Vol.2011-HCI-142 No.3
2011/3/17

アクション継続長制御を利用する POMDP 対話制御

南 泰浩†,目黒 豊美†, 東中 竜一郎‡,
堂坂 浩二†, 前田 英作†

この報告ではアクションの継続長制御を利用する POMDP による対話制御手法を提案する。我々は、これまで，POMDP による対話制御に，Trigram モデルによる統計的な対話制御を取り入れる手法を提案してきた。しかし，この手法は，対話タスクを自動的に学習することができる反面，高い確率を持っているアクションを過剰に生成する問題点があることが実験からわかってきた．本稿では，この問題点を解決するため POMDP を用いる対話制御において，アクション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導入する。 実験結果において，提案方法はアクションの Trigram 確率を高く保ちながら，偏りのないアクション生成を実現できることを確認した．

POMDP dialogue control using action durations

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

This paper proposes a dialogue control method using action durations. We previously proposed a combined method of an ordinary POMDP-based method and a probability-based method and extended it to treat trigram dialogue control. When we apply this method to less task-oriented dialogues, the method over-generates actions that have high probabilities. To avoid this problem, 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 increase means that the action generation gives new information and avoids over-gererating the same actions. This confirms that our method generates appropriate action sequences.

1. はじめに

我々の目標は，人と人との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は，POMDP (Partially Observable Markov Decision Process)[1-2]が近年注目を集めている[3-7]．POMDP は，ある状態下でシステムが行動するアクションに報酬を付与し，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，近年，鉄道乗車券の購入[3-4]，天気予報対話[5]，デジタル加入者線トラブルシューティング[6]，ロボット行動制御[7]などのタスク達成型のシステムが提案されている．これらのシステムは，事前にシステムの挙動が既知であるため，POMDP 環境を比較的簡単に設定できる．しかし，我々は，ユーザの対話の目的が対話そのものを楽しむようなタスクを研究の対象としている．このような対話では，システムの挙動を予め明確に設定することができず，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．

対話データから対話制御を学習する手法として，人間の評価に基づいて自動的に報酬の決定を行う手法を我々は提案した[8-9]．この手法では，DBN(Dynamic Bayesian Networks)を用いて，大量のデータから状態遷移確率と出力確率を求める．この学習の過程で，人間の評価値も確率変数として DBN を学習し，平均化することにより報酬を決定する．POMDP での対話制御方法すなわち方策の学習には，Value iteration を利用する[1-2]．この手法に加え，対話の自然性を考慮し，方策が予測確率の高いアクションを選択するように POMDP の報酬を設定する手法を導入した．我々は，この手法を聞き役対話に応用し，実験によりその有効性を確認した [10-11]．さらに我々は，提案する POMDP による対話制御を用いることにより，Trigram による対話処理[12-13]を包含する手法が実現できることを示した[14-15]．我々はこの手法も聞き役対話に応用して見た．しかし，予備実験を行った結果，ある特定のアクション系列のみを生成する現象が頻繁に起こることが分かった．これは，この手法が予測確率の大きいアクションを選択的に選ぶことによって生じる．本稿では，この問題を解決するため，アクションの継続長の確率分布を予め計算し，ある特定のアクションが長い間続かないようにこの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．

†NTTコミュニケーション科学基礎研究所
†NTT Communication Science Laboratories, NTT corporation
‡NTT サイバースペース研究所
‡NTT Cyber Space Laboratories, NTT corporation

2. 対象とする対話タスク

ここで，対象とするタスクは，ユーザの話を聞くことによって「話したい」「聞いてもらいたい」というユーザの欲求を満たすものである．我々は，この対話を聞き役対話と呼んでいる[10-11]．図1 は典型的な聞き役対話の例である．左の覧に発話を，右の覧にそれに対応する対話行為を記述している．図中，S は話し役，L は聞き役を表す．このような対話では，対話行為の構造をモデル化し対話行為の生成ルール（ルールによる対話制御）を，人手で書くのは容易ではない．本稿では，対話制御を自動的に学習する手法について述べる．今回の実験では，発話の生成ではなく，対話行為までの生成を対象とする．また，発話から対話行為への変換，すなわち，対話理解も人手により与えられると仮定する．

