

Machine Learning and Dialogue

Partially Observable Markov Decision Processes

James Henderson

School of Informatics
University of Edinburgh

ESSLLI 2006

<http://homepages.inf.ed.ac.uk/jhender6/esslli2006/>

Outline

- 1 Partially Observable Markov Decision Processes
 - POMDPs versus MDPs
 - Reinforcement Learning
- 2 Approximating POMDPs
 - Approximating POMDPs with other POMDPs
 - Approximating POMDPs with State Sampling
- 3 Future Research Topics

Outline

- 1 **Partially Observable Markov Decision Processes**
 - POMDPs versus MDPs
 - Reinforcement Learning
- 2 **Approximating POMDPs**
 - Approximating POMDPs with other POMDPs
 - Approximating POMDPs with State Sampling
- 3 **Future Research Topics**

Formal Models of Dialogue Management (review)

- Formal Models of dialogue management share two central properties, a dialogue is **Markovian** and a **decision process**.
- “Markovian” means that what will happen in the future does not depend on the whole past history, only the current **state**.
- “Decision Process” means that what will happen in the future is influenced by the actions which the system takes.

POMDPs versus MDPs

- The two main approaches to formalizing dialogue management take different perspectives on the state.
- **Markov Decision Process** (MDP) models assume that the state is completely known, given the dialogue history.
- **Partially Observable Markov Decision Process** (POMDP) models assume that the state is partly known and partly hidden.

See [Williams & Young, forthcoming, “Partially Observable Markov Decision Processes for Spoken Dialogue Systems”]

Markov Decision Processes (review)

- The **situation** at a time t is treated as a **state** s_t . There are a finite number of possible states S .
- A **policy** π is treated as a function of the current state s_t . (More generally, $\pi(s_t) = P(a_t|s_t)$.)

$$a_t = \pi(s_t)$$

- The **environment** is treated as a probability distribution over the next state s_{t+1} given a system action a_t and the current state s_t .

$$P(s_{t+1}|s_t, a_t)$$

- The **reward** at a time t is treated as a function of the current state s_t and action a_t . (More generally, $r(s_t, a_t) = P(r_t|s_t, a_t)$.)

$$r_t = r(s_t, a_t)$$

Partially Observable Markov Decision Processes

- The **situation** at a time t is treated as a **probability distribution** b_t over states, called the **belief state**.
- A **policy** π is treated as a function of the current belief state b_t .

$$a_t = \pi(b_t)$$

- The **environment** and **reward** are defined in the same way as for MDPs, but we also define the notion of the expected reward R_t for one step. (k is a normalizing constant.)

$$R_t(b_t, a_t) = \sum_{s \in S} b_t(s) r(s, a_t)$$

- There is also an **observation probability**, which is used to compute the next belief state.

$$b_{t+1}(s_{t+1}) = P(o_{t+1} | s_{t+1}, a_t) \sum_{s \in S} b_t(s) P(s_{t+1} | s, a_t)$$

Handling Uncertainty in POMDPs versus MDPs

- MDPs need an explicit representation of uncertainty in the state. For example, states have features for:
 - slot filled or not,
 - slot confirmed or not,
 - confidence scores from ASR.
- POMDPs can express uncertainty through the distribution over states in the belief state. For example, belief states have distributions over:
 - values for a slot,
 - multiple ASR hypotheses.

Reinforcement Learning (RL) (review)

- The **optimal dialogue management policy** is the one which, on average, gives the most reward over the whole dialogue.
- In a given state, the **optimal decision** to make is the one which maximizes the reward the system can expect from the rest of the dialogue.
- The average reward over all possible future dialogues is called the **expected future reward**.

RL with POMDPs

- Given any reasonable number of possible states-action pairs, there is no tractable way to compute the exact optimal decision for a POMDP model.
- Several known **approximation methods** can tractably estimate the optimal decision for POMDP models with several hundred state-action pairs.
- But real world tasks can easily have thousands of state-action pairs. (E.g. request and confirm actions for 2 slots with 10 values each results in 2,200 state-action pairs.)

