3C2-OS19-3

アクション継続長制御を用いたPOMDPによる対話制御

POMDP Dialogue Control using Action Duration

南泰 浩[*1] Yasuhiro Minami　目黒 豊美[*1] Toyomi Meguro　東中 竜一郎[*2] Ryuichiro Higashinaka　堂坂 浩二[*1] Kohji Dohsaka　前田 英作[*1] Eisaku Maeda

[*1]ＮＴＴコミュニケーション科学基礎研究所 NTT Communication Science Laboratories, NTT Corporation　[*2]ＮＴＴサイバースペース研究所 NTT Cyber Space Laboratories, NTT Corporation

This paper proposes a dialogue control method using action durations. We introduce duration control to our POMDP action generation process. The experimental results show that the proposed method can generate action sequences whose probability is similar to the training data and increase the entropy of the actions. This confirms that our method generates appropriate action sequences.

1. はじめに

我々の目標は，対話データから対話 (行動) 制御を自動的に獲得することである．行動制御を自動的に学習する手法として，POMDP (Partially Observable Markov Decision Process)[Sutton98, Russell03] が近年注目を集めている．POMDP は，ある状態下でシステムが行動するアクションに対して報酬を定義し，将来，最も多くの平均報酬が獲得できるアクションを選択するモデルである．POMDP を利用する対話システムとしてタスク達成型の様々な対話システムが提案されている [Williams05, Kim08, Williams07, Schmidt-Rohr2008]．我々の目的は，ユーザが対話そのものを楽しむようなタスク達成を目的としない対話制御の実現である．このような対話では，予めシステムが行うべき行動を完全に把握することはできないので，人対人あるいは人対システムの対話データから対話制御を学習しなければならない．そこで，我々は，DBN(Dynamic Bayesian Networks) を対話データから学習し，人間の評価に基づく報酬を決定し，それらを用いた POMDP による対話制御手法を提案した [Minami09, Meguro10]．この手法では，対話の自然性を考慮し，予測確率の高いアクションを選択する報酬も設定している．我々は，提案手法を用いて，Trigram による対話処理 [Hori09] を包含する手法が実現できることをも示した [Minami10]．この Trigram による対話処理を，聞き役対話 (実際の対話例を表 1 に示す) に応用し，対話行為タイプのみの制御実験を行った．その結果，表 2 のように予測確率の大きい特定の対話行為タイプのみを生成する現象が頻繁に起こることが分かった．本稿では，この問題を解決するため，ある特定のアクションが長い間続かないように継続長の確率分布に従ってアクションを生成する手法を提案する．

2. 人の評価とアクション確率に基づくPOMDPによる対話制御

我々の提案してきた対話手法は，人の評価が高くかつ，生起確率の高いアクションを生成する POMDP の方策を生成する．この手法を簡単に説明する [Minami10].

連絡先: 南泰浩，NTT コミュニケーション科学基礎研究所，〒 619-0237，京都府相楽郡精華町光台 2-4，0774-93-5323，0774-93-5245 ，minami.yasuhiro@lab.ntt.co.jp

表 1: 典型的な聞き役対話例．O は話し役，A は聞き役

発話	対話行為タイプ
O: こんばんは.	挨拶
「食事」でお願いします.	挨拶
A: はい，よろしくお願いします.	挨拶
O: 今日の夕飯はカレーでした.	自己開示 (sub: 事実)
Ｂさんはカレーは好きですか？	質問 (sub: 評価)
A: 好きです.	共感・同意
O: おお，お好きですか.	繰り返し
私も好きなんです.	共感・同意
A: 外食が主ですか？	質問 (sub: 習慣)

2.1 POMDP の構造

ここで用いる DBN および POMDP の構造を図 1 と図 2 に示す．状態として $s = (s_o, s_a)$ を導入した．s_o は通常の POMDP の状態と同じ働きをする．s_a は，アクション a の予測確率を計算し，予測確率を最大化するアクションを選択するために導入した．これにより状態遷移確率は以下のようになる．

$$\Pr(s'|s,a) \approx \Pr(s'_a|s_a,s'_o)\Pr(s'_o|a,s_o) \quad (1)$$

$$\Pr(o'|s',a) \approx \Pr(o'|s'_o) \quad (2)$$

表 2: 以前提案した POMDP で生成された対話行為例，“EPS” は何もしない対話行為

ユーザ観測値 O	システムアクション A
挨拶	挨拶
⋮	⋮
挨拶	EPS
質問 (sub:経験)	EPS
確認	EPS

これらの式を使うと 状態の確率分布 $b_t(s')$ は，以下の漸化式で求められる．

$$b_t(s') = b_t(s'_o, s'_a) = \eta \Pr(o'|s'_o) \sum_s \Pr(s'_a|s_a, s'_o) \Pr(s'_o|s_o, a) \Pr(a|s_a) b_t(s_o, s_a) \quad (3)$$

