Statistical Dialogue Management using Intention Dependency Graph

Koichiro Yoshino[1,2], **Shinji Watanabe**[1], **Jonathan Le Roux**[1], **John R. Hershey**[1]

[1]Mitsubishi Electric Research Laboratories, 201 Broadway, Cambridge, MA, 02139, USA
{watanabe,leroux,hershey}@merl.com
[2]School of Informatics, Kyoto University, Sakyo, Kyoto, 606-8501, Japan
yoshino@ar.media.kyoto-u.ac.jp

Abstract

We present a method of statistical dialogue management using a directed intention dependency graph (IDG) in a partially observable Markov decision process (POMDP) framework. The transition probabilities in this model involve information derived from a hierarchical graph of intentions. In this way, we combine the deterministic graph structure of a conventional rule-based system with a statistical dialogue framework. The IDG also provides a reasonable constraint on a user simulation model, which is used when learning a policy function in POMDP and dialogue evaluation. Thus, this method converts a conventional dialogue manager to a statistical dialogue manager that utilizes task domain knowledge without annotated dialogue data.

1 Introduction

Statistical approaches based on reinforcement learning, such as the Markov decision process (MDP) and partially observable Markov decision process (POMDP), have been successfully applied to dialogue management (Levin et al., 2000; Williams and Young, 2007; Li, 2012). These approaches allow us to consider all possible future actions of a dialogue system, and thus to obtain a new optimal dialogue strategy which could not be anticipated in conventional hand-crafted dialogue systems. Moreover, the statistical dialogue framework can be combined with conventional rule-based dialogue management in hybrid systems, (Williams, 2008; Lee et al., 2010), which combine the optimal dialogue strategy in the statistical approach with the lower cost of data and maintenance of the rule-based approach.

Our research focuses on a practical application of a hybrid statistical dialogue management based on POMDP to conventional rule-based dialogue management via the use of an intention dependency graph (IDG). The IDG derives from the conventional rule-based dialogue system (Dahl et al., 1994; Bohus and Rudnicky, 2003), and it constrains the transition matrix and provides a user simulation as a substitute for dialogue data.

The object of POMDP optimization is to produce a policy that maps from user states to system actions such that the overall expected cost of the dialogue is minimized. Such optimization typically requires data from dialogue corpora, which are manually annotated with task-oriented dialogue-act tags. On the other hand, the benefit of a hybrid approach is that human domain knowledge can be used to constrain the possible user states in the dialogue manager. We follow this idea by using an IDG, which expresses the task-domain knowledge through a directed graph of states from more general intention categories to more specific parameters of the intention categories. **Figure 1** shows an example of such a graph, where each node is associated with a (potentially partial) user intention. In previous studies, this kind of domain knowledge is used to restrict the user state and system action state space (Lemon et al., 2006; Williams, 2008; Young et al., 2010; Varges et al., 2011). However, our approach does not restrict the possible system action states, but transfers the information structure to the definition of user simulation and state transition probabilities. The system is allowed to consider all possible system actions by following the user states that reflect the IDG.

2 Statistical dialogue management

The main random variables involved at a dialogue turn t are as follows. $s^t = i \in \mathcal{I}_s$ is the hidden true user statte at turn t. It is constrained by the hidden user goal $g \in \mathcal{I}_g$ and the true user state at the previous turn. $o^t = l \in \mathcal{I}_s$ is the observation

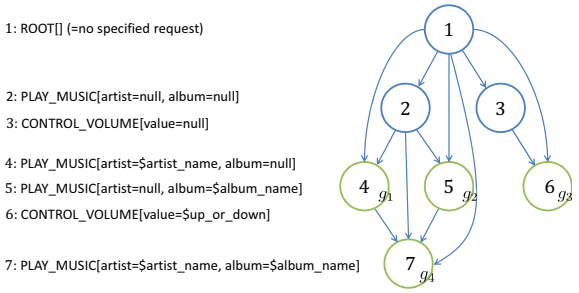

Figure 1: An example of a directed intention dependency graph.

