Vol.2009-SLP-79 No.15
2009/12/22

対話データの統計量を用いた POMDP による対話制御

南 泰浩†, 森 啓†, 目黒 豊美†, 東中 竜一郎†,
堂坂 浩二†, 前田 英作†

本研究では, ユーザに対してエージェントが適切な行動を決定する対話制御(方策)を人対人の行動系列を記録したデータから自動的に学習する手法を提案する．これを実現するため本稿では次の二つの手法を用いる．(1) エージェント設計者が実現したいデータ中の行動系列(目標行動系列)を選択し，このデータから DBN (Dynamic Bayesian Network)を学習し，POMDP(partially observable Markov decision process)に変換する．この POMDP の状態遷移確率，出力確率，報酬から方策を学習する．(2) 自然な対話を実現するため，学習データの統計的性質に基づく対話制御のための状態，報酬を(1)の DBN と POMDP に付加する．これにより，目標行動系列を達成しかつデータの統計的特徴を持つ行動を生成する対話制御を実現する．シミュレーション実験により，本手法の有効性を確認した．

Dialogue Control by POMDP using Dialogue Data Statistics

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

We propose a method that generates appropriate agent dialogue control for users by training a large amount of human to human dialogue data. We offer two technical points to resolve this issue. One is the automatic acquisition of POMDPs' (partially observable Markov decision process's) state transition probabilities, output probabilities and rewards through DBNs (Dynamic Bayesian Networks) 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.

1. はじめに

実際の物理世界(環境)の中で活動するロボットの行動制御手法として，部分観測マルコフモデル(POMDP: partially observable Markov decision process)の研究が行われている[1]．POMDP は，環境を隠れ状態としてモデル化し，この状態に依存してユーザの取る行動を確率的に出力する．このため，HMM と同様，観測値が持つ不確定性をモデル化できる [2]．しかし，扱う環境が複雑になるにつれ，状態数の増加に伴う莫大な計算量を必要とするため，応用の範囲が限られていた．近年，POMDP において，最適な行動制御(方策)に必要な計算量を近似計算により削減する様々な手法が提案され[3][4], 大規模な状態数を対象とする対話処理制御の研究が盛んになってきた[6-9]．さらに，音声だけでなく複数のモダリティ情報の入出力を可能としたコミュニケーションロボットの行動制御にも適用が拡大されている[10]．これらの研究での対象は天気予報案内やデジタル加入者線(DSL)のトラブルシューティングといったタスク達成型の対話である．タスク達成型の対話は，POMDP の確率や報酬を通常人手あるいは半自動で設定する．

これに対して，本報告では，これらの確率や報酬を大量のデータから自動的に学習することを考える．この問題に対してPOMDPの構造を統計モデル(DBN: Dynamic Bayesian Network)で近似し強化学習により方策を決定する手法が提案されている[11]．しかし，これは人工的なタスクに対する行動制御を対象としており，自然な対話データに対する行動制御を対象としたものではない．また，方策の計算にもPOMDPではなくMDPを用いている．

そこで，本報告では，人とエージェント(人間とコンピュータとのインタラクションを司るもの，ここでは実体の存在を問わない)との自然な対話を実現するため，人と人との行動系列を対象とした大量のデータからDBNを構築し, 強化学習によりエージェントの方策を決定する手法を提案する[12]．ここでは，この大量のデータからシステム設計者が実現したい行動系列の集合(目標行動系列)を選択し，エージェントの挙動を目的行動系列に近づける方策を学習することを考える．しかし，選択された目的行動系列のみから方策を学習すると実行される行動系列が必ずしもデータ全体の持つ統計的性質に従う自然な対話になるとは限らない．そこで，さらに，この目的行動系列を実現しつつも，学習データの持つ統計的な性質を反映する方策を学習する手法を提案する．これにより，エージェントがデータの統計的性質に従う自然な行動系列を生成する方策が実現できる．本報告では，以上の手法をシミュレーション実験において評価を行う．

