状態空間の分割と状態遷移の学習に基づく Parallel POMDP *

☆山田燿司, 滝口哲也, 有木康雄 (神戸大)

1　はじめに

近年、レストラン検索 [1] やカーナビゲーションシステム [2] など、タスク指向型の音声対話システムを構成する際に、部分観測マルコフ決定過程（POMDP）が多く利用されている [3][4][5]。我々はこれまで、この POMDP を階層的に配置した Hierarchical POMDP(H-POMDP) を対話システムに適用したが、POMDP 間の遷移が制限される問題があった。この点から、本研究では、POMDP を並列に配置することで、POMDP 間の遷移をより自由にした、Parallel POMDP(P-POMDP) を提案する。また、この POMDP 間の遷移の学習を行い、より適切なシステムの発話を選択する。

2　商品検索型音声対話システム

2.1　商品検索の構成

POMDP を適用するタスクとして、商品の検索を目的とした音声対話システムを考える。商品とは本や CD、食品などの売り買いの対象となる物品を指す。今回は、本の検索のみをタスクとして設定する。本の種類は小説や雑誌、参考書などに分類され、さらに小説は推理小説や歴史小説などのジャンルに分けられる。また、読者の年齢層や性別により、好まれる本が異なってくる場合もある。このように、細かくカテゴリに区分されているものの中から、システムの質問に答えることで一つずつ条件を絞り、抽象的な商品のイメージを具体化し、最終的におすすめの本を一冊紹介してくれるシステムを構築する。

2.2　対話例

このシステムの実際の対話例を以下に示す。

システム：どんな本をお探しでしょうか？
ユーザ　：小説が読みたい。
システム：ジャンルは何でしょうか？
ユーザ　：推理がいいな。
システム：年はいくつですか？
ユーザ　：22 歳です。

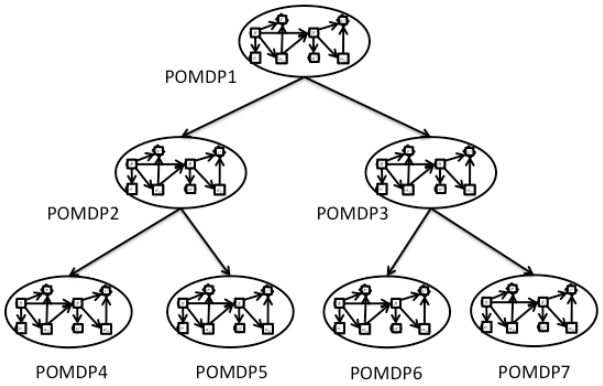

Fig. 1　H-POMDP の構造

システム：では、こちらがおすすめです。

3　Hierarchical POMDP

POMDP の問題点として、状態数が増大するにつれて計算量が指数的に増えることが挙げられる。そこで、Fig.1 のように、POMDP を階層的に構築し、状態空間を分割した Hierarchical POMDP (H-POMDP) を用いた研究がある [6][7][8]。タスクを複数の部分問題に分割し、階層的に統合することで、従来よりも複雑なタスクを実現することが可能となる。

まず、最上位の POMDP1 から開始してその内部で状態の遷移が生じ、POMDP1 における最終目的となる状態にたどり着く。その後、次の POMDP2 もしくは POMDP3 に遷移する。こうして POMDP 間の遷移を続け、最下位の POMDP において最終目的の状態にたどり着いた場合、タスク全体のゴールとなる目的の状態に到達する。

この H-POMDP を商品検索タスクに適用した例を示す。まず、最上位に位置する POMDP は“ 本の種類 ”を決定するための POMDP で、その構成として、状態を {“ 小説 ”,“ 雑誌 ”} の集合、観測値を {“ 小説 ”,“ 雑誌 ”}、行動を {“ 確認する ”,“ POMDP2 へ遷移 ”,“ POMDP3 へ遷移 ”} と設定する。“ 確認する ”とはユーザの発言を促す行動を示す [9]。例えば、最上位の POMDP に

*Parallel POMDP based on the partition of the state space and the learning of state transition.
by YAMADA, Yôji, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe University)

