Dialogue Control Algorithm for Ambient Intelligence based on Partially Observable Markov Decision Processes

Yasuhiro Minami, Akira Mori, Toyomi Meguro, Ryuichiro Higashinaka, Kohji Dohsaka, and Eisaku Maeda

NTT Communication Science Laboratories, NTT Corporation
2-4 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0237 Japan
{minami,akira,meguro,rh,dohsaka,maeda}@cslab.kecl.ntt.co.jp
http://www.kecl.ntt.co.jp/

Abstract. From the viewpoint of supporting users' natural dialogue communication with conversational agents, their dialogue management has to determine any agent's action, based on probabilistic methods derived from noisy data through sensors in the real world. We believe unique Partially Observable Markov Decision Processes (POMDPs) should be applied to such action control systems. The agents must flexibly choose their actions to reach a state suitable for the users while retaining as many statistical characteristics of the data as possible. We offer two technical points to resolve this issue. One is the automatic acquisition of POMDPs ¡' state transition probabilities through DBNs with a large amount of dialogue data, and the other is applying rewards from the emission probabilities of agent actions into POMDPs' reinforcement learning. This paper proposes a method to simultaneously achieve purpose-oriented and stochastic naturalness-oriented action controls. Our experimental results demonstrate the effectiveness of our framework, which shows that the agent can generate both actions without being locked into either of them.

Keywords: Partially Observable Markov Decision Process (POMDP), dialogue management, multi-modal interaction, Dynamic Bayesian Network (DBN), agent, reinforcement learning (RL), Hidden Markov Model (HMM), Expectation-Maximization (EM) algorithm,

1 Introduction

To activate communication between users and agents, the latter have to conversationally acquire adequate tips while recognizing and understanding the situations available through person-to-person dialogues. The systems must create and establish behavioral strategies based on a large amount of data with their communication. Markov Decision Processes (MPDs) are ordinarily applied to the acquisition of strategies with reinforcement learning (RL) if the state transitions by the agents occur stochastically, depending on their current states and

actions. When we think about conversations between users and agents in the real world, diverse varieties of data exist from participant's facial expression, behaviors, and the execution timings of their actions. Basically the data hold errors and uncertainties that originated from faults and observation. As a result, learning and behavioral acquisition under MDPs do not always work effectively.

So that both users and agents are mutually understood and activated in a multi-party dialogue system, another scheme is required for controlling their interactions while gathering ambiguous multi-modal information from sensors. In this case, such information includes the state of other participants, their behaviors, and speech with paralinguistic information in addition to the surrounding environment. Observation data from sensing a real environment essentially possess errors and uncertainty about both inputs and outputs. Therefore, difficulties exist to deterministically select and take any action suitable for the changes. That is why the dialogue system should be described statistically as an action-determining, task-executing, and observing one with such probabilistic variables as actions, states, transitions, and emission. We believe a partially observable Markov decision process (POMDP) helps formally describe the environment of any user-agent dialogue system by using the measurements of sensors and the existing characteristics of the real world.

This paper shows related works about action control under uncertainty in Section 2, a POMDP applicable to multi-party dialogue in Section 3, results and evaluations of simulation experiments on one-to-one dialogue with our action control algorithm in Section 4, and finally a conclusion.

2 Related Work

POMDPs, which play an effective role in making decisions about selecting the most probabilistically reliable actions available through observed sensor data with uncertainty and their records[1, 8], are applied to spoken dialogue management[2], dialogue support for buying train tickets [3, 9], weather information dialogues [7], dialogues for DSL trouble shooting [4] and the action control of robots by human speech and gestures [5]. The results from these cases demonstrate that POMDPs compensate for uncertainty on such observed data as speech and gestures in action-determination. As a result, they get better performance in terms of the correct accomplishment of given tasks than a conventional Markov Decision Process (MDP). Since these systems are based on purpose-oriented dialogue management and we know how the agent should work, setting rewards and calculating transition probability are easy. However, if we do not know how the agent should work, such as in person-to-person communication, we have to estimate how it should work using a large amount of data. The problem is how to make the POMDP structure from a large amount of data. Although Fujita[6] solves this problem with DBNs to model a POMDP structure with a lot of data, the task is still simple and purpose-oriented.