発話	対話行為
S: こんばんは．	挨拶
「食事」でお願いします．	挨拶
L: はい，よろしくお願いします．	挨拶
S: 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)
L: 好きです．	共感・同意
S: おお，お好きですか．	繰り返し
私も好きなんです．	共感・同意
L: 外食が主ですか？	質問(sub: 習慣)
S: いえ，自宅で作ります．	自己開示(sub: 習慣)
特に隠し味はありませんがかつおだしでのばしてうどんにもします．	自己開示(sub: 習慣)
L: うわー，それすっごい美味しそうです！	自己開示(sub: 評価(ポジティブ))

図 1 典型的な聞き役対話例．対話テーマは「食事」で，一文毎に対話行為ラベルが一つずつ付与されている．S は話し役，L は聞き役

3. POMDP

図 2 に POMDP の基本的な構造を示す. POMDP は集合のセット (S,O,A,T,Z,R,γ,b_0) で定義される[1-2]．S は状態の集合であり，その要素を s（$s \in S$）とする．

O は，観測値の集合であり，その要素を o（$o \in O$）とする．A はアクションのセットであり，その要素を a（$a \in A$）とする．T はアクション a によって状態が s から s' へ変化するときの状態遷移確率 $P(s'|s,a)$ 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観測値出力確率 $P(o'|s',a)$ 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．図 2 は POMDP を構成する変数の依存関係を示す．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，四角はシステム側が選択する固定の変数を表す．次に，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，扱えるのはその確率分布だけである．状態の確率分布は HMM と同様に一つ前の時刻の状態の確率分布から計算できる．いま，一つ前の時刻の状態の確率分布 $b_{t-1}(s)$ が分かっているものとする.この確率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．

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

ここで η は全体の和を 1 にするための正規化項を表す．また，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．この確率分布を用いて，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される．

$$V_t^{\pi} = E_{\pi}\left[\sum_{\tau=0}^{\infty} \gamma^{\tau} \sum_s b_{\tau+t}(s) r(s, a_{\tau+t})\right] \quad (2)$$

この式では正定数 $\gamma(<1)$ の設定により未来の報酬の寄与を小さくしている．ここで，π は，アクションを選択する行動制御方法，すなわち方策を表す．POMDP では，式(2)を最大にするアクション a の系列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，$b_t(s)$ の確率分布から a を返す最適な関数を求めることができる．この関数を最適な方策（このあと，ただ単に方策と記述することもある）と呼ぶことにする．もし，POMDP の環境が既知であれば，最適な方策は，Value iteration という手法で求めることができる[1-2]．ここでは，計算時間の削減のため，Exact な Value iteration の代わりに，その近似手法である Point-based value iteration を用いる[16-18].

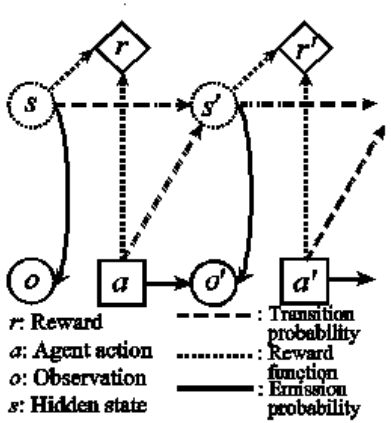

図 2 一般的な POMDP

4. アクションに対する人間の評価とアクション予測確率を利用する POMDP 対話制御

この節では，我々が提案するアクションに対する人間の評価とアクション予測確率を用いる対話制御について述べる[8-9].