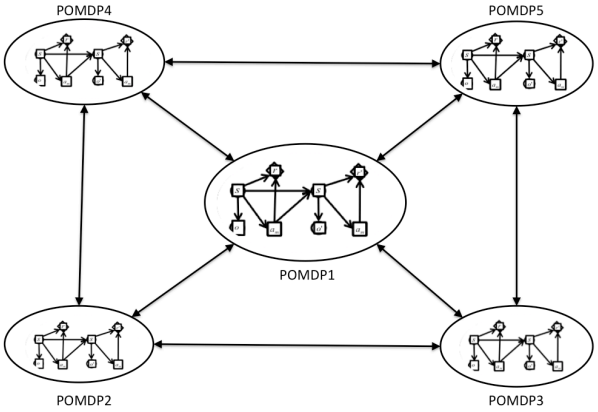

Fig. 2　P-POMDP の構造

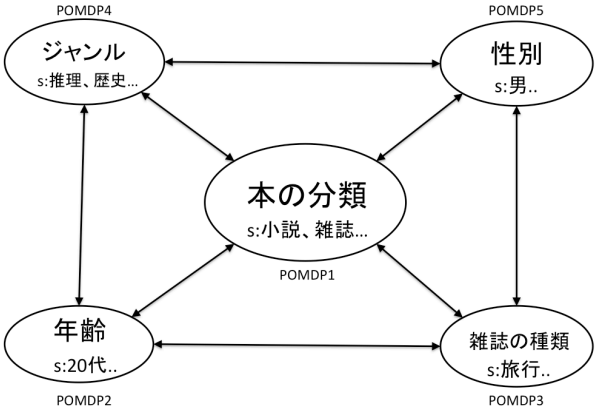

Fig. 3　P-POMDP の適用例

おいて、システムが“ 確認する ”という行動として、本の種類を尋ねる。そして、ユーザ発話を観測して信念状態 $b(s')$ に更新され、“ POMDP2 へ遷移 ”という行動を選択したとする。その後、POMDP2 の行動として、システムは小説のジャンルを尋ねる。こうして POMDP 間の遷移を続け、最下位の POMDP において、“ 本を提示する ”という行動が選択されると、システムがユーザに本を一冊紹介し、タスク完了となる。

4　Parallel POMDP

しかし、この H-POMDP の問題点として、階層構造のため、最上位の POMDP から対話を開始する必要があり、また、一つ下位の POMDP にしか遷移ができないなど、POMDP 間の遷移が制限される問題がある。これらの問題を解決するために、Fig.2 のように POMDP を並列に配置した、Parallel POMDP (P-POMDP) を提案する。これにより、さらに多くの POMDP 間の遷移が可能となり、より柔軟な対話システムが実現できる。

4.1　P-POMDP の構成

P-POMDP は、多数の POMDP を並列に配置し、POMDP 間の遷移をより自由に行うことを目的とする。この P-POMDP を商品検索タスクに適用した例を、Fig.3 に示す。“ 本の分類 ”や“ ジャンル ”など、本の検索条件を絞るための POMDP が存在し、それぞれ“ 小説 ”や“ 雑誌 ”、“ 推理 ”や“ 歴史 ”などの状態を持つ。また、“ 本の分類 ”という POMDP から、状態に応じて、“ ジャンル ”や“ 年齢 ”、“ 雑誌の種類 ”、“ 性別 ”の POMDP に遷移することを表す。しかし、これら POMDP 間の遷移において、どの POMDP から POMDP への遷移が適切かを決定する必要がある。そこで、本のラベルデータを基に、POMDP 間の遷移を学習することを考える。

