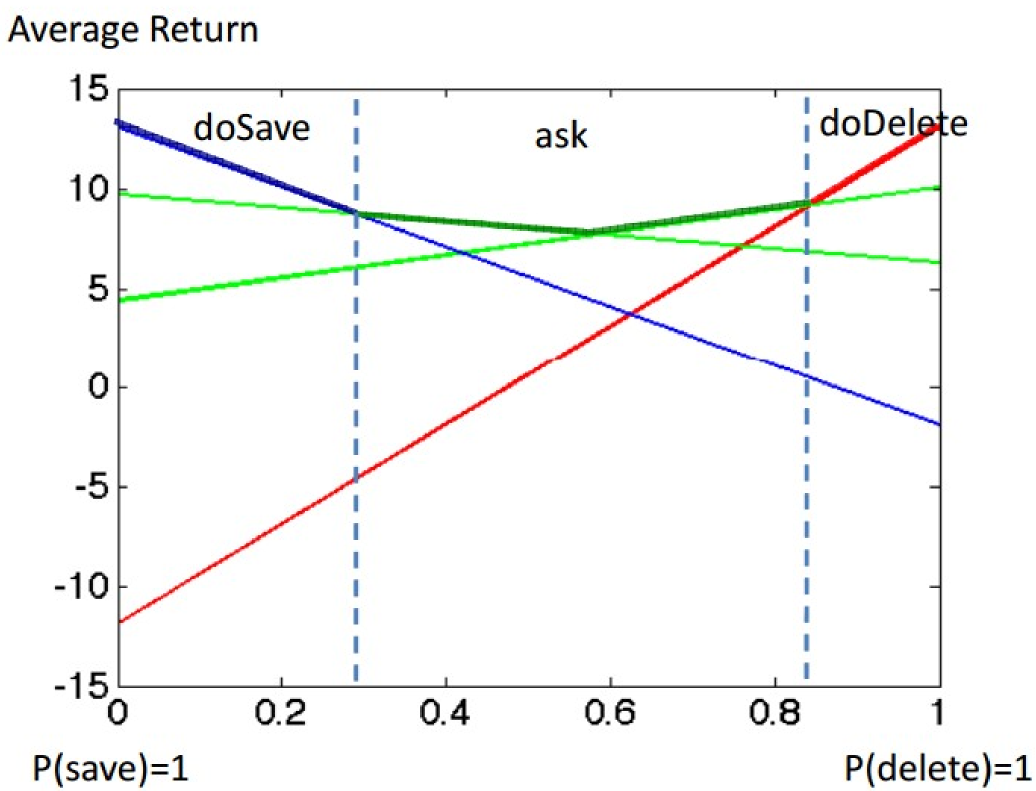
What is a POMDP?