ここで，元々の文献 [Minami10] では，$\Pr(a|s_a)$ の項を利用せず POMDP を構成していた．ここで，この項を用いるのは，アクションが決まった後に，その影響を状態に反映させる必要があるためである．

2.2 方策の生成

通常の POMDP と我々が提案している手法は以下の 3 点が異なる．

1. 目的対話のための POMDP の確率と報酬を自動的に獲得
2. POMDP へのアクション予測確率の導入
3. POMDP 報酬の統合方法，および，方策の決定

1. の報酬の計算のために，d という変数を用いた．また 3. を行うため報酬 r_1 と r_2 という変数を設定した．r_1 は s_o と a の関数で，人の評価に基づく報酬である．r_2 は s_a と a の関数で，確率に基づく対話制御のための報酬である．1. を以下のように実行する (タスク達成型ではないので，報酬は，人の評価データから学習する必要がある)．対話終了後，ユーザあるいは第三者の評価者に，アンケートに対する点数をつけてもらう．アンケート例は，例えば，”対話に満足しましたか ”というものである．この点数を変数 d の値とする．表 3 に，点数付けされたデータの例を示す．評価値がそれぞれのターンに付加されている．しかし，このように全てのターンに評価を与えることは多くの労力を必要とするため，本論文では，対話全体に一つの評価を与え，それを各ターンに均等に分配した．

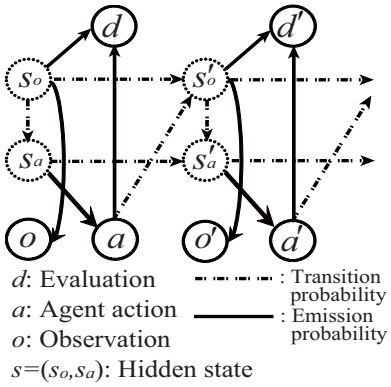

図 1: 対話のための DBN

次に，報酬を決定し，対象の対話データの POMDP を作るために以下の処理を行う．

(a) DBN を学習する．その時 d も確率変数として扱う．

(b) DBN を POMDP に変換し，変数 d と d の確率から以下の式により POMDP の報酬を求める．

$$r_1(s_o, a) = \sum_d d \cdot \Pr(d|s_o, a) \quad (4)$$

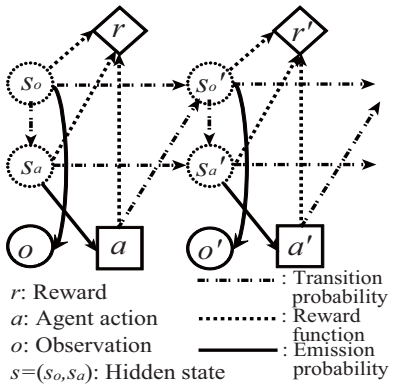

図 2: 対話のための POMDP

表 3: 対話データの例．O は，ユーザ発話の観測値 (対話行為)，A はシステム側のアクション (対話行為) を示す．(注) 評価の値は例であり実際のデータから抽出したものではない．

ユーザ観測値 O	システムアクション A	評価値 d
挨拶	EPS	3
挨拶	挨拶	3
自己開示 (sub:事実)	EPS	3
質問 (sub:評価)	共感・同意	5
繰り返し	EPS	3
共感・同意	質問 (sub:習慣)	5

次に 2. について述べる．$s = (s_o, s_a)$ を使った目的関数を以下に示す．

$$V_t^\pi = E^\pi[\sum_{\tau=0}^{\infty} \gamma^\tau \sum_s b_{\tau+t}(s_o, s_a) r((s_o, s_a), a_{\tau+t})] \quad (5)$$