4.2　POMDP 間の遷移の学習

Table 1 のように、“ 状態 ”と“ その状態を含む POMDP ”の組のラベルデータを持つ本がある。この状態は、その本の特徴を表している。例えば、Book 1 は“ 小説 ”、“ 推理 ”、“ 20 代向け ”の 3 つの状態を持つ。これらはそれぞれ、“ 本の分類 ”、“ ジャンル ”、“ 年齢 ”の POMDP に対応する。この本を検索するためには、これら 3 つの POMDP に遷移し、それぞれの状態を特定する必要がある。そのため、この 3 つの POMDP には関連があり、これらの POMDP 間での遷移が可能であることが言える。例えば、“ 本の種類 ”の POMDP に含まれる“ 小説 ”という状態において、システムは行動として、“ ジャンルの POMDP に遷移 ”もしくは“ 年齢の POMDP に遷移 ”を選択する。このように、本のラベルデータを基に POMDP 間の遷移を決定し、さらにその遷移の数を合計することで、遷移表を作成する。

Table 1 の 2 つの本データから、Table 2 のような遷移表を作成する。例えば、現在選択されている POMDP が“ 本の種類 ”であり、状態が“ 小説 ”の場合を考える。このとき、“ 小説 ”という状態に対し、次に遷移する POMDP として、“ ジャンル ”を含む本が Book1 と Book2 の 2 つ、“ 年齢 ”を含む本が Book1 の 1 つ、“ 性別 ”を含む本が Book2 の 1 つである。これは、“ ジャンル ”の POMDP に遷移する確率、“ 年齢 ”の POMDP に遷移する確率、“ 性別 ”の POMDP に遷移する

Table 1　本のラベルデータ例

Book 1		Book 2	
状態 s	POMDP	状態 s	POMDP
小説	本の分類	小説	本の分類
推理	ジャンル	歴史	ジャンル
20	年齢	男性	性別

Table 2　遷移表 $p(s|a)$ の例

現在の POMDP	状態 s	行動 a:次の POMDP に遷移			
		分類	ジャンル	年齢	性別
本の分類	小説	-	0.5	0.25	0.25
ジャンル	推理	0.5	-	0.5	0
:	歴史	0.5	-	0	0.5
年齢	20	0.5	0.5	-	0
性別	男	0.5	0.5	0	-

確率の 3 つ割合が、2:1:1 であることを表している。さらに、この割合について、総和が 1 になるように正規化を行うと、0.5:0.25:0.25 となる。このように、各状態に対して、可能な遷移をラベルデータに含んでいる本を数え、遷移の比率を算出する。これにより、状態 s において行動 a を選択する確率を表す、遷移表 $p(s|a)$ を作成することができる。そして、この遷移表を基に、報酬を設定する。

報酬の例を Table 3 に示す。まず、“ 確認する ” という行動は、何度も発話を行うユーザの煩わしさを考慮し、状態に関わらず-1 の報酬を与える [9]。また、$p(s|a) = 0$ の値は、その状態 s における行動 a の選択が不可能であることを表しているので、それに対応する $r(s|a)$ は-10 とする。そして、$p(s|a) > 0$ の値を取る各状態 s と各行動 a に対しては、$r(s|a) = 10 \times p(s|a)$ の式を適用し、正の報酬を与える。このような条件の下、報酬の設定を行う。

5　システム評価

この P-POMDP を評価するために、商品検索型の音声対話システムをスマートフォンアプリケーションとして開発した。今回、本のデータを書店のホームページより 200 冊用意し、それを基に 57 個の POMDP を作成した。この 57 個の POMDP の中から、最初のユーザの発話を観測値として含む POMDP に焦点を当て、その POMDP から対話を開始する。また、各 POMDP の平均状態数は 3.6、平均行動数は 3.3 である。これらの POMDP 間の遷移から得られる報酬を、ラベルデータを基に設定し、PBVI より方策の学習を行った [11]。このシステムの動作画面を Fig. 4 に示す。Fig.4 のように、ユーザはシステムからの質問に答えることで、検索の条件を絞り、条件に適合する本が見つかれば、その本を提示する。また、5 回目の対話で条件に適合する本が見つかっていなければ、それまでに得られた条件と、最も適合率が高い本を提示する。

Table 3　報酬 $r(s|a)$ の例

現在の POMDP	状態 s	行動 a				
		次の POMDP に遷移				確認する
		分類	ジャンル	年齢	性別	
本の分類	小説	-	5	2.5	2.5	-1
ジャンル	推理	5	-	5	-10	-1
:	歴史	5	-	-10	5	-1
年齢	20	5	5	-	-10	-1
性別	男	5	5	-10	-	-1