In this paper, we propose a new type of POMDP-based action control algorithm with unique protocol acquisition and data property characteristics. One

automatically acquires a protocol specific to the conversations available through a large amount of dialogue data among users and agents. The other automatically reflects the statistical characteristics of the data upon action selection in the agent's decision processes. For English conversation lessons, such acquisition resembles how students instantly respond to a teacher's messages while learning typical conversational protocols like greetings and hand-shaking. To learn and perform like this, we have to consider the following two issues.

1. Automatic POMDP learning probabilities of internal states and acquiring a specific conversational protocol, based on Expectation-Maximization (EM) and reinforcement learning algorithms by giving a large amount of data.
2. Reflecting emission probabilities of actions upon action-selection rules through reinforcement learning.

We propose two methods to resolve the above issues.

1. Automatically acquiring the probabilities of internal states and outputting observed values with a dynamic Bayesian network (DBN)
2. Selecting actions based on emission probabilities by making internal states that match actions with one-to-one correspondence and reflecting their probabilities upon POMDP rewards

3 POMDP Controller Created from Large Amount of Data

3.1 POMDP Model for Dialogue Control

A POMDP is defined as (S, O, A, T, Z, R, γ, and b_0). S is a set of states described by $s \in$ S. O is a set of observation o described by $o \in$ O. A is a set of actions a described by $a \in$ A. T is a set of the state transition probabilities from s to s', given a, $\Pr(s'|s,a)$. Z is a set of the emission probabilities of o' at state s', given a, $\Pr(o'|s',a)$. R is a set of expected rewards when the agent performs action a at state s, $r(s,a)$. The basic employed structure is shown in Fig. 1.

Before referring to γ and b_0, we explain the state transition probability update method. In POMDP, since states are directly unobservable like in HMMs, we can only treat their distribution. Here, suppose that the distribution of states $b_{t-1}(s)$ is known. Using the transition and emission probabilities, the distribution update is performed by

$$b_t(s') = \eta \cdot \Pr(o'|s',a) \sum_s \Pr(s'|s,a) b_{t-1}(s), \tag{1}$$

where η is a factor so that the distribution summation is one. If the initial value of b is set as b_0, $b_t(s')$ can be obtained iteratively using a recursive equation.

Using this distribution, the average discounted reward at time t can be obtained as

$$V_t = \sum_{\tau=0}^{\infty} \gamma^{\tau} \sum_s b_{\tau+t}(s) r(s, a_{\tau+t}), \tag{2}$$

where γ is a discount factor. POMDP obtains a policy that is a function from $b_t(s)$ to a by maximizing the average discounted reward at infinite time. The policy, which is independent of time, is obtained by reinforcement learning.

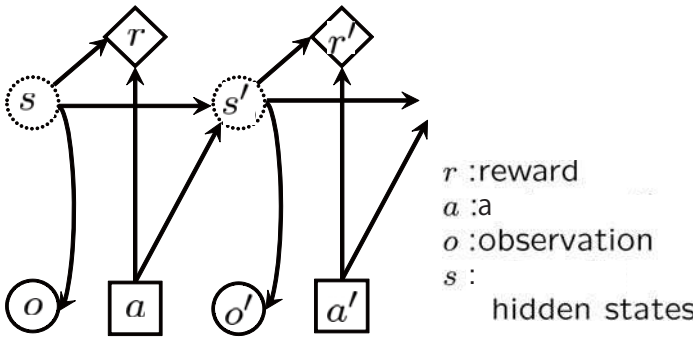

Fig. 1. POMDP structure

3.2 Training acquired specific conversational protocols

The POMDP is required to train the transition probabilities, the emission probabilities, and the rewards described in 3.1. Ordinary dialogue systems assume that the probabilities and the rewards are given. In this paper, these parameters are automatically trained from data, examples of which are shown in Table 1. The agent and user perform eight actions: shaking hands, greeting, laughing, moving, speaking, nodding their heads, shaking their heads, and doing nothing. The user and agent alternately perform an action from among the eight to have a dialogue. After the dialogue, the user evaluates whether the dialogue was a typical conversational protocol by looking at its sequence. Based on this result, the user scores it. In this example, the user shows the period of a typical conversational protocol by setting one for a certain length. We used variable d for these scores as shown in the final row in Table 1. The corresponding DBN shown in Fig. 2 is used to train the probabilities for variables o, a, d by the EM algorithm. After training the DBN, it is converted into a POMDP. Each probability in DBN is used for a corresponding probability in POMDP without modification. Since POMDPs use rewards that DBN do not have, rewards should be obtained. The objective of the POMDP is to perform typical conversation. To attain this, the rewards are obtained from the d variable by