of the user state by the system. It includes errors caused by automatic speech recognition (ASR), natural language understanding (NLU) and intention understanding (IU). Uncertainty on the observation o^t caused by errors in the preprocessor (ASR, NLU, and IU) is encompassed in the conditional probability $O^t_{li} = p(o^t = l|s^t = i)$. $a^t = k \in \mathcal{K}$ is the system action. $\hat{k}$ is the optimal system action that is acquired in the learning step. The goal of statistical dialogue management is to output an optimal system action $\hat{a}^t = \hat{k}$ given an observation o^t, based on the probability of s^t in a soft decision manner. The probability of the user state s^t given an observation sequence $o^{1:t}$ from 1 to t with confidence $O^{1:t}$ is denoted by $b^t_i = p(s^t = i|o^{1:t}; O^{1:t})$, and referred to as "belief". To avoid clutter, we will usually omit $O^{1:t}$.

2.1 Belief update

We consider a belief update equation based on the graphical model shown in **Figure 2**, assuming that the system actions $a^{1:t}$ are given. We can obtain the following update equation from b^t_i to $b^{t+1}_{i'}$:

$$b^{t+1}_{i'} = p(s^{t+1} = i'|o^{1:t+1}) \quad (1)$$

$$\propto \sum_i p(o^{t+1}, i'|i)(b^t_i)^\beta, \quad (2)$$

where β is a forgetting factor for the belief, and $0 \leq \beta \leq 1$. Then, by introducing the system action $a^t = k$ based on the sum rule, we can rewrite $p(o^{t+1}, i'|i)$ in Eq. (2) as follows:

$$\sum_k p(o^{t+1}, i', k|i) = \sum_k p(o^{t+1}, i'|i, k)\delta_{\hat{k}k}$$

$$= p(o^{t+1}|i')p(i'|i, \hat{k}) \quad (3)$$

where $p(k|i) = \delta_{\hat{k}k}$ is obtained by the decision making step in the POMDP. We rewrite the distributions in Eq. (3) as follows: $p(i'|i, \hat{k}) = T_{ii'\hat{k}}$

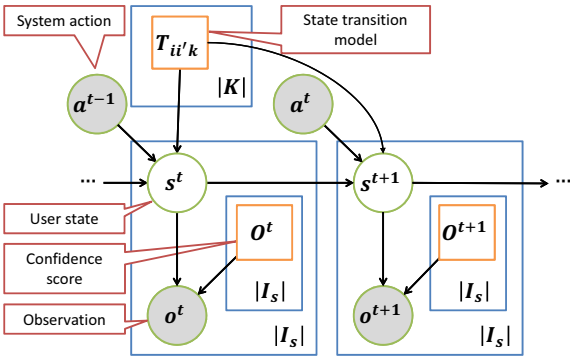

Figure 2: Graphical model of user state sequences given system actions a^{t-1} and a^t. This graphical model shows user behavior that is observed from the system.

and $p(o^{t+1} = l|i') = O^{t+1}_{li'}$. $T_{ii'\hat{k}}$ are the user state transition probabilities given system action $\hat{k}$, and $O^{t+1}_{li'}$ are the confidence scores given by the pre-processor. In conventional studies, the state transition probabilities $T_{ii'\hat{k}}$ are learned from annotated data. In our scheme, the probabilities can be obtained by using the IDG, as described in Section 3.4. We finally obtain

$$b^{t+1}_{i'} \propto O^{t+1}_{li'} \sum_i T_{ii'\hat{k}}(b^t_i)^\beta. \quad (4)$$

Once the system estimates the belief b^t_i, it can output the optimal action $\hat{a}^t$ as $\hat{a}^t = \pi^*(\{b^t_i\}^{|\mathcal{I}_s|}_{i=1})$. π is called a policy function, and π^* is an optimal policy function pre-computed in the learning step described in the following Section.