このシステムの評価として、システムと被験者が対話を行い、平均タスク達成率と平均対話回数を調べた。システムが提示した本が、被験者の満足のいく本であれば、タスク達成とする。結果は、平均タスク達成率=36 %、平均対話回数=4.8 回であった。平均タスク達成率が低い理由としては、被験者の発話の特徴として、具体的な作品名や著者名を発話することが多く、今回、本のデータが 200 冊と少ないため、これらの発話に対応できなかったからと考えられる。そのため、今後、本のデータを増やし、より多くのユーザの発話に対応できるように、システムの改良を行う予定である。

6　おわりに

今回、H-POMDP における POMDP 間の遷移の制限を改善するため、POMDP を並列に配置した P-POMDP を提案した。さらに、P-POMDP の遷移をデータより学習し、報酬の設定を行った。これにより、さらに自由な POMDP 間の遷移を行うことができるようになり、状況に応じた、より適切なシステムの行動を選択できるようになったと考えられる。そのため、今後の課題として、各 POMDP のタスク達成率を比較し、P-POMDP の方が H-POMDP よりも適切な対話が行えていることを示す。

また、この P-POMDP の特徴として、状態空間を分割し、複数の小さな POMDP をそれぞれ学

Fig. 4　システム動作画面

習しているため、状態数が大きい1つのPOMDPに比べ、方策の計算時間が少ないことが考えられる。そのため、今後の課題として、学習データを増やしたときに、P-POMDPの方が単体のPOMDPよりも方策の計算時間が短いことを示し、P-POMDPの有用性を証明する[11]。

参考文献

[1] S. Young, M. Gasic, S. Keizer, F. Mairesse, J. Schatzmann, B. Thomson and K. Yu, “ The Hidden Information State model: A practical framework for POMDP-based spoken dialogue management, ”Computer Speech and Language, 24, 150-174, 2010.

[2] 岸本　康秀,　滝口　哲也,　有木　康雄, “ 階層的強化学習を適用したPOMDPによるカーナビゲーションシステムの音声対話制御, ”　電子情報通信学会信学技報, 110(143), 49-54, 2010.

[3] S. Young, M. Gasic, B. Thomson and J. Williams, “ POMDP-Based Statistical Spoken Dialog Systems: A Review, ”Proceedings of the IEEE, 101, 1160-1179, 2013.

[4] Jason D. Williams and Steve Young, “ Partially observable Markov decision processes for spoken dialog systems, ”Computer Speech and Language, 21, 393-422, 2007.

[5] R.S. Sutton and A.G. Barto, “強化学習,” 森北出版社, 2000.

[6] 岸本　康秀,　滝口　哲也,　有木　康雄, “ 階層的強化学習を適用したPOMDPによる音声対話制御, ”　電子情報通信学会信学技報, 110(356), 121-126, 2010.

[7] Joelle Pineau, Nicholas Roy, and Sebastian Thrun, “ A Hierarchical Approach to POMDP Planning and Execution, ”Proceedings of the ICML Workshop on Hierarchy and Memory in Reinforcement Learning, 110, 121-126, 2001.

[8] Heriberto Cuayahuitl, Ivana Kruijff-Korbayova, Nina Dethlefs, “ Nonstrict Hierarchical Reinforcement Learning for Interactive Systems and Robots, ”ACM Transactions on Interactive Intelligent Systems, 4(3), article 15, 2014.

[9] 南　泰浩, “ 部分観測マルコフ決定過程に基づく対話制御, ”日本音響学会誌, 67(10), 482-487, 2011.

[10] Joelle Pineau, Geoffrey J. Gordon, and Sebastian Thrun, “ Point-based value iteration: An anytime algorithm for POMDPs, ”roceedings of the International Joint Conferences on Artificial Intelligence, 1025-1032, 2003.

[11] Minlue Wang and Richard Dearden, “ Run-Time Improvement of Point-Based POMDP Policies, ”Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence, 2408-2414, 2013.