π は方策を表す．方策は，状態の分布からアクションを返す関数である．もし $a = s_a$ なら，DBN(図 1) において，$\Pr(a|s_a) = 1$ と設定することにより，s_a が a と一対一に対応するようにする．これにより，$a_t = s_a$ が成り立つとき，以下の式を得る．

$$\Pr(a_t|o_1, a_1, \ldots, a_{t-1}, o_t) = \sum_a \Pr(a_t|s'_a) \Pr(s'_a|o_1, a_1, \ldots, a_{t-1}, o_t) \quad (6)$$

$$= \Pr(s_a|o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t) = \sum_{s_o} b_t(s) \quad (7)$$

ここでの目的は，$o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t$ が与えらているときに，a_t の確率が高くなるように a_t を選択することである．すなわち，式 (7) を大きくするようアクションを選ぶことである．これを実現するため (5) において，$b_{\tau+t}(s_a)$(ただし，$a_t = s_a$) の値が大きい場合に，POMDP が高い報酬を得らるようにする．これを実現するために，$r_2(s = (*, s_a), a) = 1$ という報酬を導入する．ここで $s_a = a$ であり，また，*は s_o が任意の値であることを示す．これ以外のときは $r_2(s = (*, s_a), a) = 0$ とする．

3. で実現する最終的な報酬すなわち (5) の r は以下のようになる．

$$r(s, a) = r_1((s_o, *), a) + w \cdot r_2((*, s_a), a); \quad (8)$$

これにより，新しい目的関数 V_t^π を得る．POMDP の方策はこの報酬を用いて Value Iteration により求められる．この定式化を用いると，POMDP は予測確率が高く，かつ，評価の高いアクションを選択するようになる．

2.3 Trigram を用いた対話処理への拡張

実験では，対話の確率モデルとして，Trigram のモデルを利用する．この手法について簡単に説明する．Trigram を我々のモデルで利用するに際し，以下の設定を行う．$|\mathrm{O}| = |S_o|$ とする．$o = s_o$ のとき，$\Pr(o|s_o) = 1$ と設定する．これにより s_o は o と 1 対 1 に対応するようになる．この設定により，$a = a_{t-1}$，$s_o = o_{t-1}$，$s'_o = o_t$，$a_t = s'_a$，$a_{t-1} = s_a$，$a_{t-1} = s_a$ のとき，次の式を得る．

$$\begin{aligned}\Pr(s'|s,a) &= \Pr(s'_o\,|s_o,a)\Pr(s'_a\,|s'_o\,,s_a)\\ &= \Pr(o_t\,|o_{t-1},a_{t-1})\Pr(a_t\,|a_{t-1}\,,o_t\,).\end{aligned} \tag{9}$$

これは，文献 [Hori09] での Trigram 確率に対応する．

2.4 対話行為誤りのモデル化

2.3 節では，$\Pr(o|s_o) = 1$ とし，s_o を固定した値として扱っているので，POMDP は現実には MDP となっていた．対話行為の認識誤りをモデル化するために，我々は，DBN から POMDP に変換する際に $\Pr(o|s_o)$ に対話行為の認識率を設定する．この設定により，この枠組みは認識誤りを含んだ対話を扱う POMDP となる．我々は，認識誤りのある状況で，この枠組みがうまく働くことを実験的に示した [Minami10].

3. 継続長制御

あらかじめ，対話データから各アクションが l 回続く継続確率 $\mathrm{P}_a(l)$ をヒストグラムとして計算する．対話の過程で，対話中での継続長の確率の履歴 $\bar{\mathrm{P}}_a(l)$ を計算する．ここでは，$\bar{\mathrm{P}}_a(l)$ が $p_a(l)$ に近くなるように，アクションを選択する．これを行うため，確率差分を以下のように定義する．

$$\Delta\mathrm{P}_a(l) = P_a(l) - \bar{\mathrm{P}}_a(l) \mid \Delta\mathrm{P}_a(l) < 0 \quad \text{ならば} \quad \Delta\mathrm{P}_a(l) = 0 \tag{10}$$

この $\Delta\mathrm{P}_a(l)$ が小さくなるようにアクションを生成する．ターン毎に，これを実現するため以下のような式を定義する．

$$\Delta\mathrm{P}'_a(n) = \frac{\Delta\mathrm{P}_a(n)}{\sum_{l=n}\Delta\mathrm{P}_a(l)} \tag{11}$$