†日本電信電話株式会社
†Nippon Telegraph and Telephone Corporation

Vol.2009-SLP-79 No.15
2009/12/22

2. POMDPの概念

ここでは，POMDP の概念について述べる．POMDP は 集合のセット(S,O,A,T,Z,R, γ，b_0)で表現される．ここで S は状態の集合であり，s はSの各要素($s \in$S)である．O は観測値(観測されたユーザの行動)の集合であり o は Oの要素($o \in$O)である．A はアクション(エージェントの行動)のセットであり，a はA の要素($a \in$A)である．T はアクションa によって状態がs からs'へ変化するときの遷移確率$\mathrm{P}(s'|s,a)$の集合， Z は状態s'でアクションa によって観測値o'が観測されるときの出力確率$\mathrm{P}(o'|s',a)$の集合．R は状態s でアクションa を実行したときの報酬$r(s,a)$ の集合である． ここで用いる POMDP を構成する変数の依存関係を図 1 に示す．図は HMM と類似しているが，a とR でシステムの挙動を制御するところが HMM とは異なる．実線の円は確率変数を示し，点線の円は隠れ変数を表す．ひし形は固定値を表し，四角はエージェント側の選択する固定の変数を表す．ここで γ と b_0を説明する前に，状態の確率分布の更新式について述べる． POMDP では，HMM と同様に状態 s が直接観測できない．そのため，扱えるのはその分布だけである．この状態の分布は HMM と同じように一つ前の時刻の分布から計算できる．いま，一つ前の時刻の分布$b_{t-1}(s)$がわかっているものとする．この分布と遷移確率および出力確率から$b_t(s)$は以下の漸化式で記述される．

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

ここでη は全体の和を 1 にするための正規化項を表す．bの初期値を b_0と置くと，$b_t(s')$ は式(1)により再帰的に計算できる．この分布を使うと，時刻 t で将来獲得する平均割引報酬は以下のように定義できる．

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

正定数$\gamma(<1)$により未来の報酬の寄与は小さくなる．POMDP では，式(2)を最大にするアクション系列aを求めることにより行動制御を実現する．強化学習を用いると，式(2)を直接計算せずに$b_t(s)$の分布からa を返す関数を求めることができる[1][2]．これを方策と呼ぶ．

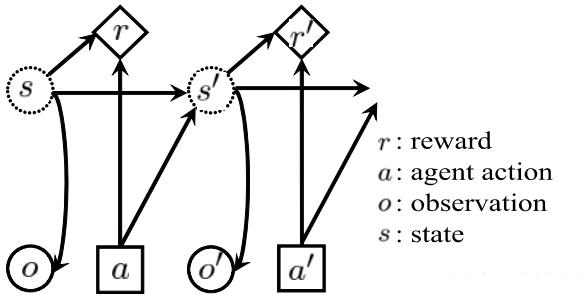

図 1　対話制御に用いる POMDP

3. 大量データからの対話制御手法

図 2 に，行動系列データから POMDP の方策を導出する提案手法のフローを示す．図 3 に図 1 の POMDP のパラメータを学習するための DBN を示す．この DBN を行動系列データから EM アルゴリズムにより学習する．学習された DBN の状態遷移確率と出力確率を POMDP の状態遷移確率，出力確率，報酬に変換する．この POMDP を用いて強化学習により方策(最適な行動制御)を作成する．ここでは，この方策の学習に次の二つの目的を設定する．一つは設計者が実現したい行動系列の集合(目標行動系列)を大量の学習データから抽出し，その行動を実現する方策を学習することであり．もう一つは，この目的行動系列を実現しつつも，学習データの持つ統計的な性質を反映する方策を学習することである．この二の目的設定は，例えば，英語を学習するときに，挨拶のように典型的な一連の決まり切った系列を学習するとともに，相手の行動に対する適切な応答も同時に学習することに似ている．以上の二つの学習を実現するため以下に示す二つの手法を用いた．

(1) 目標行動系列のデータを含む学習データからEMアルゴリズムによりDBNの遷移確率，出力確率を自動的に学習し，これらをPOMDPの遷移確率，出力確率，報酬へ変換する．このとき，目標行動系列を生成する報酬の設定をDBNの出力確率から統計的に計算する．

(2) (1)のDBNに対してアクションと1対1に対応する隠れ状態を付加し，その隠れ状態の出現確率にアクションの出現確率を反映する．次に，アクションの出現確率がPOMDPの平均報酬になるように報酬を決定する．