- Solving the POMDP (computing an optimal policy): $\Pi(b) \rightarrow A$

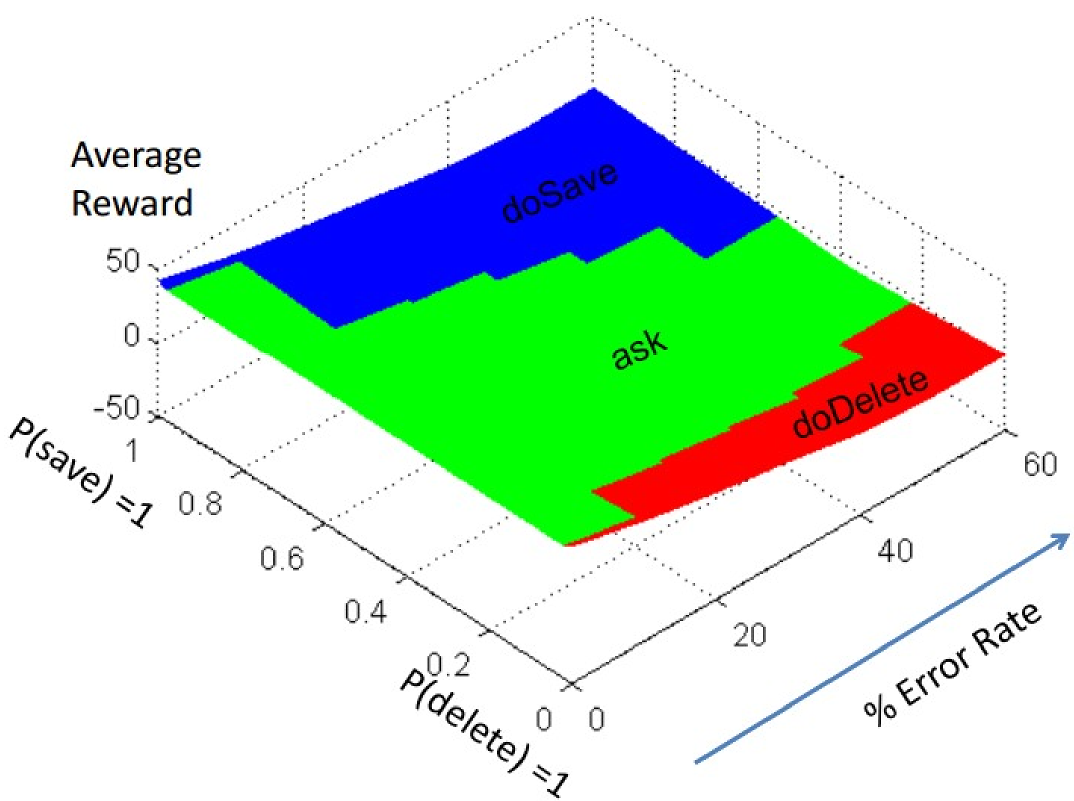
From Jason Williams (AT&T/MSR)

Policy Value Function at 30% Error Rate



From Jason Williams (AT&T/MSR)

Policy Value Function vs Error Rate



From Jason Williams (AT&T/MSR)

Example Dialog

action	observation	belief [save delete]	reward	
		[0.65 0.35]		Prior for save vs delete
ask	mumble	[0.65 0.35]	-1	No ASR output
ask	“Delete”	[0.28 0.72]	-1	ASR Correct
ask	mumble	[0.28 0.72]	-1	No ASR output
ask	“Save”	[0.65 0.35]	-1	ASR Error
ask	“Delete”	[0.28 0.72]	-1	ASR Correct
ask	“Delete”	[0.08 0.92]	-1	ASR Correct
doDelete			+5	Correct action taken

SDS-POMDP Formulation

- States, S: 62
- Actions, A: 126 (62 “submit-s”, 62 “confirm-s”, ask-initial question, terminate-dialog
- Observations, O: contain a discrete concept and a continuous confidence score)
  - 65 discrete concepts (62 possible states, YES, NO, NULL)
  - Confidence score between 0 and 1

- Transition function, T = P(S'|S,A) (62 X 62 X 126): 126 identity matrices
- Observation function: learned from labeled training set
- Reward function R(S,A) (62X126): rewards based on dialog length

Observation Model, Ω

- Note: Redefine variables as states S, actions A, and observations Z
- Observations consist of both a discrete (z_d) and a continuous (z_c) component
- z_d: concept (e.g. <weather today>, <dinner tomorrow>)
- z_c: confidence score ($0 < z_c < 1$)
- $P(Z \mid S, A) = P(z_d , z_c \mid S, A)$

Observation Model, Ω

- Multiplication rule [P(a,b) = P(a)P(b|a)]

$$P(z_d, z_c | s, a) = P(z_d | s, a) P(z_c | s, a, z_d)$$

- Discrete part: take counts from labeled training set

$$P(z_d^* | s, a) = \frac{c(z_d^*, s, a)}{\sum_{z_d} c(z_d, s, a)}$$

CSAIL

Spoken Dialog Systems at The Boston Home (TBH)

- 96-bed specialized-care residence for adults with multiple sclerosis
- Goal: voice-commanded control of wheelchair functions
- Targeted functions: weather, activities schedules, lunch/dinner menus, Skype calls