2.2 Learning step

The aim of the learning step in reinforcement learning is to acquire the best policy π^*. Many algorithms formulated to solve the reinforcement learning problem have been proposed (Shani et al., 2013). While most advanced algorithms require transition probabilities $T_{ii'\hat{k}}$ that are calculated using annotated corpora, our approach aims at learning a POMDP without any data. We thus use one of the most basic algorithms, Q-learning (Watkins and Dayan, 1992), as it can acquire the policy without using transition probabilities. Q-learning relies on the estimation of a Q-function $Q(b^t, a^t)$, which computes the expected future reward of a system action a^t at dialogue turn t given the current belief $b^t = \{b^t_i\}^{|\mathcal{I}_s|}_{i=1}$ of the user state. The Q-function can be obtained by iterative updates on training dialogue data. The up-

dates do not involve the transition probabilities $T_{ii'\hat{k}}$, thus we can acquire the optimal policy without requiring knowledge of this function. Given the Q-function, the optimal policy is determined as $\pi^*(b^t) = \arg\max_{a^t} Q(b^t, a^t)$.

3 Dialogue management using intention dependency graph

3.1 Intention dependency graph

An intention dependency graph (IDG) is a representation of a user's intention in a hierarchy, with broad categories of the intention at the top, and specific instantiations of those categories at the bottom, as shown in **Figure 1**. A child node in the graph represents a more specific intention than the parent node, so that the flow from top to bottom represents the completion of the full specification of an intention. However, the graph is not necessarily a tree, and hence there may be multiple paths from a parent node to any descendent node. A node that is fully specified and actionable by the system can be considered a user goal. In node 7, which is a child of node 2, both the album and artist are specified, and the system has enough information to perform the desired action. Such a graph is automatically generated from task knowledge that is usually designed by hand for a conventional rule-based dialogue manager (Dahl et al., 1994; Bohus and Rudnicky, 2003), as a graphical user interface, and can be obtained by forming a taxonomy of the possible system actions. In our context, a node in this graph represents a hypothesis of the user's intention and/or goal.

3.2 User simulator

Training a statistical dialogue management system in the absence of large amounts of dialogue data requires a user simulator to ensure adequate coverage of possible user states. In a general dialogue, the system action and the user state would follow a dialogue history and lead toward a user goal. The simulator thus samples user states $s^{t+1} = i$, at every time step, tending toward a user goal g, and depending on the previous system action $a^t = k$. Thus, our approach defines the sampling distribution $p(i|g, k)$ by using IDG. Our approach gives uniform distribution to hypotheses that are outputted by the IDG. We show an example IDG in **Figure 1** and a dialogue example in **Figure 6** of Appendix.

3.3 Learning without any annotated data

We discuss the learning for the POMDP that uses our IDG. In our task, no data can be referred to and we cannot calculate the transition probability that is generally calculated from an annotated data for the belief update. This property makes it impossible to establish the exact value of the state-value function. In standard POMDP learning, sampling belief point approaches that select a small set of representative belief points such as point-based value iteration (PBVI) can be applied (Pineau et al., 2003). However, it is difficult to sample a small set of belief points without any tagged data. Therefore, we calculate the action-value function $Q(b^t, a^t)$, and simulate the noise with a grid-based approach (Lovejoy, 1991; Bonet, 2002). The grid-based approach can select points in accordance with a grid and a noise parameter η that is released from the data. In our learning approach, a sample of belief $b_i^t = p(s^t = i|o^{1:t}, O^{1:t})$ is given by $p(o^{t+1} = l|i') = O_{li'}^{t+1}$ where

$$O_{li'}^{t+1} = \begin{cases} 1 - \eta & l = i' \\ \frac{\eta}{|\mathcal{I}_s| - 1} & l \neq i'. \end{cases} \quad (5)$$

We tried noise $\eta = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$. The resulting policy does not reflect the belief update, but we can use the belief update method that follows the IDG.