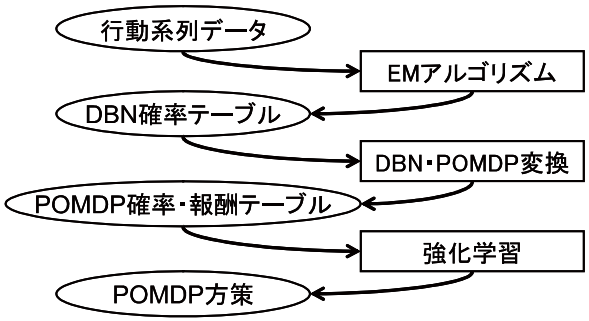

図 2　本手法の基本フロー

4.　目標行動系列実現のためのDBNとPOMDPの学習

POMDP では，式(1)で示すように，状態遷移確率と観測値の出力確率の学習が必要である．タスク達成型の対話において，これらのパラメータは既知である，あるいは，あらかじめ容易に計算できると仮定している． 本報告では，これらのパラメータを自動的にデータから学習する．ここでは，表 1 のような対話のシミュレーションをデータとして用いる．エージェント，ユーザとも握手，挨拶，笑い，移動，話す，頷き，首ふり，無行動の 8 個の行動を実行する．対話では，エージェントとユーザは交互に行動を実施する．対話の後で，その対話が目標行動系列であったかどうかをその系列を見ながら人が評価すると仮定する．この評価結果に基づいて，個々の対話にスコア d をつける．表 1 の一番最後の列に d を記入している．このデータから図 1 とほぼ同じ構造である図 3 の DBN を使って，状態遷移確率，観測値の出力確率を EM アルゴリズムにより学習する．ただし，図 3 では図 1 の r の代わりに d を確率変数として扱う．

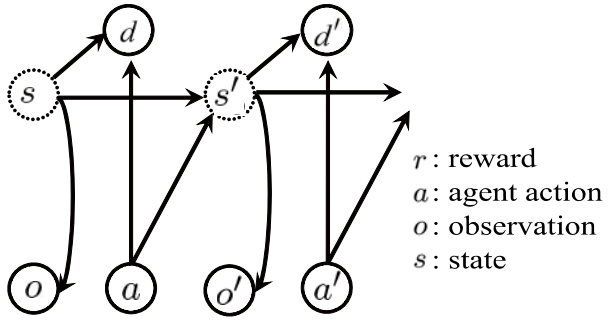

図 3 POMDP に対応する DBN

表 1　対話データの例

観測値 o	アクション　a	スコア　d
無行動	話す	1
頷き	頷き	1
握手	握手	1
挨拶	挨拶	1
笑い	話す	1
首ふり	話す	1
挨拶	挨拶	1
握手	握手	1
無行動	握手	1
挨拶	無行動	1

DBN の学習後，DBN を POMDP に変換する．図 1 と図 3 を見比べて遷移確率と観測値の出力確率はほぼ同じ構造なのでそのまま利用する．しかし，POMDP では報酬という固定の値を設定しているのに対し，DBN では，これに代わってスコアという確率変数を設定している．そこで， DBN の d の値とその確率値から POMDP の報酬を計算する．ここでの POMDP の報酬の目的は目標行動系列を実現することである．スコア d は，目標行動系列に対して 1 を割り当てたものであるから，その平均は妥当な報酬となる．そこで報酬として，次式を定義する．

$$r_1(s,a) = \sum_{d=0}^{1} d \cdot \mathrm{P}(d \mid s,a) \qquad (3)$$

この報酬と遷移確率，出力確率を用いて POMDP の強化学習を行うことにより，目標行動系列に対する方策を求めることができる．

5.　アクションの出現確率を反映する行動制御

4 章で求めた POMDP を使って実験を行うと，確かに目標行動系列を実現できることが分かった．しかし，ユーザは必ずしも常時，目標行動系列の実現を望むとは限らないが，その場合でもエージェントは目標行動系列実現に向けて行動することも分かった．このような場合でも自然な対話を実現するために，エージェントが生成するアクションの確率が学習データ中のエージェントのアクション出現確率に従うように，エージェントの行動制御を行う．これを実現する機構としてエージェントのアクションの出現確率を推定する状態 s_a を図 3 の POMDP に導入する(図 4 参照). この POMDP

では，従来までの状態を状態 s_o と呼ぶことにする．これにより，状態 s をユーザ・システム共通の隠れ状態 s_o とエージェントのアクション生成のための隠れ状態 s_a との組 $s=(s_o,s_a)$ と再定義する．