$\Delta\mathrm{P}'_a(n)$ はアクションを n の長さで止めるための指標である．そして，$1 - \Delta\mathrm{P}'_a(n)$ を次のアクション a を実行するための指標と定義する．この指標を使って，アクション選択のルールを以下のように定義する．ここで，$n-1$ 回の連続するアクション a が実行されていて，POMDP はいま次のアクション a を実行しようとしていると仮定する．次に，確率変数 x(0<x<1.0) を生成する．もし x が $\Delta\mathrm{P}'_a(n-1)$ より大きければ，同じアクション a を実行する．そうでない場合は，$n-1$ 回連続したアクション a を終了させ，a ではないセカンドベストのアクション候補を選択する．

4. 対話シミュレーション実験

表 1 のような聞き役対話を用いて，提案する継続長制御の評価を行った．

4.1 実験設定

対話行為タイプは 32 種類である．これらのタイプを観測値とアクションのラベルとして利用する．実際の対話では，表 1 のように同じターンで複数の発話をしている．しかし，POMDP ではこのような複数の観測値，アクションを扱えない．これを避けるために，”EPS ”という対話行為を加えて，疑似的にターン単位で交互に対話行為を行うようにした．各対話に対して，二人のアノテータが主観的な満足度の評価を付けた (7 段階のリッカートスケールであり，d を 0 から 6 のレンジで設定した)．DBN と Trigram の学習するため 1259 の対話を利用した．各対話の平均ターンは 27.7 である (”EPS”対話行為タイプを含む)．この Trigram 確率を使ってユーザの行動をシミュレートした．このシミュレーションでは，ユーザとエージェントの対話行為タイプまでの生成を行った．各エポック (対話) の長さは，50 ターンである．提案手法を使って，1000 のシミュレーション対話を作成し評価した．また，対話行為タイプの認識誤りを一律 40%と仮定した．

4.2 評価尺度

従来の POMDP で生成した対話は同じ対話行為タイプを高い確率で生成する．この現象を定量的に評価するため，次の二つのメジャーを利用した．

$$Entropy = \sum_a -\mathrm{P}_{generated}(a)\log(\mathrm{P}_{generated}(a)) \tag{12}$$

$$Distance = \sum_a \left(\mathrm{P}_{generated}(a) - \mathrm{P}_{training}(a)\right)^2 \tag{13}$$

ここで，$\mathrm{P}_{generated}(a)$ は，生成されたデータ中でのアクションの確率であり，$\mathrm{P}_{training}(a)$ は，学習データ中のアクションの確率である．*Entropy* は，情報源の情報量を示している．もし，情報源が同じ対話行為タイプばかりを生成すればエントロピーは小さい．このため *Entropy* は，以上の問題を発見するよいメジャーとなる．また，生成されたデータと学習データとの間の *Distance* もよいメジャーである．これは，同じ対話行為タイプばかり生成すれば，生成されたデータと学習データの *Distance* は大きくなるからである．元々のパフォーマンスを評価するために，次の二つの評価も利用する．一つは，生成されたアクションの平均の Trigram 確率であり，以下の式で定義する

$$Trigram = \frac{1}{N}\sum_i^N \frac{\sum_t \Pr(a_{t+1}|a_t, o_{t+1})}{L_i}, \tag{14}$$

ここで，N は対話の数であり,L_i は，各対話の長さである．$\Pr(a_{t+1}|a_t, o_{t+1})$ は，a_t, o_{t+1} が与えられている時の a_{t+1} の学習データの Trigram 確率である．この尺度は，生成されたデータがどれだけ，学習データの Trigram 確率を大きくするかを示すものである (確率に基づく対話制御がどれだけ良いかを示す評価)．別の方法として，平均推定ユーザ評価を次のように定義する．

$$Satisfaction = \frac{1}{N}\sum_i^N \frac{\sum_t \bar{d}(a_t^i, o_t^i)}{L_i} \tag{15}$$

ここで，$\bar{d}(a, o)$ は観測値とアクションのペアに対するユーザ評価値である．これは，POMDP に基づく対話制御がどれだ

け推定されたユーザの評価を向上させるかを示す．ユーザの評価は，対話の履歴に影響を受けるが，ここでは，一番最後のユーザの観測値に最も影響を受けると仮定し，この値だけを用いた．

4.3 実験結果

w=15, w=10, w=5, w=0 と 4 つの条件で評価を行った．同じ環境下で，継続長制御を行うものと行わないものとを比較した．w=0 は POMDP の方策がユーザの満足度を強調していることを意味し，w=15 は POMDP の方策がアクションの確率を強調していることを意味する．

