Modelling and Solving Human - Robot collaborative tasks using POMDPs

Nakul Gopalan
Advisor: Prof. Stefanie Tellex
Work with Izaak Baker

May 15, 2015

BROWN

Motivation

- 

Robot and Frank

Robot and Frank

- collaborative tasks between humans and agents with joint actions
- solve tasks without complete state information in real time
- deal with large observation spaces like from speech and gestures

BROWN

Outline

POMDPs

- world state s not known to the agent when solving the problem
- Partially Observable Markov Decision Process - $\langle S, A, T, R, O, \Omega \rangle$, Kaelbling et al. [1998]
- O observation function, T transition dynamics, Ω set of observations
- MDP over belief space where belief state b updates as:

$$b(s') = \frac{O(o|s', a) \sum_{s \in S} T(s'|a, s) b(s)}{\sum_{s' \in S} O(o|s', a) \sum_{s \in S} T(s'|a, s) b(s)},$$

BROWN

Related work

- elderly care Montemerlo et al. [2002], museum robot Burgard et al. [1999], caregiver wheelchair Doshi and Roy [2008]
- most of these works done with relatively simple observation space and state space
- they lack a joint action and state space for both partners
- POMDP based dialogue models, Young et al. [2013] have a joint state and action space
- lack physical state information which is important