$$r_1(s, a) = \sum_{d=0}^{1} d \times \Pr(d|s, a). \tag{3}$$

Table 1. Example of dalogue data

observation o	agent action a	user evaluation d
doing nothing	doing nothing	0
nodding	speaking	0
shaking hands	shaking hands	1
greeting	greeting	1
laughing	speaking	1
shaking head	speaking	1
greeting	greeting	1
shaking hands	shaking hands	1
doing nothing	shaking hands	0
greeting	doing nothing	0

Here we call such a conversation a purpose-oriented dialogue.

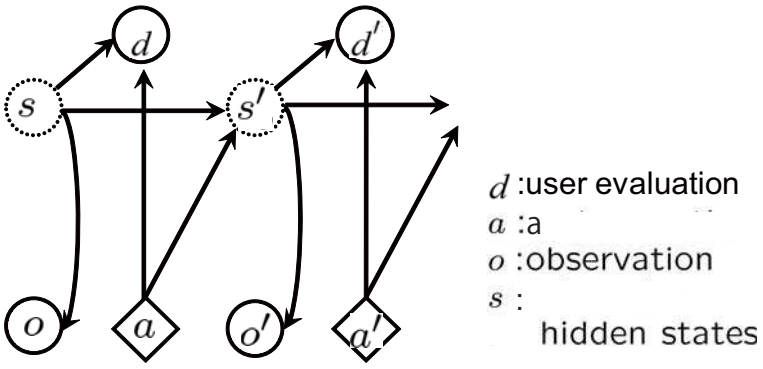

Fig. 2. DBN structure corresponding to POMDP

3.3 Reflecting Action Emission Probability on POMDP Rewards

Our goal is to make an appropriate policy using the interaction data between the users and the agent. Our target interaction characteristic is that the interaction should be processed based on probabilistic characteristics; however sometimes typical protocols occur, and the agent should obey them. The problem is how to obtain the policies that achieve this behavior by reinforcement learning.

We propose the following methods to solve this problem. First we introduce an extra hidden DBN and POMDP states to the ordinary states as $s = (s_o, s_a)$ (Fig. 3-4).

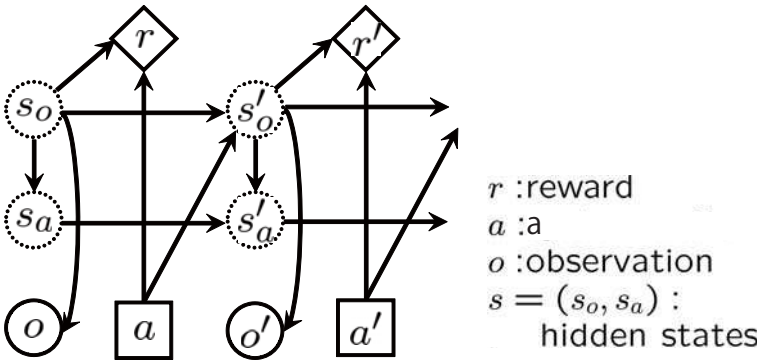

Fig. 3. POMDP structure employed in this paper

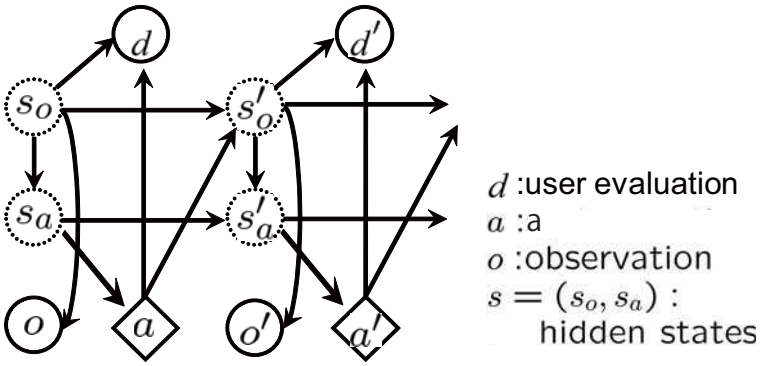

Fig. 4. DBN structure employed in this paper

The training algorithm's procedure is as follows (Fig. 5).