4.1 対話制御の目的

ここでは次の二つの条件を仮定する．

1） データの統計量が未知であり，大量のデータからこれを学習できる．

2） このデータ中には実現したい目標の対話の系列が含まれており，人間がその部分あるいは全体の対話を評価し点数化できる．

これらの条件の下，次の二つの目的を同時に設定する．

一つは，アクションに対する人間の評価を最大化することである．もう一つは，生成されたアクション系列の予測確率を最大化することである．通常はこれらを同時に実現するのは難しいので，ここでは，これらを重み付きでバランスさせる方法を提案する．

我々はこの実現のため以下に示す四つの手法を用いた（注：最初の二つは，人間の評価を最大化するための手法であり，後の二つはその手法を前提にアクション予測確率を最大化するための手法を述べたものである.）

(1) DBN による対話構造の学習

対話の確率的構造を POMDP と等価な DBN を使って学習する．このとき，人間の評価も確率変数として扱う.

(2) 人間の評価から報酬への変換

DBN から POMDP への変換, および, 目標対話の評価を POMDP の報酬へ変換する.

(3)POMDP へのアクション予測確率の導入，およびその最大化のための報酬設定

POMDP においてアクションの予測確率を求める状態の導入，および，その予測確率を最大にする報酬の設定を行う.

(4)POMDP 報酬の統合方法，および，方策の決定

(1)を説明するために, 一般的な POMDP とそれに対応する DBN を図 2, 図 3 に示す. この図を参照しながら 4.2 で説明を行う．また，(2)と(3)に関しては，それぞれに，報酬 r_1 と r_2 を設定する. これらの報酬の計算方法については 4.3 および 4.4 で説明する. (4)に関しては 4.5 において説明を行う.

4.2 DBN による対話構造の学習

POMDP では，環境すなわち，状態遷移確率，観測値出力確率と報酬の設定が必要である．ここでは，これらのパラメータをデータから学習する．このパラメータを学習するために POMDP に対応する DBN(図 3)を用いる．ここでは，人の目的対話への評価を示す変数 d を POMDP の報酬 r を計算するために用いる．この d は，人と人または人とシステムの対話が終了した後，その対話系列を見ながら，人がその対話を評価した値である．この評価には，アンケートの回答を利用する．たとえば，”システムに親近感を感じましたか”などの質問を行い，その結果を評価値 d として用いる．この値が高い対話を目標の対話とする. DBN の確率を EM アルゴリズムによって学習し対話の統計的構造を学習する.

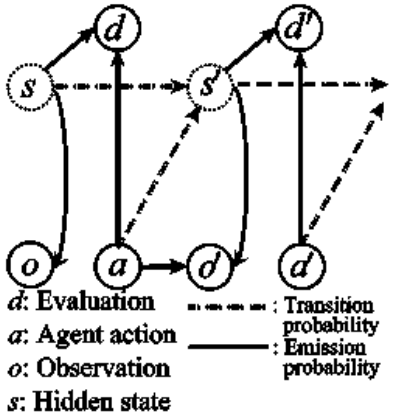

図 3 POMDP に対応させた DBN

4.3 人間の評価からの報酬変換

DBNの確率構造からPOMDP確率構造への変換はほぼ同じ構造のためそれほど困難ではない．確率構造の依存関係を考慮しながら変換すればよい．すなわち，報酬を求めることが問題であるが，ここでは，評価値 d から目標対話に対する r_1 という報酬を d の確率を使って以下のように変換する．

$$r_1(s,a)=\sum_d d\times \Pr(d\mid s,a) \tag{3}$$

この式は， s,a に対する目標対話の平均評価を表している．すなわち， s,a がどれだけ目標対話に貢献するかを示している．

4.4 POMDP へのアクション予測確率の導入，およびその最大化のための報酬設定

タスク達成を目的としない対話制御においては，目標の対話を実現するだけなく，対話の自然性を考慮することが必要である．我々は，学習データの確率に従ってシステムがアクションを生成することで自然な対話を実現できると考えた．ここでは，アクションの予測確率をアクション制御に反映するために，我々は，図 2 の POMDP と図 3 の DBN の構成を図 4 と図 5 のように変えた(変換時に近似計算も導入している)．ここでは， a に一対一に対応する s_a という状態を付加した（状態を $s=(s_o,s_a)$ と記述する）．ここで s_o は図 2 の状態と同じ役割をする． s_a の役割は，アクション a の予測確率を推定することである．また，その予測確率を最大化するアクション a を選択するためのものでもある．我々は s_a を a に一対一に対応させるために，図 3 の DBN において $a=s_a$ のとき， $\Pr(a\mid s_a)=1$ と設定した．この定式化により， $a_t=s_a$ のときに以下の式が得られる．

$$\Pr(a_t\mid o_1,a_1,\ldots,a_{t-1},o_t)$$

$$=\Pr(s_a\mid o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t)=b_t(s_a) \tag{4}$$