表 4: 継続長制御あるなしでの対話シミュレーション実験の結果

	継続長制御なし				継続長制御あり			
$Weight(w)$	15	10	5	0	15	10	5	0
$Entropy$	0.27	0.29	0.42	1.24	1.23	1.24	1.35	1.62
$Distance$	0.42	0.42	0.38	0.48	0.12	0.12	0.10	0.35
$Trigram$	0.30	0.30	0.30	0.01	0.31	0.31	0.30	0.01
$Satisfaction$	2.84	2.84	2.91	3.81	2.95	2.96	3.13	3.72

表 5: 提案手法によって生成された対話行為

ユーザ観測値 O	システムアクション A
挨拶	挨拶
質問 (sub:要求)	質問 (sub:事実)
自己開示 (sub:経験)	EPS
自己開示 (sub:事実)	自己開示 (sub:評価 (ポジティブ))
情報	確認
共感・同意	EPS

表 4 に評価結果を示す．*Entropy* に関しては，提案手法が従来手法をどの場合でも上回っている．*Distance* に関しても，同様である． 継続長制御を行わない場合でも，Trigram 確率が高い確率を示している．しかし，このほとんど全ての場合，表 2 で示したように特定の対話行為タイプを過剰に生成していた．

満足度に関しては，w=0 の場合を除いて，提案手法が良い結果を示している．表 5 に実際の生成された対話行為タイプを示す．これを見ると，表 2 と比べて，より自然な対話行為タイプを生成していることが分かる．w=0 以外では，継続長制御なしの場合，EPS が過剰に生成される．しかし，w=0 では，フィラーや感嘆が過剰に生成されていた．このことから本手法は EPS を導入したことによる EPS の過剰生成に特化した手法ではなく，そのほかの対話行為の過剰生成に対しても良好な結果を示すことが分かった．

5. 結論

提案してきた POMDP による対話制御において，確率の高いアクションばかりを生成してしまうという問題に対して，アクション継続長制御の導入する手法を提案した．この手法は，あらかじめ学習対話データからアクション継続長の確率分布を計算し，この継続長確率分布を考慮して，最終的なアクションの選択を行う．本手法を用いて，聞き役対話のアクション生成すなわち対話行為生成に応用した結果，提案手法の有効性を確認した．

6. 謝辞

本研究の一部は，科研費（新学術領域）「人とロボットの共生による協創社会の創成」における計画研究「ロボットのコミュニケーション戦略の生成」(21118004) の助成を受けたものである．

参考文献

[Sutton98] Sutton, R. S. and Barto, A. G.: Introduction to Reinforcement Learning, The MIT Press, (1998).

[Russell03] Russell, S and Norvig, P.: Artificial Intelligence: a Modern Approach Second Edition, Prentice Hall, (2003).

[Williams05] Williams, J., Poupart, P. and Young, S.: Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management, Proc. SIGdial, pp. 25-34, (2005).

[Kim08] Kim, K., Lee, C., Jung, S. and Lee, G. G.: A Frame-Based Probabilistic Framework for Spoken Dialog Management using Dialog Examples, Proc. SIGdial, pp. 120-127, (2008).

[Williams07] Williams, J.: Using Particle Filters to Track Dialogue State," Proc. ASRU, pp. 502-507, (2007).

[Schmidt-Rohr2008] 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, pp. 53-62, (2008).

[Minami09] Minami, Y., Mori, A., Meguro, T., Higashinaka, R., Dohsaka, K. and Maeda, E.: Dialogue Control Algorithm for Ambient Intelligence based on Partially Observable Markov Decision Processes, Proc. ISCA IWSDS, pp. 254-263. (2009).

[Hori09] Hori, C., Ohtake, K., Misu, T., Kashioka, H. and Nakamura, S.: Weighted Finite State Transducer based Statistical Dialog Management, Proc. ASRU, pp. 490-495, (2009).

[Minami10] Minami, Y., Higashinaka, R., Dohsaka, K., Meguro, T. and Maeda, E. : Trigram Dialogue Control Using POMDPs, Proc. SLT, pp. 336-341, (2010).

[Meguro10] Meguro, T., Higashinaka, R., Minami, Y. and Dohsaka, K.: Controlling Listening-Oriented Dialogue Using Partially Observable Markov Decision Processes, Proc. COLING, pp. 761-769, (2010).