3.4 State transition and belief update

The state transition probability $T_{ii'\hat{k}}$ is one of the most important components of the belief update in the POMDP framework. To obtain the transition probabilities, we usually require user state and system action data with annotated tags. However, we cannot calculate the probability because of the lack of annotated data. Therefore, we define the state transition probability by using an IDG similar to user simulation, as discussed in Section 3.2. By employing time-invariant user goal g, time-variant user state $s^t = i$ and time-variant best system action $a^t = \hat{k}$ in Section 2, we can represent the state transition probabilities, as follows:

$$p(i'|i, \hat{k}) = \sum_g p(i'|g, i, \hat{k}) p(g|i, \hat{k}) \quad (6)$$

We approximate $p(i'|g, i, \hat{k})$ by user simulator $p(i'|g, \hat{k})$. This means that the next user state i' does not depend on the previous user state i. We approximate $p(g|i, \hat{k}) \simeq p(g|i)$ because the user

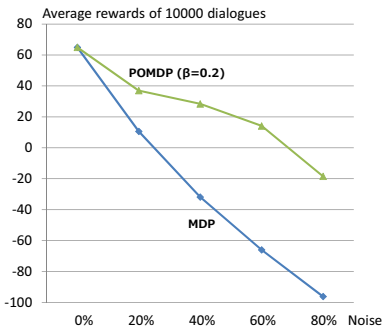

Figure 3: Average rewards of 10000 dialogues between the obtained dialogue manager and the user simulator.

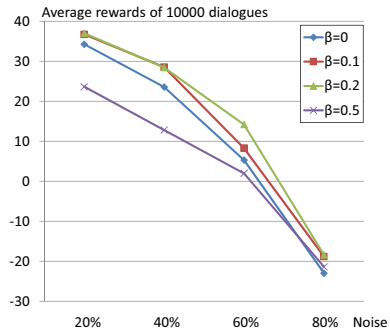

Figure 4: The effect of forgetting factor β as regards the average rewards of 10000 dialogues.

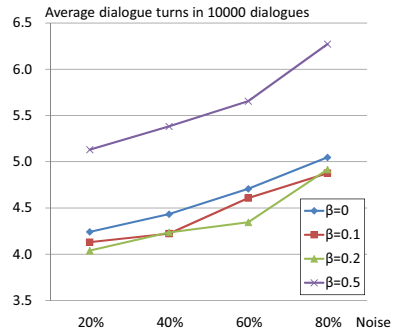

Figure 5: The effect of forgetting factor β as regards the average dialogue turns of 10000 dialogues.

goal g can be estimated from the user state i by using the IDG. As a result, Eq. (6) is approximated as,

$$(6) \cong \sum_{g} \underbrace{p(i'|g,\hat{k})}_{\textbf{simulator}} \underbrace{p(g|i)}_{\textbf{goal model}} \qquad (7)$$

Here, **simulator** is the user simulator that is defined in Section 3.2, and **goal model** is a goal estimation model that can be calculated from an IDG. Our user simulator does not perform in accordance with $p(s^{t+1} = i'|g, a^t = \hat{k})$ exactly, but our model uses the track back of the user simulator that is defined in Section 3.2. The probability of a goal estimation model is defined as $p(g|i)$, which expresses possible goals given a user state $s^t = i$.

4 Evaluations

We evaluate our statistical dialogue management approach, which uses the IDG. These are experimental evaluations with the user simulator that follows Section 3.2. In the experiment, we used an IDG that had 957 states including 667 goals.