ここで，

$$b_t(s_a)=\Pr(s_a\mid o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t)=\sum_{s_o} b_t(s). \tag{5}$$

である．ここで， s_a の時刻 t での分布を $b_t(s_a)$ とする．以上の仮定を置くと，アクション予測確率が隠れ状態の生起確率 $b_t(s_a)$ と等価になることがわかる．次に状態遷移確率について述べる．ここでは，状態遷移確率には計算量の削減を考え，次の近似を利用する．

$$\Pr(s'\mid s,a)\approx \Pr(s'_o\mid s_o,a)\Pr(s'_a\mid s_a,s'_o) \tag{6}$$

$$\Pr(o'\mid s',a)\approx \Pr(o'\mid s'_o) \tag{7}$$

この依存構造は図 4 に示されている．ここで，これらの確率を用いて $b_t(s')$ の更新式を以下のような漸化式で定義する．

$$\begin{aligned} b_t(s') &= b_t(s'_o,s'_a)\\ &=\eta\Pr(o'\mid s'_o)\sum_s \Pr(s'_a\mid s_a,s'_o)\Pr(s'_o\mid s_o,a)\Pr(a\mid s_a)b_t(s_o,s_a)\end{aligned} \tag{8}$$

s_a の導入により，目的関数は以下のようになる．

$$V_t^{\pi}=E^{\pi}\left[\sum_{\tau=0}^{\infty}\gamma^{\tau}\sum_s b_{\tau+t}(s_o,s_a)r((s_o,s_a),a_{\tau+t})\right] \tag{9}$$

ここでの我々の目標は， $o_1,a_1,\ldots,o_{t-1},a_{t-1},o_t$ が与えられている時に， a_t の予測確率が将来に渡って最大になるように a_t を選択することである．そこで我々は $b_{\tau+t}(s_a)$ が大きくなるときに POMDP が高い報酬を得るように報酬を設定する．

これを行うために， $s_a=a$ のときに $r_2(s=(*,s_a),a)=1$ とし，それ以外は $r_2(s=(*,s_a),a)=0$ となる報酬を設定する．ここで，*は任意の状態 s_a を表す．

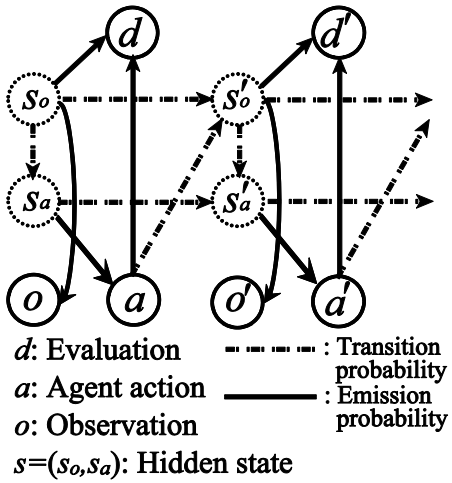

図 4 アクション予測確率を反映する DBN

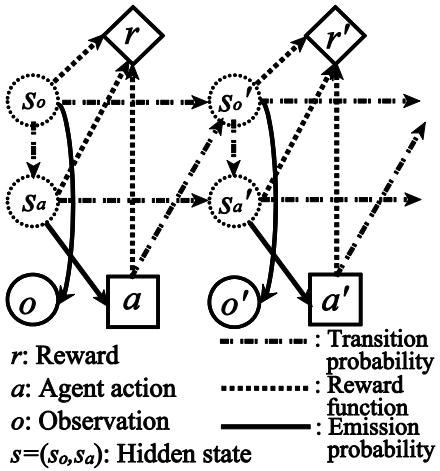

図 5 アクション予測確率を反映する POMDP

4.5 POMDP 報酬統合方法，および，方策の決定

以上で二つの報酬の設定方法を述べたが，最終的な方策は式(9)の報酬 r を以下の式で置き換えたものとなる．

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

ここで w は報酬の重みを表す．この報酬の設定により POMDP の平均報酬 V_t^{π} が得られる．最終的に，Value iteration を行うことによって POMDP の方策が学習できる．この定式化を用いると POMDP の方策は高い予測確率を示し，かつ，同時に目標の対話を実現するようなアクションを選ぶようになる．