1. Positive reward 1.0 is set to the typical protocol data and 0.0 reward to the other data (see Table 1).
2. A DBN is trained whose probabilities are $\Pr(s'|s,a) \approx \Pr(s'_o|s_o,a)\Pr(s'_a|s'_o,s_a)$, $\Pr(d|s_o,a)$, and $\Pr(o'|s'_o)$ (Fig. 4). d is also treated as a random variable.
3. The DBN is converted into a POMDP, where we convert an evaluation random variable into POMDP fixed rewards by Eq. (3).
4. We set reward $b_{\tau+t}(s_a)r(s_a,a_{\tau+t})$, so that if $b_{\tau+t}(s_a)$ is high, POMDP may obtain a higher reward.
   If $a = s_a$, we set $\Pr(a|s_a) = 1$ in the DBN (Fig. 4) so that s_a corresponds one-on-one with a. Based on this, if $a_t = s_a$ is given, we obtain

$$\begin{aligned}&\Pr(a_t|o_1,a_1,\ldots,a_{t-1},o_t)\\ &\quad=\sum_{s'_a}\Pr(a_t|s'_a)\Pr(s'_a|o_1,a_1,\ldots,a_{t-1},o_t) \qquad (4)\\ &\quad=\Pr(s_a|o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t)=b_t(s_a). \qquad (5)\end{aligned}$$

   This is for propagating the emission probabilities of the actions into the probabilities of the hidden states. Our objective here is to select a_t so that the probability of a_t is maximized when $o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t$ is given. The reward should be set to satisfy this. This means the rewards should be set by maximizing Eq. (5). To do this ¡/we set $r_2(s=(*,s_a),a)=1$ when $s_a=a$, where * is arbitrary s_o. Otherwise, $r_2(s=(*,s_a),a)=0$. Replacing r in Eq. (2) into r_1+r_2 as

$$r(s,a)=r_1((s_o,*),a)+r_2((*,s_a),a), \qquad (6)$$

   we obtain new objective function V_t. We modify rewards r_1 described in 3.1 using * so that we can treat extra hidden states s_a.
   The POMDP is then trained by reinforcement learning to generate the policy. Using this formulation, the POMDP can select the action that simultaneously gives higher probability of the action and obeys the purpose-oriented action control.

4 Evaluation and Result

We prepared two types of patterns as typical conversational protocols between the agent and the user. The following is the sequence of one protocol. They shake hands and greet each other. Then they talk randomly, laugh, and nod their heads. Finally they greet and shake hands again. In the other sequence, first the user moves and the agent does nothing. Then they greet each other, speak randomly, laugh, and nod their heads. Next they greet each other. Finally the user moves and the agent does nothing. The amount of these data is one tenth of the total data.