BROWN

Outline

BROWN

Human task representation

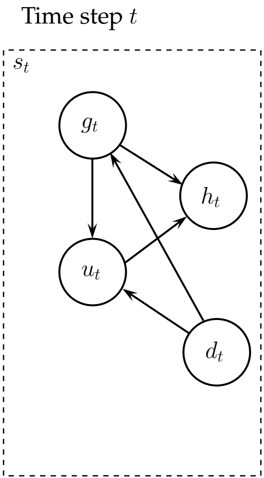

Figure : Human MDP influence diagram

Human task representation

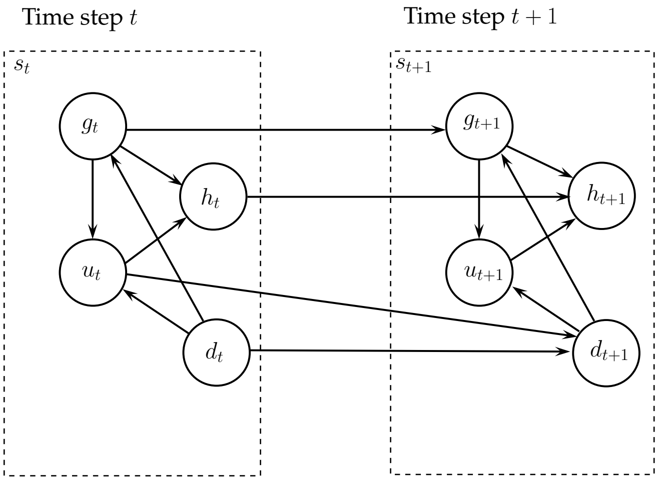

Figure : Human MDP influence diagram

Human-Robot task POMDP influence diagram

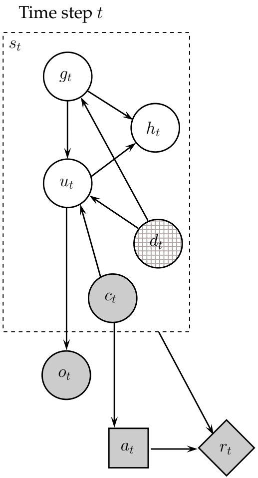

Figure : Robot-human POMDP influence diagram

Human-Robot task POMDP influence diagram

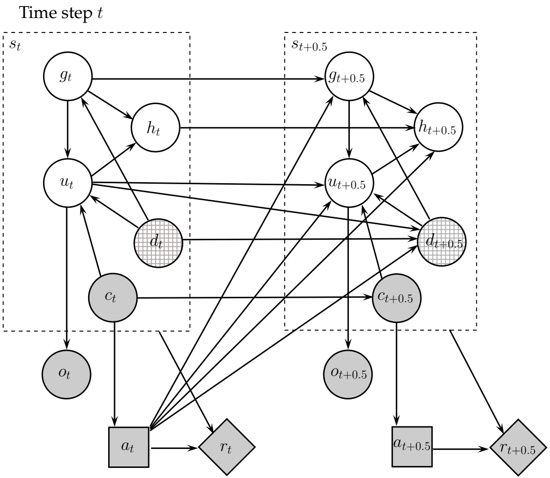

Figure : Robot-human POMDP influence diagram

Human-Robot task POMDP influence diagram

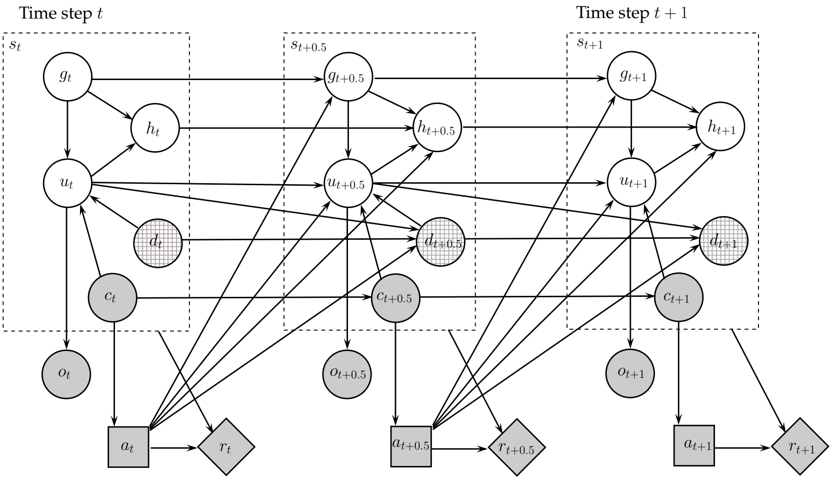

Figure : Robot-human POMDP influence diagram

Conversion to a robot-human POMDP

- human MDP - $\langle S_h, U_h, T_h, R_h \rangle$,
- S_c : $S_h, pose_r, loc_r, objects_r$
- A_r : The robot's available actions
- R_r : $f(R_h)$.
- T_c : $T_h \times T_r = P(s'|s, a_c) = \sum_{s_{0.5} \in S} P(s_{0.5}|s, a) \times P(s'|s_{0.5}, u)$
- O_c : $O_h \times O_r = P(o_c|s', a_c) = \sum_{s_{0.5} \in S} P(o_h|s', u) \times P(o_r|s_{0.5}, a)$
- robot-human POMDP $\langle S_c, A_c, T_c, R_r, O_c, \Omega \rangle$
-

$$b(g_t, u_t, h_t, d_t) = \eta P(o_t|u_t) \sum_{u_\tau} P(u_t|a_\tau, c_t, g_t, d_t, u_\tau) \sum_{u_\tau, d_\tau} P(d_t|u_\tau, d_\tau, a_\tau)$$
$$\times \sum_{g_\tau} P(g_t|g_\tau, a_t, d_t) \sum_{h_\tau} P(h_t|g_t, u_t, h_\tau, a_\tau) b_\tau(g_\tau, u_\tau, h_\tau, d_\tau)$$

BROWN

Outline

Point Based Value Iteration

- an approximate solver PBVI Pineau et al. [2003]
- performs backups in belief spaces learning piece-wise linear functions that approximate the value function over belief space
- training time cubic in the size of the state space, linear in observation space
- the method needs an initialization of belief points, which in a large state space might not be trivial to find

Natural Belief Critic

- Modified Natural Actor Critic algorithm on belief space, Young et al. [2013]
- policy gradient approach, Monte-Carlo method in the off policy/ episodic version
- basic idea: perturb policy parameters, compute value of new policy with rollouts next change policy parameters again using gradient descent
- features are hard to engineer and belief spaces can be computationally intractable