5. Trigram 対話制御

我々は，これまで述べてきた手法において確率構造を注意深く構成することにより，Trigram を用いた対話制御を我々の提案する POMDP による対話制御に取り込むことができることを確認した[12-13]．Trigram を用いた理由は，以前提案していた手法では確率構造に大きな制限を設けていたが，これらの制限を部分的に外し，精度のよい対話モデルを実現したかったからである．この対話制御方法を聞き役対話に応用した時に，今回のアクション継続長制御に関する問題点が発生した．そこで，今回，提案する POMDP による対話制御においてその問題の解決を図る．本節では，最初に，その前提となる Trigram 対話制御について説明を行う．

5.1 Trigram アクション制御

文献[12-13]に基づく定式化では，Trigram による対話制御を以下の式で与える．

$$a_t = \arg\max_{a_t} \left[\max_{a_t,o_t} \{\Pr(o_t \mid o_{t-1}, a_{t-1}) \Pr(a_t \mid a_{t-1}, o_t)\} \right] \tag{11}$$

ここでは，o_t は観測値ではなく観測値の属するカテゴリであり不確定性が存在する．この式において，o_t を確定的な観測値だとすると，次式を得る．

$$a_t = \arg\max_{a_t} \{\Pr(a_t \mid a_{t-1}, o_t)\} \tag{12}$$

また，一つ先のターンまで Trigram を考慮し，Trigram 確率を最大とする観測値 o_{t+1} 選択する代わりに，Trigram 確率の和を利用すると以下の式に示す対話制御が得られる．

$$a_t = \arg\max_{a_t} [\max_{a_t,a_{t+1}} \{\Pr(a_t \mid a_{t-1}, o_t) \cdot \sum_{o_{t+1}} \Pr(o_{t+1} \mid o_t, a_t) \Pr(a_{t+1} \mid a_t, o_{t+1})\}] \tag{13}$$

5.2 Trigram 対話制御への拡張

5.1 で述べた Trigram を提案する POMDP に導入する手法を示す．これを実現するため，我々は，図 5 おいて，o と s_o を一対一に対応させるため，$o = s_o$ のとき $\Pr(o \mid s_o) = 1$ と設定した．これにより，$o_t = s_o$ と $a_t = s_a$ が成り立つときに，次式を得る．

$$\begin{aligned} \Pr(s' \mid s,a) &= \Pr(s'_O \mid s_O, a) \Pr(s'_A \mid s_A, s'_O) \\ &= \Pr(o' \mid o, a) \Pr(a' \mid a, o') \end{aligned} \tag{14}$$

この式から，5.1 で述べた Trigram による対話制御式(13)で使われている確率と等価であることが分かる．ここで $r_1((s_o,*),a) = 0$ と設定すれば，我々は将来獲得するアクションの Trigram の統計量の和の目的関数が得られる．

Trigram 対話制御と提案手法の違いは，提案手法が将来の報酬の各時刻の和を計算していることである．Trigram による対話制御は，これに対し，Trigram 確率の最大パスを探し，そのパスのアクションを選択している．

6. アクション継続長制御

我々は 5 節で述べた手法を聞き役対話に応用してみた．しかし，予備実験を行った

結果, ある特定のアクション系列のみを生成する現象が頻繁に起こることが分かった. この問題を解決するため, アクションの継続長の確率分布を予め計算し, ある特定のアクションが長い間続かないようにこの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. 以下ではその手法について述べる.