ここで，図 4 の関係に示す各変数の依存関係から，状態遷移確率を求めると以下の式になる．

$$P(s'|s,a)\approx P(s'_a|s'_o,s_a)P(s'_o|s_o,a) \tag{4}$$

図 4 では状態の遷移がアクションに依存するので，出力確率はアクションに依存しないと仮定している．図 4 の関係図から出力確率は以下の式となる．

$$P(o'|s',a)\approx P(o'|s'_o) \tag{5}$$

これらの式を用いると式(1)は以下のようになる．

$$\begin{aligned} b_t(s') &= b_t(s'_o,s'_a)=\eta\cdot P(o'|s',a)\sum_s P(s'|s,a)b_{t-1}(s) \\ &= \eta P(o'|s'_o)\sum_s P(s'_a|s'_o,s_a)P(s'_o|s_o,a)b_t(s_o,s_a) \end{aligned} \tag{6}$$

また，時刻 t 以降に獲得できる平均報酬 V_t は， 次式で表される．

$$V_t=\sum_{\tau=0}^{\infty}\gamma^{\tau}\sum_s b_{\tau+t}((s_o,s_a))r((s_o,s_a),a_{\tau+t}) \tag{7}$$

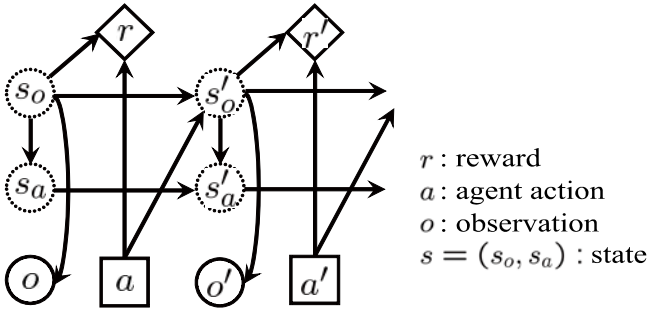

図 4　アクションの出現確率を反映するための POMDP

次に，データ中に出現する統計情報に従ってアクションを選択する方策のための報酬の設定手法を説明する．まず $b_t(s)$ の定義から次式が得られる．

$$b_t(s_a)=P(s_a|o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t)=\sum_{s_o}b_t(s) \tag{8}$$

図 4 の POMDP を学習するために，図 5 に示す DBN を構成する．ここで，アクション a の出現確率を s_a に反映させることを考える．これを実現するため，図 5 では，s_a と a を接続し，1 対 1 に対応させている．すなわち，$a=s_a$ の時に限り，$P(a|s_a)=1$ とする．これにより，$a_t=s_a$ の時に以下の式を得る．

$$\begin{aligned} &P(a_t|o_1,a_1,\ldots,a_{t-1},o_t) \\ &=\sum_{s'_a}P(a_t|s'_a)P(s'_a|o_1,a_1,\ldots,a_{t-1},o_t) \\ &=P(s_a|o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t) \end{aligned} \tag{9}$$

これは，過去に $o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t$ が観測された時のアクション a_t の予測確率を表している．この接続および確率の導入により，アクション a の出現確率を s_a に反映できる．本報告では，この出現確率を最大化する行動を選択するように方策を決定する．すなわち，式(9)を最大化するように，報酬を設定する．これを実現するため，ここでは，$s_a=a$ の時に 1 となり，他の時には 0 となる報酬 $r_2(s=(*,s_a),a)$ を定めた．ここで，*は任意の s_o を指す．これにより式(7)のある時刻 $\tau+t$ の平均報酬は $b_{\tau+t}(s_a)$ となる．これは a の $\tau+t$ における出現確率となっている．この報酬を設定することで，アクションの出現確率が最大となるアクションを選択する方策を作成することができる．

最後に，4 章で述べた報酬とここでの報酬の両方を考慮し，目的行動系列とデータの統計的性質のどちらにも従うエージェントの行動が実現できるように報酬を決定する．ここで，4 章で述べた式(3)の報酬を $r_1((s_o,*),a)$ と書き換える．これは，目標行動の報酬が s_o とアクション a にしか依存しないことを示している．以上の二つの報酬を用いて，式(2)の r を r_1+r_2 に置き換える．この式を以下に示す．

$$r(s,a)=r_1((s_o,*),a)+r_2((*,s_a),a), \tag{10}$$

これにより最終的な目的関数 V_t を得る．この式は，目標行動系列を実現する報酬とアクションの出現確率を反映する報酬との和になっている．すなわち，得られる方策は目標行動系列の報酬，アクションの出現確率の最大化による報酬の最大化の両方