Partially Observable Monte Carlo Planning

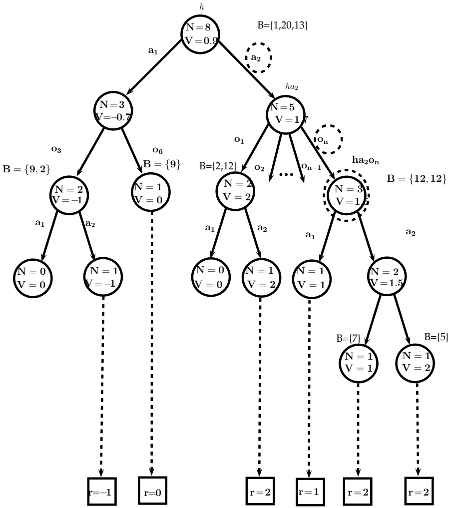

Figure : POMCP

Partially Observable Monte Carlo Planning

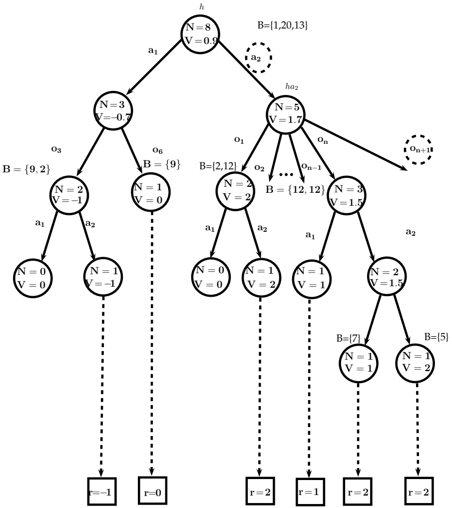

Figure : POMCP

Limited Branching Likelihood Weighted - POMCP

- Rejection Sampling: only those samples accepted that agree with the evidence in our case observations
- Likelihood Weighting: fix the evidence variables, and sample rest of the Bayes net, weight of a new sample calculated based on product of probability of evidence variables given its parents
- instead of random planning if a completely new observation is seen we weigh next state particles based on the observation probability of the new observation
- we use likelihood weighting instead of rejection sampling (LWPOMCP)
- the problem of tasseling still remains
- use ideas from sparse sampling to limit the number of observations considered (LBLWPOMCP)

BROWN

Limited branching likelihood weighted POMCP

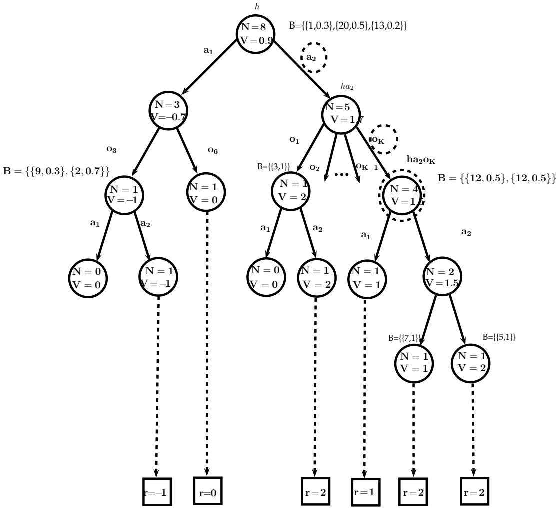

Figure : LBLWPOMCP

Outline

Childcare domain - Human MDP

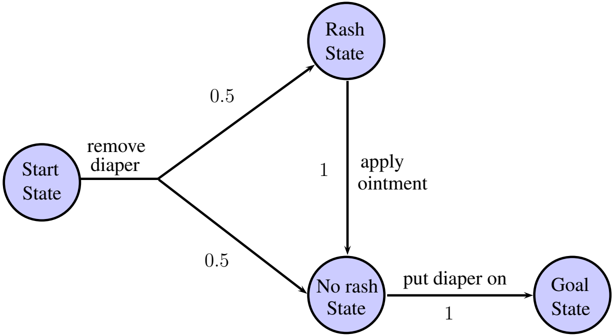

Figure : 14 state, 3 actions, average reward with increased observations, -1 reward for all actions