6.1 対話制御へのアクション継続長制御の導入

対話アクションの継続長制御を行うために, アクションの継続長確率分布 $p_a(l)$ を学習データからあらかじめ計算しておく. ここで l は, アクションの継続回数である. このような確率を使う手法は, HMMを用いた音声認識での継続時間制御として知られている[19]. しかし, HMMでの継続時間長制御は, 隠れ状態で行うのに対して, ここでは, 隠れ変数ではなく, 決定的なアクション系列に対して行う. このため, アクション継続長の確率は, 極めて安定的に求めることができる. 本手法では, この継続長確率分布と対話の履歴から得られる継続長確率 $\bar{p}_a(l)$ から, 以下の差分確率を計算する.

$$\Delta \Pr_a(l) = \Pr_a(l) - \bar{\Pr}_a(l) \,|\, if\ \bar{\Pr}_a(l) < 0\ then\ \bar{\Pr}_a(l) = 0 . \quad (15)$$

ここで, この差分確率を利用するのは, 同じ継続長のアクション系列が何回も続かないようにするためである. アクションの継続長が増える毎に, (15)の確率に従ってアクションを生成する. これには, 図6に示すルールを利用する.

IF　継続長== 1:
　　最適なアクションを選択
　　継続長を一つ増やす
ELSE:
　　IF０～１までの乱数 $< \Delta \Pr_a(duration) / \sum_{d=duration} \Delta \Pr_a(d)$:
　　　最適なアクションを選択
　　　継続長を一つ増やす
　　Else:
　　　２番目に最適なアクションを選択
　　　継続長を１に設定

図6 アクション選択ルール

7. 実験結果

7.1 実験設定

聞き役対話をタスクとして評価を行った. 聞き役対話の対話行為を32個定義し, アクションと観測値のラベルとした. データ中の対話では, 人間は一つのターンで複数の発話を生成するが, POMDPでは, このような複数の発話を同時に扱えない. ここでは, これらの発話を扱えるように, 二つの連続する発話があった場合には, その二つの発話の間に, もう一方の人間の”eps”というラベルを挿入した. ”eps”の意味は何もしない対話行為を行ったということである. 対話の後で, 一人のアノテータ（実験参加者でない）が７レベルのリッカートスケールで対話に関するアンケートに答えた. アンケートはユーザの満足度に関して聞いた. 250の対話を学習データとしてDBNとTrigramを計算した. 平均の対話の長さは, 29.1である. シミュレーションでは, 対話行為のレベルの生成しか行っていない. 各対話は50ターンからなり, 継続長制御を行なうものと行わないものとで評価を行った. ユーザの観測値を生成するため, 学習データより得られたTrigram確率を利用した. この確率に従って, 観測値をランダムに生成した. 実験では, 1000のシミュレーション対話を評価のために作成した.

7.2 評価尺度

予備実験で, 聞き役対話は, シミュレーションの結果に多くの”eps”ラベルのアクションを生成することが分かった. これは, ”eps”の生起確率が極めて高いからに他ならない. この現象の相対的な頻度を効率的に評価するため, 我々は以下の二つの距離尺度を利用した.

$$Entropy = \sum_a -\Pr_{generated}(a) \log(\Pr_{generated}(a)) \quad (16)$$

$$Distance = \sum_a (\Pr_{generated}(a) - \Pr_{training}(a))^2 \quad (17)$$

ここで, $\Pr_{generated}(a)$ は, 生成されたデータのアクションの生起確率である. $\Pr_{training}(a)$ は, 学習データ中のアクションの生起確率である. どちらも, ユニグラムの確率を使って計算した. *Entropy* は, 情報源の情報量を表している. もし, 情報源がいつも同じラベルを出力するのであれば情報量は小さい値を示す. そのため, *Entropy* は, ”eps”ラベルを沢山生成しているかどうかをチェックする良い指標となる. *Distance* は, 生成されたデータの生起確率と学習データの生起確率の間の距離であり, 生成された統計が学習データの統計に近いかどうかの指標である.