を同時に実現するものとなる．図 5 の DBN による学習結果の統計量と，式(10)を用いて POMDP での強化学習を行うことにより最終的な方策が得られる．

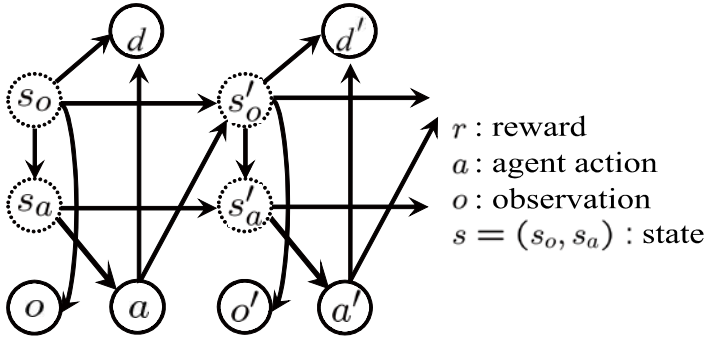

図 5　アクションの出現確率を反映する POMDP に対する統計量を学習する DBN

6. シミュレーション対話実験

1 対 1 の対話を想定し，大量の行動系列データからエージェントの対話制御を学習するシミュレーション実験を行った．

ここで用いる全ての変数は複数のシンボルを取る離散変数である．4 章と同様にアクションと観測値のシンボルとして，握手，挨拶，笑い，移動，話す，頷き，首ふり，無行動の 8 種類を用意した．ここでは，ユーザの意図を観測できない隠れ状態として扱った．隠れ状態 S_o の数は 16 である．また，アクションに 1 対 1 に対応する隠れ状態 S_a の状態数をアクションの数と同じ 8 とした．目標行動系列としては 2 種類の系列を用意した．一つは，お互いに握手をし，お互いに挨拶をし，その後，笑いと話しと頷きを数回ランダムに相互に繰り返し，最後に挨拶をし合い，握手をし合うという系列である．もう一つは，片方が移動し，片方が無行動で，その後，挨拶をし合い，笑いと話しと頷きを数回ランダムに繰り返し，挨拶をし合い，最後に片方が無行動で，片方が移動するというものである．これらの目標行動系列は全体の学習データの数(10000)に対して 10 分の 1 とした．この行動系列が出現する場合には，各時刻にスコアとして 1 を割り当てた．残りのデータでは，ユーザ観測値とシステム行動の自然な行動として，握手－握手，挨拶－挨拶，笑 い－笑い，移動－移動，話す－話す，頷き－話す，首ふり－話す，無行動―無行動の対の出現確率が統計的に多くなるようにサンプルを作成した．このデータでは，各対話にスコアとして 0 を割り当てた．全てのデータの長さを 10 とした．目標行動系列では，途中のランダムの回数によってこの長さに満たないサンプルが生じるが，この部分には，自然なふるまいをシミュレートしている残りのデータからサンプルを付加した．この学習データを使って DBN を学習した後，式(10)の報酬を計算し POMDP による強化学習を行った．

比較手法は，POMDP において，目標行動系列だけに報酬を設定する方法とした．すなわち，報酬として以下の式を使った．

$$r(s,a) = r_1((s_o,*),a) \qquad (11)$$

評価 には，2000 サンプルのデータを用い，目標行動系列の学習データを生成した手法，および，その他の系列の学習データを生成した手法に従ってユーザの観測値のシミュレーションデータを生成した．比較手法，提案手法のどちらも，目標行動系列 200 サンプルに対して全て正しく(設定上誤りのない．すなわち，ランダムと仮定した変数は取り得る値であれば正しいとした)行動を生成した．これにより，どちらの手法も目標行動系列に対しては正しい行動を生成することが確認された．2000 サンプルの学習データに対するユーザ観測値とエージェントアクションの対の出現確率の多いものを表 2 に示した．表 2 の“学習データ”という項にその出現確率を示した．この頻度に近づくほど 自然な対話であるとみなすことができる(このシミュレーションでは，観測値とアクションの共起確率だけを対象とした).