Childcare domain

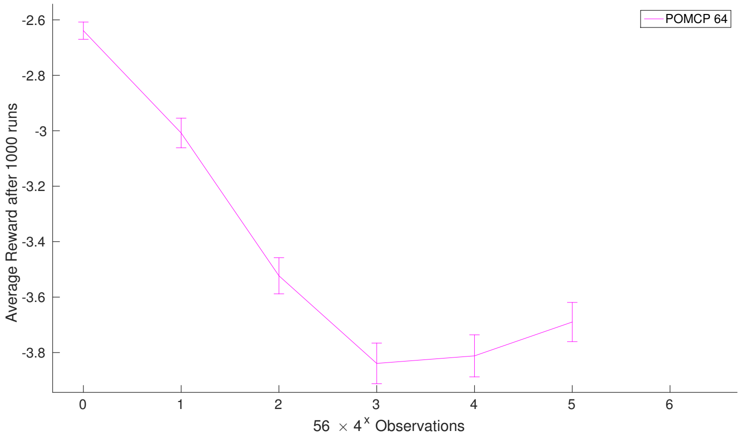

Figure : 14 state, 3 actions, average reward with increased observations, -1 reward for all actions

Childcare domain

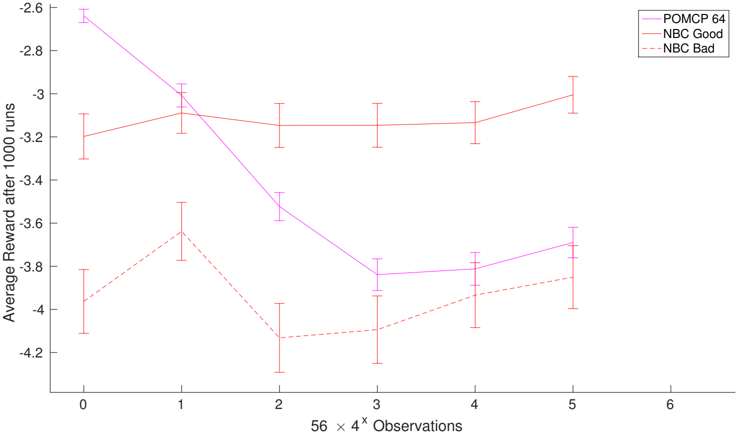

Figure : 14 state, 3 actions, average reward with increased observations, -1 reward for all actions

Childcare domain

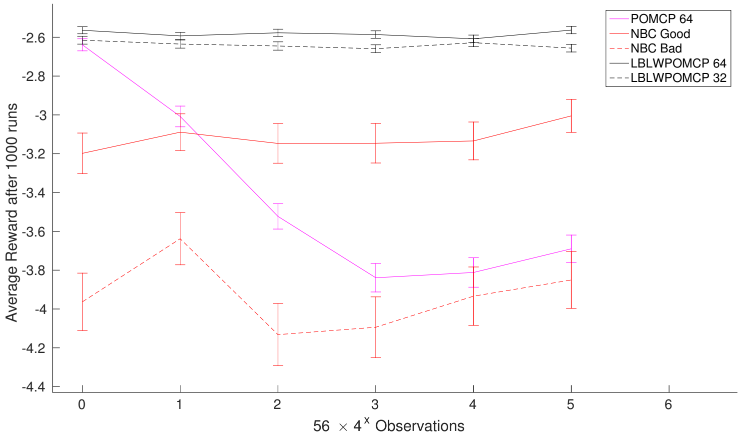

Figure : 14 state, 3 actions, average reward with increased observations, -1 reward for all actions

Confirmation based Childcare domain

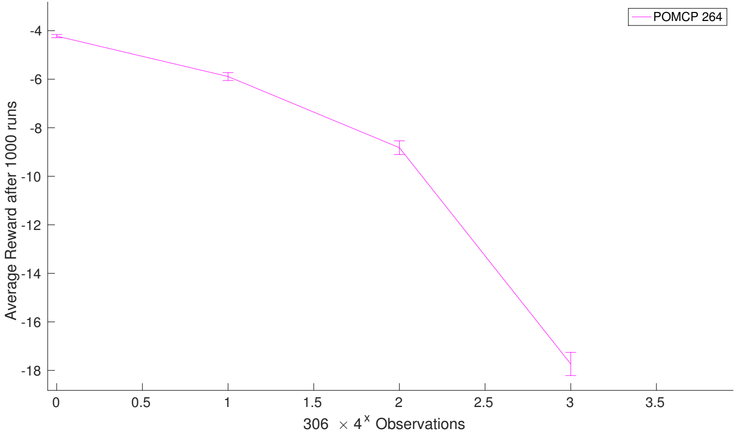

Figure : 34 states, 5 actions, average reward with increased observations, -10 reward for not asking before getting ointment, -1 reward for all other actions