4.1 Evaluation of average reward

We evaluated dialogue managers in terms of the average reward for 10000 dialogues between the user simulator and the obtained dialogue manager. We simulated uniformly distributed noises that are defined on Eq. (5) for observation. We tried six grids that suppose a uniform distribution given by Eq. (5). The parameters ($\eta = \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$) were sampled in the Q-learning of the POMDP. We used parameters $\gamma = 0.8$ and $\epsilon = 0.2$. The belief update defined in Section 3.4 was used for the dialogue evaluation. For comparison, we prepared an MDP based dialogue manager that learned from observations without any noise. The average rewards result is shown in **Figure 3**. In this experimental result, the POMDP dialogue manager performed better than the MDP based dialogue manager (MDP) in noisy cases. The effects of forgetting factor β in terms of average reward and average dialogue turn are shown in **Figure 4** and **Figure 5**. In this graph, the proposed POMDP framework, which includes state transition probabilities, works best at the point $\beta = 0.2$. These figures show that the approach depended on the forgetting factor and the robust setting of β is left to future work.

Figure 7 in Appendix shows an example of dialogue between the user simulator and the dialogue manager. This example was obtained with $\eta = 0.8, \beta = 0.2$.

5 Conclusion and discussion

We have proposed a dialogue management framework that uses a directed IDG. The IDG is handcrafted during the construction of the conventional rule-based dialogue system, and our approach can easily adapt rule-based systems to a statistical dialogue management framework. The proposed framework does not require annotated dialogue data in the initial deployment that are essential for the typical statistical dialogue management framework, and this enables rapid and easy adaptation. The proposed scheme is developed purely based on a probability process, and the framework can be extended to use annotated data to estimate model parameters, which will be future work. Ongoing work includes evaluation with real user or realistic user simulator that is constructed from dialogue logs.

References

Dan Bohus and Alexander I. Rudnicky. 2003. Ravenclaw: Dialog management using hierarchical task decomposition and an expectation agenda. In *Proc. of EUROSPEECH*.

Blai Bonet. 2002. An e-optimal grid-based algorithm for partially observable Markov decision processes. In *Proc. of ICML*, pages 51–58.

Deborah A. Dahl, Madeleine Bates, Michael Brown, William Fisher, Kate Hunicke-Smith, David Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the ATIS task: The ATIS-3 corpus. In *Proc. of the workshop on Human Language Technology*, pages 43–48.

Lucie Daubigney, Matthieu Geist, and Olivier Pietquin. 2012. Off-policy learning in large-scale POMDP-based dialogue systems. In *IEEE-ICASSP*, pages 4989–4992.

M. Gašić, F. Jurčíček, S. Keizer, F. Mairesse, B. Thomson, K. Yu, and S. Young. 2010. Gaussian processes for fast policy optimisation of POMDP-based dialogue managers. In *Proc. of SIGDIAL*, pages 201–204.

F. Jurcicek, B. Thomson, S. Keizer, F. Mairesse, M. Gasic, K. Yu, and S. Young. 2010. Natural belief-critic: a reinforcement algorithm for parameter estimation in statistical spoken dialogue systems. In *Proc. of INTERSPEECH*.

Cheongjae Lee, Sangkeun Jung, Kyungduk Kim, Donghyeon Lee, and Gary Geunbae Lee. 2010. Recent approaches to dialogue management for spoken dialog systems. *Journal of Computer Science and Engineering*, 4(1):1–22.

Oliver Lemon, Xingkun Liu, Daniel Shapiro, and Carl Tollander. 2006. Hierarchical reinforcement learning of dialogue policies in a development environment for dialogue systems: Reall-dude. In *Proc. of the 10th Workshop on the Semantics and Pragmatics of Dialogue*, pages 185–186.

Esther Levin, Roberto Pieraccini, and Wieland Eckert. 2000. A stochastic model of human-machine interaction for learning dialog strategies. *Speech and Audio Processing, IEEE Transactions on*, 8(1):11–23.

William Li. 2012. Understanding user state and preferences for robust spoken dialog systems and location-aware assistive technology. Master's thesis, Massachusetts Institute of Technology.

William S Lovejoy. 1991. Computationally feasible bounds for partially observed Markov decision processes. *Operations research*, 39(1):162–175.