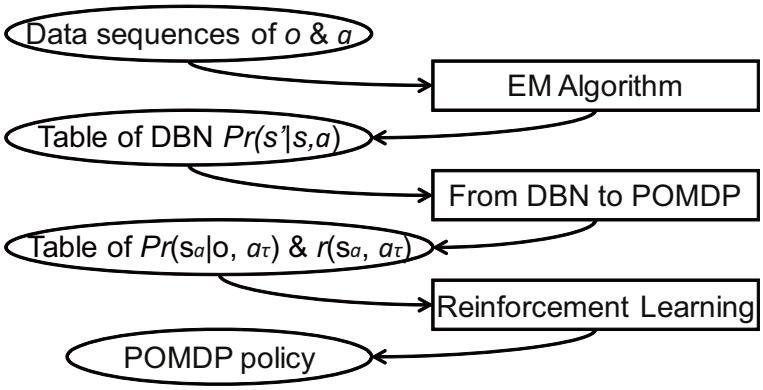

Fig. 5. Policy generation procedure by POMDP

The rest of the data is generated so that the joint probabilities of the observation and action pairs (shaking hands/shaking hands, greeting/greeting, laughing/laughing, moving/moving, speaking/speaking, nodding/speaking, shaking head/speaking, doing nothing/doing nothing) have the highest probabilities. We call these everyday dialogue data. The lengths of the sequences of all samples were identical. When the lengths of the typical protocol samples were shorter than the fixed length, we added everyday dialogue data at the start and the end of the samples. 10,000 samples were made for the training data. For the typical protocols, we set the reward values to one per frame. For everyday dialogue data we set the reward values to zero and trained a DBN using these data. Then the DBN was converted to POMDP by the proposed method. 2,000 samples were used for the evaluation data. Only observation data were generated using the same algorithm to generate the training data. We evaluated the action generation results of two POMDPs: our proposed POMDP and a purpose-oriented POMDP that only gives rewards to typical conversational protocol training data.

The experimental results show that both methods generated complete sequences for all the data of the typical conversational protocols. Table 2 shows the results of the joint probabilities of the observation and agent action pairs. The second column shows the joint probabilities for the training data. Although we can tune the weight value for the rewards of and , weight tuning was not performed for the proposed method. The purpose-oriented POMDP tunes the typical conversational protocol data and always tries to attract the user to the typical conversational protocols. It did not generate any laughing/laughing, moving/moving, nodding/speaking, shaking head/speaking, and shaking head/speaking pairs. The far right column shows that the proposed method improved the joint probabilities, confirming that the proposed method simultaneously achieved purpose-oriented and stochastic naturalness-oriented action control.

Table 2. Joint probability of observation and action pairs

Obs.-act. pairs	Training sample	Purpose-oriented POMDP	Proposed POMDP
shaking hands/shaking hands	0.09	0.13	0.13
greeting/greeting	0.10	0.11	0.13
laughing/laughing	0.08	0.00	0.02
moving/moving	0.08	0.00	0.002
speaking/speaking	0.04	0.00	0.00
nodding/speaking	0.09	0.00	0.08
shaking head/speaking	0.09	0.00	0.05
doing nothing/doing nothing	0.10	0.00	0.05

5 Conclusion

In this paper, we presented a POMDP-based dialogue control scheme that can automatically acquire a conversational protocol typical of daily dialogues with a reinforcement learning algorithm, which can reflect statistical characteristics automatically acquired with a large amount of dialogue data based on the agent's decision processes in selecting actions. Our experiment results indicate that the action control algorithm functions effectively through our simulation experiments.

References

1. Pineau, J., Gordon, G., and Thrun, S.: Point-Based Value Iteration: an anytime algorithm for POMDPs. In: IJCAI, pp.1025-1032, (2003)
2. Roy, N., Pineau, J., and Thrun, S.: Spoken Dialogue Management Using Probabilistic Reasoning. In: Proceedings of 38th ACL 2000, Oct.(2000)
3. Williams, J., Poupart, P., and Young, S.: Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management. In: 6th SIGDial Workshop on Discourse and Dialogue, pp.25-34, Sept.(2005)
4. Williams, J.: Using Particle Filters to Track Dialogue State. ASRU 2007, pp.502-507, (2007)
5. Schmidt-Rohr, S. R., Jäkel, R., Lösch, M., and Dillmann, R.: Compiling POMDP models for a multimodal service robot from background knowledge. In: Proceedings of European Robotics Symposium 2008, 44, pp.53-62 (2008)
6. Fujita, H.: Learning and decision-planning in partially observable environments, Ph.D. dissertation, Nara Institute of Science and Technology, (2007)
7. Kim, K., Lee, C., Jung, S., and Lee, G. G.: A Frame-Based Probabilistic Framework for Spoken Dialog Management Using Dialog Examples. the 9th SIGdial Workshop on Discourse and Dialogue, pp.120-127, (2008)
8. Poupart, P.: Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes. Ph.D. dissertation, University of Toronto, (2005)

9. Williams, J., Poupart, P., and Young, S.: Factored partially observable markov decision processes for dialogue management. In: IJCAI Wkshp. on K&R in Practical Dialogue Systems, pp.75-82, (2005)