Confirmation based Childcare domain

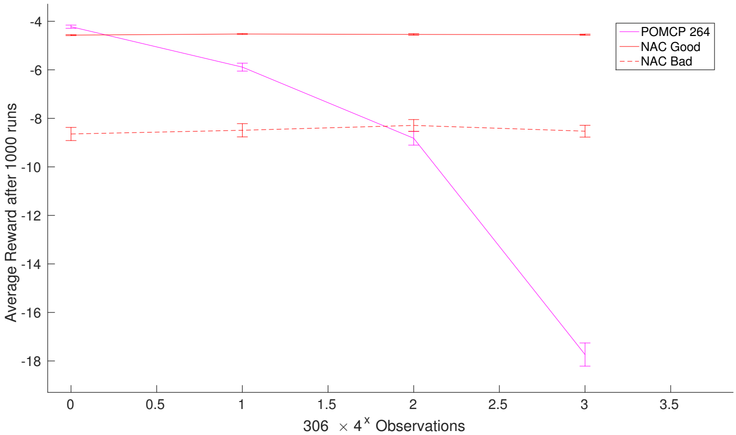

Figure : 34 states, 5 actions, average reward with increased observations, -10 reward for not asking before getting ointment, -1 reward for all other actions

Confirmation based Childcare domain

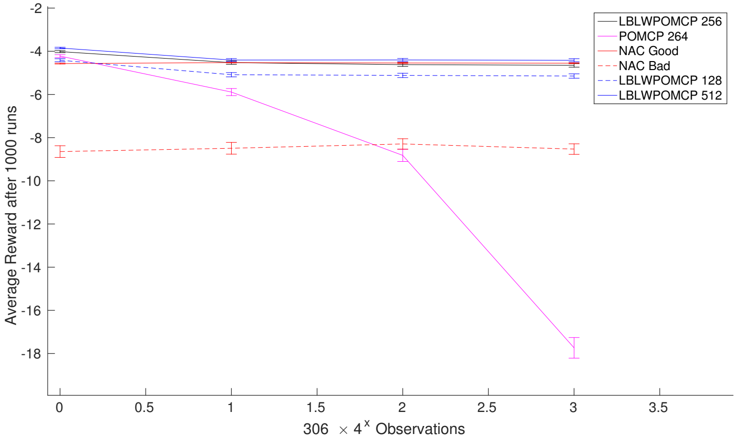

Figure : 34 states, 5 actions, average reward with increased observations, -10 reward for not asking before getting ointment, -1 reward for all other actions

RockSample domain

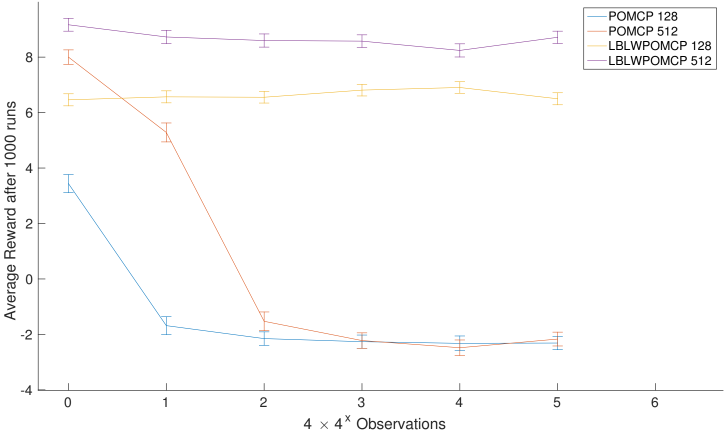

Figure : 247,808 states, 15 actions, average reward with increased observations,+10/-10 reward for mining good/bad rock