Teruhisa Misu and Hideki Kashioka. 2012. Simultaneous feature selection and parameter optimization for training of dialog policy by reinforcement learning. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 1–6. IEEE.

George E Monahan. 1982. State of the art? a survey of partially observable Markov decision processes: Theory, models, and algorithms. *Management Science*, 28(1):1–16.

Sébastien Paquet, Ludovic Tobin, and Brahim Chaib-draa. 2005. An online POMDP algorithm for complex multiagent environments. In *Proc. of AAMAS*, pages 970–977.

Joelle Pineau, Geoff Gordon, Sebastian Thrun, et al. 2003. Point-based value iteration: An anytime algorithm for POMDPs. In *Proc. of IJCAI*, volume 18, pages 1025–1032. LAWRENCE ERLBAUM ASSOCIATES LTD.

ShaoWei Png and Joelle Pineau. 2011. Bayesian reinforcement learning for POMDP-based dialogue systems. In *Proc. of IEEE-ICASSP*, pages 2156–2159. IEEE.

Guy Shani, Joelle Pineau, and Robert Kaplow. 2013. A survey of point-based POMDP solvers. *Autonomous Agents and Multi-Agent Systems*, 27(1):1–51.

Matthijs TJ Spaan and Nikos Vlassis. 2005. Perseus: Randomized point-based value iteration for POMDPs. *Journal of artificial intelligence research*, 24(1):195–220.

Sebastian Varges, Giuseppe Riccardi, Silvia Quarteroni, and Alexei V Ivanov. 2011. POMDP concept policies and task structures for hybrid dialog management. In *Proc. of IEEE-ICASSP*, pages 5592–5595.

Christopher JCH Watkins and Peter Dayan. 1992. Q-learning. *Machine learning*, 8(3):279–292.

Jason D Williams and Steve Young. 2007. Partially observable Markov decision processes for spoken dialog systems. *Computer Speech & Language*, 21(2):393–422.

Jason D. Williams. 2008. The best of both worlds: Unifying conventional dialog systems and POMDPs. In *Proc. of INTERSPEECH*.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for POMDP-based spoken dialogue management. *Computer Speech & Language*, 24(2):150–174.

A Dialogue examples

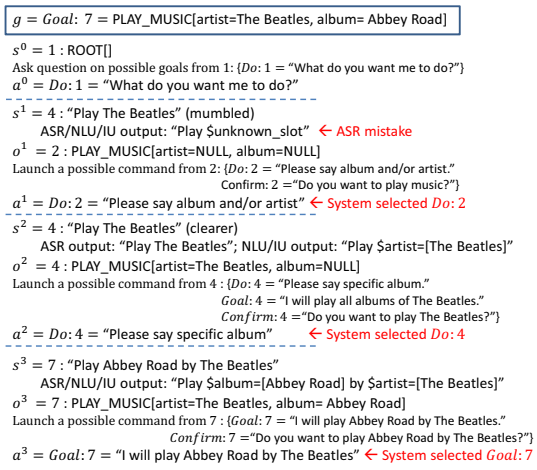

Figure 6: A dialogue example.

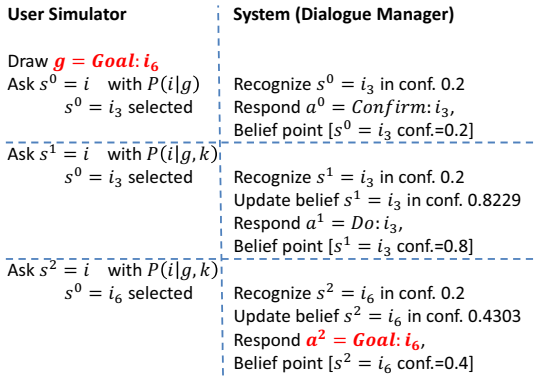

Figure 7: An example of the obtained dialogue between the user simulator and the our system.