また, もともとの性能をチェックするため, ここでは以下の２つの評価尺度を用意

した．

一つは，次のように定義する平均の Trigram 確率である．

$$Trigram = \frac{1}{N}\sum_{i}^{N}\frac{\sum_{t}\Pr_{trigram}(a^{i}_{t+1} \mid a^{i}_{t}, o^{i}_{t+1})}{L_i}, \quad (18)$$

ここで，N は，対話の総数であり L_i は各対話の長さである．そして $\Pr_{trigram}(a^{i}_{t+1} \mid a^{i}_{t}, o^{i}_{t+1})$ は学習データの Trigram 確率に対して，生成されたアクション観測値 $a_{t+1}\ a_{t}, o_{t+1}$ を代入したものである．この指標は，生成されたアクションがどれだけ学習データの確率を大きくするかを示すものとなっている（確率ベースの対話制御の適切さを示す指標）．もう一つの指標として，以下に定義するユーザの評価尺度を利用する．

$$Satisfaction = \frac{1}{N}\sum_{i}^{N}\frac{\sum_{t}\bar{d}(a_t, o_t)}{L_i}, \quad (19)$$

ここで，$\bar{d}(a,o)$ はユーザ観測値とアクションのペアの平均の評価値である．この尺度は，観測値が観測された後，あるアクションを生成したときの人間の評価の平均を示している．もちろん，ユーザの評価は過去の対話履歴に大きく依存するが，ここでは，ユーザの評価はユーザの観測値に対する直後のシステムのアクションが最も大きな影響を与えると仮定する．また，この評価は，POMDP の目的関数と近いので，システムそのものの評価にはならないが，他の新たな機能を加えた場合に評価結果が落ちないことを調べる上では有効な値である．

7.3 実験結果

我々は，4.5 で述べた重みを w=10, w=15, と w=20 のように 3 種類使って実験を行った．この条件で，アクション継続長制御ありとなしで評価を行った．ここで，w=10 の場合，POMDP の方策はユーザの評価に重きをおき，w=20 の場合，アクションの生起確率に重きを置いていることを意味する．表 1 は，以上の 6 種類の手法の各評価尺度の値を示す．また，この表には学習データの値も示す．

表 1　アクション継続長制御ありとなしの場合の POMDP 対話シミュレーション結果

	アクション継続長制御なし			アクション継続長制御あり			学習データ
Weight	w=20	w=15	w=10	w=20	w=15	w=10	
Entropy	1.41	1.63	0.19	1.56	1.71	1.49	2.63
Distance	0.070	0.041	1.04	0.046	0.035	0.306	
Trigram	0.363	0.356	0.018	0.337	0.334	0.029	0.063
Satisfaction	3.68	3.74	5.91	3.73	3.76	4.60	3.70

表 1 から，*Entropy* に関しては，アクション継続長制御ありの方が，ない方より良いことがわかる．また *Distance* についても，アクション継続長制御ありの方が生成されたアクションの確率分布が学習データの確率分布に近いことがわかる．このときの Trigram の確率は若干下がるものの, 同じラベルを出力するという行動が大幅に減っていることから本手法の有効性がわかる．人間の評価に関しては，アクション継続長制御を使うことで，Trigram の確率が高いところでは，満足度も高くなることがわかる．

8. 考察

我々の提案する POMDP による対話制御方法は，アクションの確率，アクションに対する人間の評価，アクションの継続長確率の 3 つのファクタを考慮して，アクションの選択を行っている．今回，焦点を当てたアクション継続長による対話制御は，システムが同じアクションを何回も繰り返すことを回避するものである．この評価の中で，*Entropy* すなわち情報量という評価尺度を利用した．本報告では，情報量を同じ系列が続くか続かないかを調べるために利用した．しかし，本来，情報量は，ユーザにとって情報が重要かを測る指標である. 聞き役対話などのタスクにおいて, 対話を「継続」させるためには, ユーザにとってもの珍しい情報(重要な情報)を提供することが，重要だと考えられる．このことから，今回のテーマであるヒューマンコンピュータインタラクションの「継続」という観点からこの問題を考えると，このようなタスクでは，情報量という概念が対話の「継続」を考える上で，極めて大きなファクタになることがいえる．