Wheelchair-based Spoken Dialog Systems at The Boston Home (TBH)

ASR Concept Error Rates

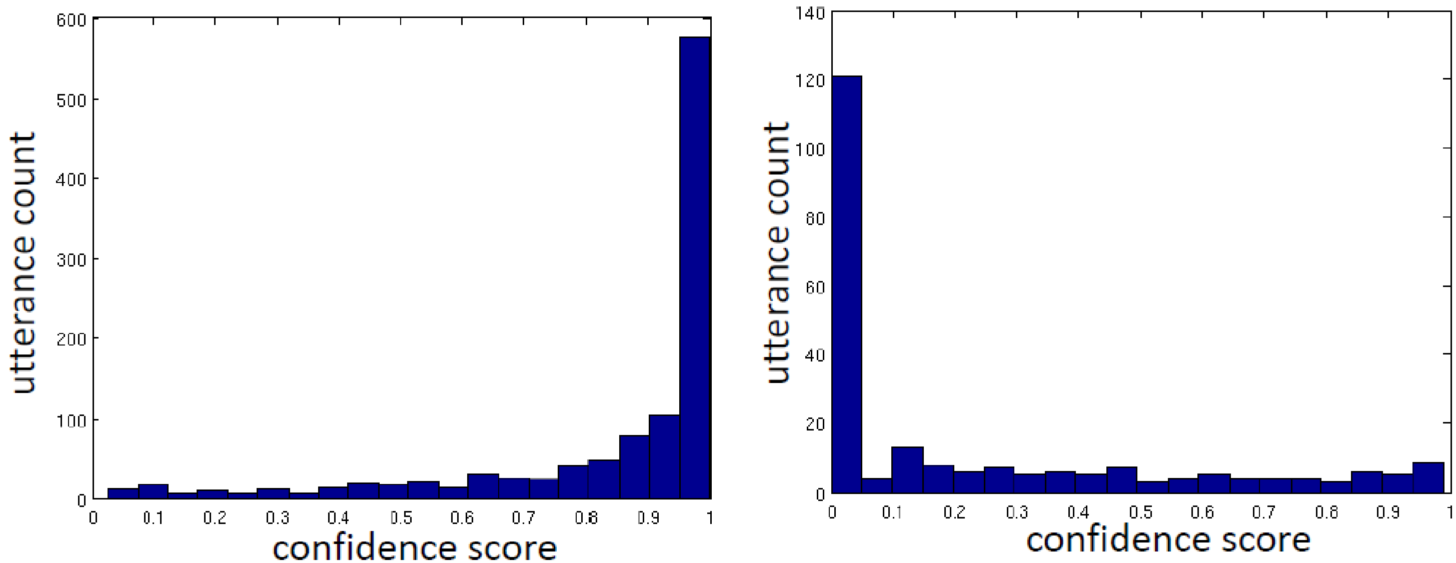
Lab Speakers	Error Rate
lab01	4.0%
lab02	7.4%
lab03	10.9%
lab04	4.3%
lab05	12.7%
lab06	3.3%
lab07	3.8%
mean	**5.7%**
std. dev.	**4.5%**

TBH Speakers	Error Rate
tbh01	12.0%
tbh02	3.7%
tbh03	5.1%
tbh04	34.6%
tbh05	57.1%
tbh06	26.1%
tbh07	9.4%
mean	**25.1%**
std. dev.	**19.5%**

Observation Model, Ω

Continuous part:

$$P(z_c|s,a,z_d) = \begin{cases} P(z_c|\text{correct observation}) & \text{if } z_d \mapsto s \\ P(z_c|\text{incorrect observation}) & \text{otherwise} \end{cases}$$