表 2　POMD から自動生成された観測値アクションペア対の共起確率

観測値・アクション対	学習データ	比較手法	提案手法
握手−握手	0.09	0.13	0.13
挨拶−挨拶	0.10	0.11	0.13
笑い−笑い	0.08	0.00	0.02
動き−動き	0.08	0.00	0.002
話す−話す	0.04	0.00	0.00
頷き−話す	0.09	0.00	0.08
首振り−話す	0.09	0.00	0.05
無行動−無行動	0.10	0.00	0.05

この表から，比較手法では，学習された 目標行動系列に含まれる観測値・アクシ

ョン対を高頻度で生成しているのがわかる．しかし，学習データの観測値・アクション対の統計パターンとは大きな違いがある．これは，目標行動系列に対してだけ報酬を与える比較手法は，アクションを決定する際に，この系列のみを生成しようとするからである．これに対して，アクションの出現確率を報酬に導入する提案手法は，9倍ある目標行動系列以外の学習データの統計量に近づいており，自然な対話を実現している．

7. 考察

課題としては次のことがあげられる．図 5 の DBN と図４の POMDP はアクションに対する出力確率の有無が異なる．このため，今回のモデルでは二つのモデルは全く同じ統計構造とはなっていない．このことから，今回の定式化では若干の近似が含まれていることがわかる．この近似がどの程度，結果に影響を与えているのかは今後の課題である．

また，今回は，スコアとして目標行動系列が含まれるデータの全対話に 1 を付与したが，この与え方には他にも様々なものが考えられる．例えば，目標行動系列の区間のみに 1 を付与する方法，目標行動系列の最後の対話に 1 を付与する方法，などである．現状どのような報酬の付与方法が適切なのかに関してはまだ分かっていない．今後この点に関して研究を進めていく必要がある．

8. おわりに

Williams らの研究[6]を契機としてタスク達成型の対話制御に POMDP を持ち込んだ研究が注目されつつある．本報告では，人と人との大量のデータから目標行動系列を実現しつつ，自然な行動を行うエージェントの最適な行動制御すなわち方策を学習する手法を提案した．この実現のため次の二つの手法を提案する．(1) 実現したい行動系列の集合(目標行動系列)にラベルを付け, DBN を学習し, POMDP の状態遷移確率，出力確率，報酬に変換し，強化学習により方策を学習する．(2) 自然な対話を実現するため，学習データ中のエージェントのアクション出現確率から得られる報酬を POMDP の強化学習に利用する．これにより，目的行動系列とデータの統計的性質のどちらにも従うエージェントの行動が実現できる．二つのアルゴリズムを実装し，シミュレーション実験により，本手法の有効性を確認した．

参考文献

[1] Thrun, S., Burgard, W., Fox, D.: Probabilistic Robotics, The MIT Press, (2005).
[2] Smallwood, R.D. and Sondik, E.J.: The Optimal Control Of Partially Observable Markov Decision Processes Over A Finite Horizon, Operations Research, 21, 1071-1088, (1973).
[3] Pineau, J., Gordon, G., and Thrun, S.: Point-Based Value Iteration: An Anytime Algorithm for POMDPs, IJCAI, pp.1025-1032, (2003) .
[4] Poupart, P.: Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes. Ph.D. dissertation, University of Toronto, (2005).
[5] Roy, N., Pineau, J., and Thrun, S.: Spoken Dialogue Management Using Probabilistic Reasoning, ACL 2000, pp. 93-100, (2000).
[6] Williams, J., Poupart, P., and Young, S.: Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management,,SIGDial Workshop on Discourse and Dialogue, pp.25-34, (2005).
[7] Williams, J.: Using Particle Filters to Track Dialogue State, ASRU 2007, pp.502-507, (2007) .
[8] Kim, K., Lee, C., Jung, S., and Lee, G. G.: A Frame-Based Probabilistic Framework for Spoken Dialog Management Using Dialog Examples, SIGdial Workshop on Discourse and Dialogue, pp.120-127, (2008) .
[9] Williams, J., Poupart, P., and Young, S.: Factored Partially Observable Markov Decision Processes For Dialogue Management, IJCAI Workshop on K&R in Practical Dialogue Systems, pp.76-82, (2005).
[10] Schmidt-Rohr, S. R., Jäkel, R., Lösch, M., and Dillmann, R.: Compiling POMDP Models For A Multimodal Service Robot From Background Knowledge, European Robotics Symposium 2008, 44, pp.53-62, (2008).
[11] Fujita, H.: Learning And Decision-Planning In Partially Observable Environments, Ph.D. dissertation, Nara Institute of Science and Technology, (2007) .
[12] 南 泰浩, 目黒 豊美, 東中 竜一郎, 森 啓, 堂坂 浩二, 前田英作: 統計的モデルを用いた POMDP による対話制御, 2009 年秋季研究発表会, 3-1-6, (2009).