9. まとめ

今まで提案してきた POMDP による Trigram を利用する対話制御において，Trigram 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．

10. 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．

文献

[1] R. S. Sutton and A. G. Barto, “Introduction to Reinforcement Learning,” The MIT Press, 1998.
[2] S. Russell and P. Norvig, “Artificial Intelligence: a Modern Approach Second Edition,” Prentice Hall, 2003.
[3] J. Williams, P. Poupart, and S. Young, “Partially Observable Markov Decision Processes with Continuous Observations for Dialogue Management,” *Proc. SIGdial*, pp. 25-34, 2005.
[4] J. Williams, P. Poupart, and S. Young, “Factored Partially Observable Markov Decision Processes for Dialogue Management,” *Proc. IJCAI*, pp. 75-82, 2005.
[5] K. Kim, C. Lee, S. Jung, and G. G. Lee, “A Frame-Based Probabilistic Framework for Spoken Dialog Management using Dialog Examples,” *Proc. SIGdial*, pp. 120-127, 2008.
[6] J. Williams, “Using Particle Filters to Track Dialogue State,” *Proc. ASRU*, pp. 502-507, 2007.
[7] S. R. Schmidt-Rohr, R. Jakel, M. Losch, and R. Dillmann, “Compiling POMDP Models for A Multimodal Service Robot from Background Knowledge,” *European Robotics Symposium*, pp. 53-62, 2008.
[8] Y. Minami, A. Mori, T. Meguro, R. Higashinaka, K. Dohsaka, and E. Maeda, “Dialogue Control Algorithm for Ambient Intelligence based on Partially Observable Markov Decision Processes,” *Proc. ISCA IWSDS*, pp. 254-263. 2009.
[9] Y. Minami, A. Mori, T. Meguro, R. Higashinaka, K. Dohsaka, and E. Maeda, “Dialogue Control by POMDP Using Dialogue Data Statistics,” Spoken Dialogue Systems Technology and Design, G. Lee, J. Mariani, W. Minker and S. Nakamura Ed., New York, Springer-Verlag Inc., 2010.
[10] T. Meguro, R. Higashinaka, Y. Minami, and K. Dohsaka, “Controlling Listening-Oriented Dialogue Using Partially Observable Markov Decision Processes,” *Proc. COLING*, pp. 761-769, 2010.
[11] 目黒豊美, 東中竜一郎, 南泰浩, 堂坂浩二, “POMDPを用いた聞き役対話システムの対話制御”, 言語処理学会第17回年次大会, 2011.
[12] C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, “Weighted Finite State Transducer based Statistical Dialog Management,” *Proc. ASRU*, pp. 490-495, 2009.
[13] C. Hori, K. Ohtake, T. Misu, H. Kashioka, and S. Nakamura, “Recent Advances in WFST-based Dialog System,” *Proc. Interspeech*, pp. 268-271, 2009.
[14] 南泰浩, 東中竜一郎, 堂坂浩二, 目黒豊美, 森啓, 前田英作, “POMDPによるtrigram対話制御”, 電子情報通信学会技術研究報告, SP2010-90, pp. 37-42 2010.
[15] Y. Minami, R. Higashinaka, K. Dohsaka, T. Meguro, and E. Maeda, “Trigram Dialogue Control Using POMDPs,” *Proc. SLT*, pp. 336–341, 2010.
[16] P. Poupart, “Exploiting Structure to Efficiently Solve Large Scale Partially Observable Markov Decision Processes,” Ph.D. dissertation, University of Toronto, 2005.
[17] T. Smith and R. G. Simmons, “Point-Based POMDP Algorithms: Improved Analysis and Implementation,” *Proc. UAI*, pp. 542-547, 2005.
[18] J. Pineau, G. Gordon, and S. Thrun, “Point-Based Value Iteration: An Anytime Algorithm for POMDPs,” *Proc. IJCAI*, pp. 1025-1032, 2003.
[19] B. H. Juang, L. R. Rabiner “Mixture Autoregressive Hidden Markov Models for Speaker Independent Isolated Word Recognition,” *Proc. ICASSP*, pp. 41-44, 1986.