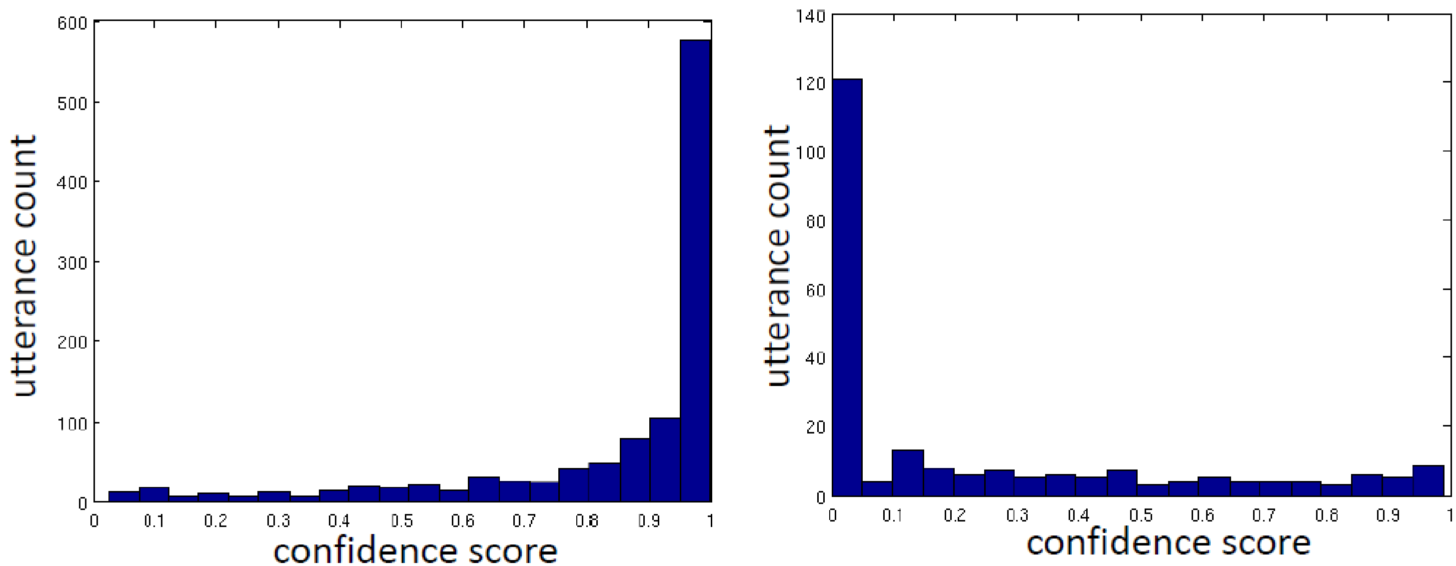


Histograms of correct (left) and incorrect (right) hypotheses

Observation Model, Ω

$$P(z_d, z_c|s, a) = P(z_d|s, a)P(z_c|s, a, z_d)$$

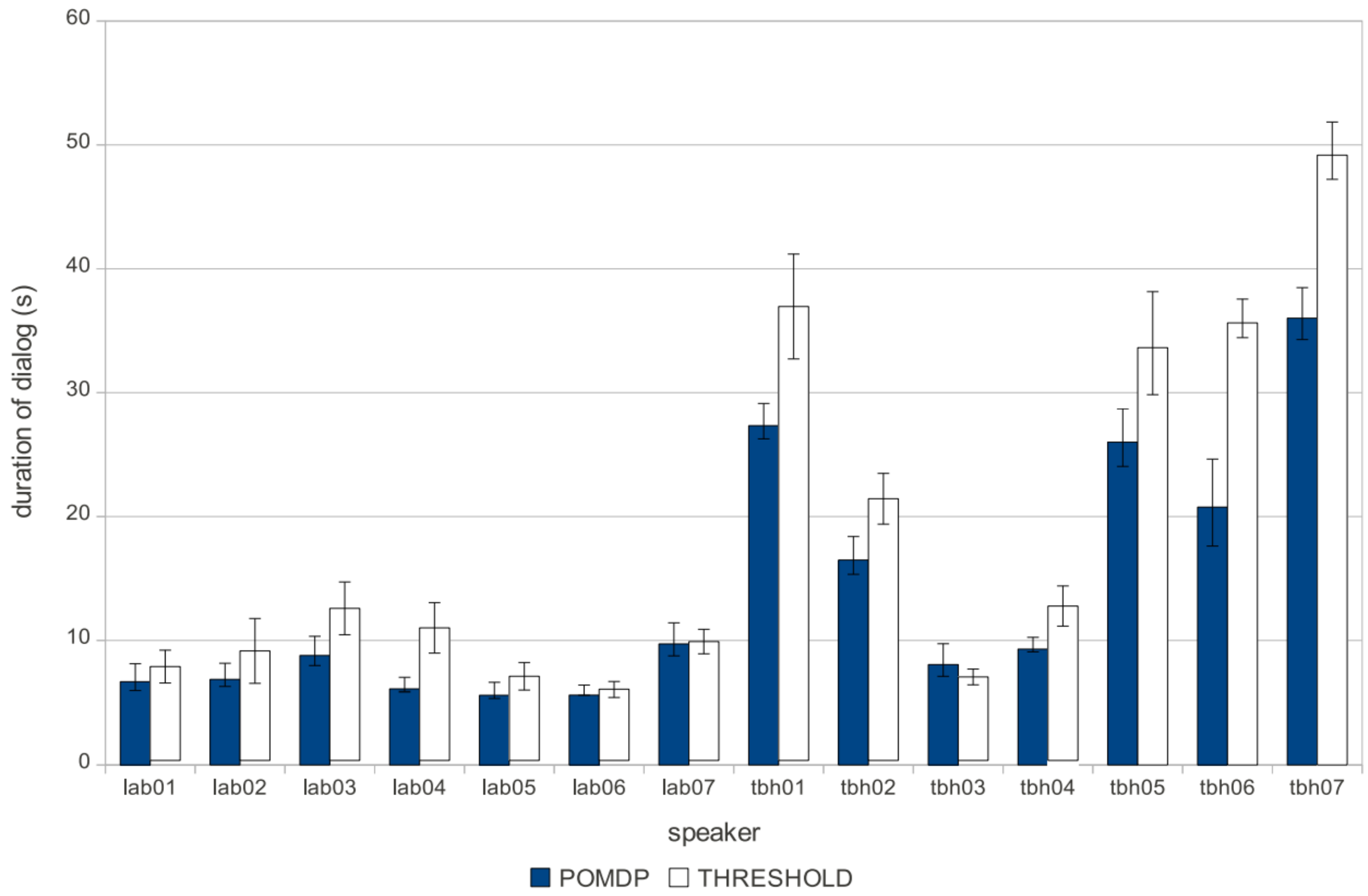
The confidence score provides information about the correctness of the hypothesis



Histograms of correct (left) and incorrect (right) hypotheses

Baseline Threshold Dialog Manager vs. POMDP Dialog Manager

- Threshold-based baseline (threshold=0.7)
- 20 dialogs/user (goals given to user by prompts)



Baseline Threshold Dialog Manager vs. POMDP Dialog Manager

- Threshold-based baseline (threshold=0.7)
- 20 dialogs/user (goals given to user by prompts)

	completed dialogs (/20)	
	POMDP	THRESHOLD
tbh01	18	13
tbh02	17	16
tbh03	20	20
tbh04	19	18
tbh05	13	5
tbh06	18	10
tbh07	17	10

Conclusions

- Partially observable Markov decision processes (POMDPs):
  - Explicitly model the user's intent as a latent variable
  - Handle uncertainty in a principled manner
  - Maximize expected reward according to some reward function
- Useful for handling speech recognition for challenging populations
- Further research directions:
  - Hierarchical/slot-based/factored state spaces
  - Scaling to more states, actions, and observations
  - Learning POMDP parameters (model-uncertainty)

Voice Interface Usage

- Possible extension: user-specific/spatial/temporal models of topics
- Example: phone-related utterances peak at 3pm (mid-afternoon) and 8pm (after dinner)