Vocab domain results

- Unigram model used for $P(o|s, a)$

Solvers	Avg. Cost	95% CI	Avg. Time (ms)	95% CI
PBVI	-2.4	0.3036	194666.5	3237.4857
POMCP	-4.2	1.5874	45.8	29.0024
LBLWPOMCP	-2.4	0.3036	25.2	9.2299
NBC	-2.6	0.3036	2504.4	149.34

Table : Result of training a vocabulary model with 35 sentence as observations over 10 runs. PBVI takes a long time to train and performs as well as LBLWPOMCP with 64 particles, and NBC slightly worse.

BROWN

Demo videos

Conclusion

- we developed methods to model human-robot collaborative tasks using POMDPs
- we looked at different solvers that can be used to solve such a POMDP model with increased observations
- LBLWPOMCP solver was found to produce better results with almost no overheard of feature engineering, hyper parameter selections, and was considerable fast when solving the domains.
- LBLWPOMCP was implemented on the robot as part of a demo

BROWN

Wolfram Burgard, Armin B. Cremers, Dieter Fox, Dirk Hähnel, Gerhard Lakemeyer, Dirk Schulz, Walter Steiner, and Sebastian Thrun. Experiences with an interactive museum tour-guide robot. *Artif. Intell.*, 114(1-2):3–55, October 1999. ISSN 0004-3702. doi: 10.1016/S0004-3702(99)00070-3. URL http://dx.doi.org/10.1016/S0004-3702(99)00070-3.

Finale Doshi and Nicholas Roy. Spoken language interaction with model uncertainty: an adaptive human-robot interaction system. *Connect. Sci.*, 2008.

Leslie Pack Kaelbling, Michael L. Littman, and Anthony R. Cassandra. Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 1998.

Michael Montemerlo, Joelle Pineau, Nicholas Roy, Sebastian Thrun, and Vandi Verma. Experiences with a mobile robotic guide for the elderly. In *AAAI/IAAI*, 2002.

Joelle Pineau, Geoff Gordon, and Sebastian Thrun. Point-based value iteration: An anytime algorithm for pomdps, 2003.

Steve Young, Milica Gasic, Blaise Thomson, and Jason D. Williams. Pomdp-based statistical spoken dialog systems: A review. *Proceedings of the IEEE*, 2013.

Observations as sentences

- sentence examples:

```
scene i-
S: who has a dirty diaper ? not me , it is you , let us change it !
R: oh oh , there is a rash , it needs meds
NR: let us clean you up and put on a new diaper
G: done ! yay ! toys for everyone

scene ii-
S: diaper change time !
NR: yes you have a poopy diaper , put on a new one
G: finished !
```

- 35 sentences with about 100 different words

Natural Belief Critic

- Modified Natural Actor Critic algorithm on belief space, Young et al. [2013]
- policy gradient approach, Monte-Carlo method in the off policy/ episodic version
- basic idea: change policy, compute value of new policy with rollouts next change policy again in the direction of improving value
- parameterized policy $\pi(\mathbf{a}_t|\mathbf{s}_t) = Pr(\mathbf{a}_t|\mathbf{s}_t, \theta)$, e.g.: $\pi(\mathbf{a}_t|\mathbf{s}_t) = \frac{e^{\theta^T \phi_{sa}}}{\sum_b e^{\theta^T \phi_{sb}}}$
- each iteration of learning a policy involves a series of rollouts, computation of the parameterized critic function in this case Q-values for the current policy
- the weight vectors of the critic function are used to update policy parameters using gradient descent
- features are hard to engineer and belief spaces can be computationally intractable

Confirmation Childcare domain run examples

```
type.start: null , Observation: obs.null#0_8,
reward: -1.0, diaper: sidetable, ointment: sidetable
type.rash: askOintment , Observation: obs.ointment_mentioned_needed#0_8,
 reward: -1.0, diaper: sidetable, ointment: sidetable
type.rash: confirmOintment , Observation: obs.ointment_confirmed_needed#0_
reward: -1.0, diaper: sidetable, ointment: sidetable
type.rash: bringOintment , Observation: obs.no_rash#0_0,
reward: -1.0, diaper: sidetable, ointment: sidetable
type.noRash: bringDiaper , Observation: obs.goal#0_7,
 reward: -1.0, diaper: sidetable, ointment: changingtable
```