Outline

- 1 Partially Observable Markov Decision Processes
 - POMDPs versus MDPs
 - Reinforcement Learning
- 2 **Approximating POMDPs**
 - Approximating POMDPs with other POMDPs
 - Approximating POMDPs with State Sampling
- 3 Future Research Topics

Approximating POMDPs with other POMDPs

- Two approaches taken by Steve Young's group at Cambridge apply POMDP approximation methods to a *different* POMDP model at each step of the dialogue.
- They transform a POMDP for the full dialogue system into a much smaller POMDP that captures the uncertainty at that point in the dialogue.
- This transformation exploits domain knowledge about what types of uncertainty do and don't occur in dialogues.

Hidden Information State Models

- Hidden Information State (HIS) [*Williams & Young 2006*] models exploit the fact that, at any given point in a dialogue, very few pieces of information about the state are partially known.
- If a piece of information is completely known, then all alternative values have zero probability, and thus do not effect the complexity of POMDP approximation.
- If a piece of information is **completely unknown**, then all alternative values can be treated as **equivalent**.
- HIS models transform the POMDP to one which does not distinguish between states which are equivalent due to completely unknown pieces of information.

Summary POMDP Models

- Summary-POMDP models [Williams & Young 2005] are for slot filling dialogues.
- They exploit the fact that when a slot is filled, there is usually one value which is particularly likely, and the other values have a fairly uniform distribution.
- The POMDP is transformed into one with only two possible values for that slot, the most likely value or anything else.

Approximating POMDPs with State Sampling

- Current work at Edinburgh approximates the belief state of a POMDP as a short list of the **most probable states**.
- We compute this list from the most probable states from the previous step.
- States will be like those of MDPs, so a few states can encode a wide range of possible worlds through the explicit encoding of uncertainty.
- The expected future reward will also be computed as in MDPs, but averaged over the list of possible states weighted by their probabilities.

Summary of POMDPs

- POMDPs provide a principled and uniform treatment of **uncertainty** in dialogue modeling.
- Exact methods for choosing the optimal action given a POMDP belief state are **not tractable** for dialogue systems.
- **Approximate methods** for choosing the optimal action given a POMDP belief state are only tractable if knowledge about the types of uncertainty which occur in the domain are exploited in the approximation method.
- **Sampling methods** promise to provide the advantages of POMDPs while deviating minimally from traditional MDP dialogue models.

Outline

- 1 Partially Observable Markov Decision Processes
 - POMDPs versus MDPs
 - Reinforcement Learning
- 2 Approximating POMDPs
 - Approximating POMDPs with other POMDPs
 - Approximating POMDPs with State Sampling
- 3 Future Research Topics

Future Challenges in RL for Dialogue Management

- Determining:
 - additional useful features of the ISU dialogue context
 - useful combinations of features (e.g. kernel induced feature spaces)
 - more accurate measures of reward
- Learning generic strategies which can be applied to many domains (e.g. generalizing from COMMUNICATOR to an in-car scenario [*Lemon et al. submitted "Evaluating Effectiveness and Portability of Reinforcement Learning Dialogue Strategies with Real Users: The TALK TownInfo Evaluation"*]).
- Tractable policy exploration with simulated users
- Tractable POMDP Reinforcement Learning

Other Current Challenges in ML for Dialogue

- Dialogue parsing (e.g. rhetorical structure, discourse structure)
- Natural Language Understanding for dialogue systems
- ...

Dialogue Related Work at University of Edinburgh

Several groups at the University of Edinburgh are investigating different aspects of dialogue and/or dialogue systems.

- **Prof Johanna Moore**
- **Prof Bonnie Weber**
- **Dr Alex Lascaridas**
- **Drs Oliver Lemon, James Henderson, Kallirroi Georgila**