Confirmation Childcare domain run examples

```
type.start: null , Observation: obs.null#0_3,
reward: -1.0, diaper: changingtable, ointment: sidetable
type.noRash: askOintment , Observation: obs.ointment_mentioned_in_negation
 reward: -1.0, diaper: changingtable, ointment: sidetable
type.noRash: null , Observation: obs.goal#0_6,
reward: -1.0, diaper: changingtable, ointment: sidetable
```

```
type.start: null , Observation: obs.null#0_1,
reward: -1.0, diaper: changingtable, ointment: changingtable
type.noRash: null , Observation: obs.goal#0_7,
reward: -1.0, diaper: changingtable, ointment: changingtable
```

UCB 1

$$Q(s,a) = Q(s,a) + c\sqrt{\frac{logN(s)}{N(s,a)}}$$

Timing data for different solvers for Childcare domain

Obs.	LBLWPOMCP 64	POMCP 64	NBC	LBLWPOMCP 32
56	948.67 (10)	1566.67 (17)	8701.13 (273)	566.12 (3)
224	1185.51 (12)	1878.26 (26)	10182.41 (83)	805.71 (7)
896	1384.11 (14)	3538.17 (172)	20754.95 (207)	946.68 (12)
3584	2306.10 (29)	13510.26 (1652)	80480.12 (1219)	1410.17 (48)
14336	6444.13 (79)	51519.87 (11817)	386865.48 (6636)	3443.83 (79)
57344	22673.28 (260)	10235.87 (171)	1184992.66 (14482)	19200.79 (1682)

Table : Average time results over 1000 runs in ms, in brackets are the 95% confidence intervals.

BROWN

Timing data for different solvers for Confirmation based Childcare domain

Obs.	LBLW 256	POMCP 256	NBC	LBLW 512
306	17990 (2665)	44208 (3555)	784842 (7096)	6459 (83)
1224	21589 (2100)	119784 (5352)	2100273 (19757)	88626 (7031)
4896	26966 (373)	728517.27 (20430)	5805935.30 (40649)	81197 (1394)
19584	130120 (2299)	60357 (11673)	17694362 (164820)	331295 (10839)

Table : Average time results over 1000 runs in ms, in brackets are the 95% confidence intervals.

BROWN

Dialogue Models modelled as a POMDP

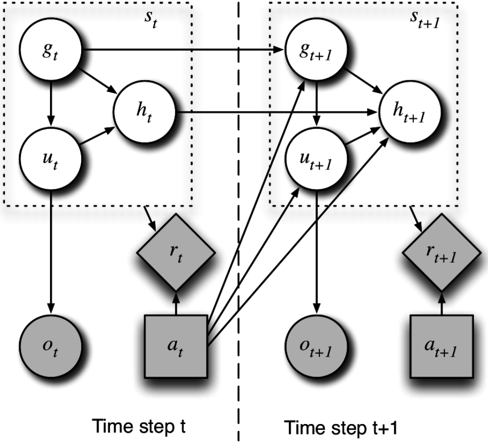

Figure : Dialogue model from Young et al. [2013]

Mathematical description of the dialogue model POMDP

$$b_{t+1}(g_{t+1}, u_{t+1}, h_{t+1}) = \eta Pr(o_{t+1}|u_{t+1}).Pr(u_{t+1}|g_{t+1}a_t).\sum_{g_t} Pr(g_{t+1}|g_t, a_t)$$

$$.\sum_{h_t} Pr(h_{t+1}|g_{t+1}, u_{t+1}, h_t, a_t).b_t(g_t, h_t)$$

- $b_{t+1}(g_{t+1}, u_{t+1}, h_{t+1})$ probability of being in a factorized state
- g_t goal, u_t user intention, h_t is the dialogue history, and o_{t+1} speech utterance, at time t
- there is an observation model, compared to a normal POMDP the transition dynamics are a combination of last three terms each of which describe the transition dynamics of individual factors.

BROWN