

タスク指向型対話システムにおける強化学習と ニューラルネットワークの比較*

☆山田耀司, 滝口哲也, 有木康雄 (神戸大)

1 はじめに

近年, タスク指向型の音声対話システムを構成する際に, POMDP 等の強化学習 [1] やニューラルネットワーク [2] が多く用いられている. 我々はこれまで, この POMDP を階層的に配置した Hierarchical POMDP(H-POMDP) や, 並列に配置した Parallel POMDP(P-POMDP) を対話システムに適用した. 本研究では, word2vec を用いた対話制御を提案し, タスク成功率や学習時間について, 従来の強化学習との比較を行う.

2 商品検索型対話システム

2.1 商品検索型対話システムの構成

対話システムのタスクとして, 商品の検索を目的とした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. 例えば, 本は小説や雑誌, 参考書などのカテゴリに区分される. 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, 好まれる本が異なる場合もある. このように, 本は“カテゴリ”や“ジャンル”などの slot, それら slot に対する“小説”や“推理”などの value に分類することができる. ユーザがシステムからの質問に答えることで, これら value を特定し, slot を埋め, 本の検索の条件を絞る. こうして, ユーザの意図を推定し, 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. 3つの value, “小説”, “推理”, “20代”が20代向けの推理小説を特定している例を Fig.1 に示す.

2.2 学習データ

本を検索する対話システムを構築するにあたり, Table 1 のような本のラベルデータを用いて学習を行う. 例えば, Book1 は“小説”, “推理”, “20代”の3つの value を持つ. これらはそれぞれ, “カテゴリ”, “ジャンル”, “年齢”の slot に対応する. この本を検索するためには, システムがこれら3つの slot に対応する質問をユーザに行い, 各 value を特定する必要がある. つまり, システムは本のカテゴリを尋ねた後, 本のジャンルを尋ねる, もしくは, ユーザの年齢を尋

```
Sys: どんな本をお探しでしょうか? (<value>="", <slot>="カテゴリ")
Usr: 小説が読みたい. (<value>="小説", <slot>="カテゴリ")
Sys: ジャンルは何でしょうか? (<value>="", <slot>="ジャンル")
Usr: 推理がいいな. (<value>="推理", <slot>="ジャンル")
Sys: 年はいくつですか? (<value>="", <slot>="年齢")
Usr: 22歳です. (<value>="20代", <slot>="年齢")
Sys: では,こちらがおすすめです.
```

Fig. 1 対話例

Table 1 本のラベルデータ例

Book 1		Book 2	
value	slot	value	slot
小説	カテゴリ	小説	カテゴリ
推理	ジャンル	歴史	ジャンル
20代	年齢	男性	性別

ねるという行動が適切であると考えられる. このように, 本のラベルデータを用いてシステムの行動を選択することが可能である.

3 強化学習

3.1 Hierarchical POMDP

強化学習の中でも, POMDP を用いた対話制御が多く研究されている. しかし, この 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.2 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 [3][4]. 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, タスク完了となる.

3.2 Parallel POMDP

しかし, この H-POMDP の問題点として, 一つ下位の POMDP だけ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. この問題を解決す

*Comparison of reinforcement learning and neural networks in the task-oriented dialogue system. by YAMADA, Yôji, TAKIGUCHI, Tetsuya, ARIKI, Yasuo (Kobe University)

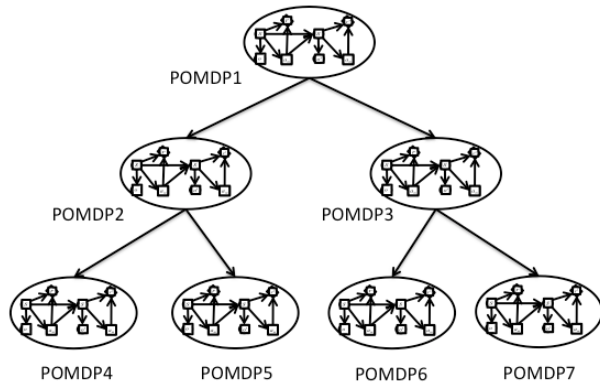


Fig. 2 H-POMDP の構造

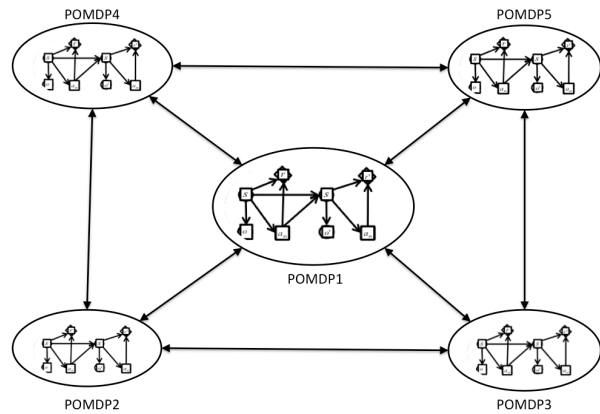


Fig. 3 P-POMDP の構造

るために、我々は Fig.3 のように POMDP を並列に配置した Parallel POMDP (P-POMDP) を提案した [5]. 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. この P-POMDP を商品検索タスクに適用した例を、Fig.4 に示す. “カテゴリ”や“ジャンル”など、slot に相当する POMDP が存在し、それぞれ “小説”や “雑誌”, “推理”や “歴史”などの value に相当した状態を持つ. また, “カテゴリ”という POMDP から、状態に応じて, “ジャンル”や “年齢”, “雑誌”, “性別”の POMDP に遷移することを表す. POMDP 間の遷移を続け、それぞれの value を特定し、特定された value を全て含む本が見つければ、システムはそれを提示する.

しかし、これら H-POMDP や P-POMDP の行動の選択において、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、Table 1 のような本のラベルデータを基に、POMDP 間の遷移を学習する.

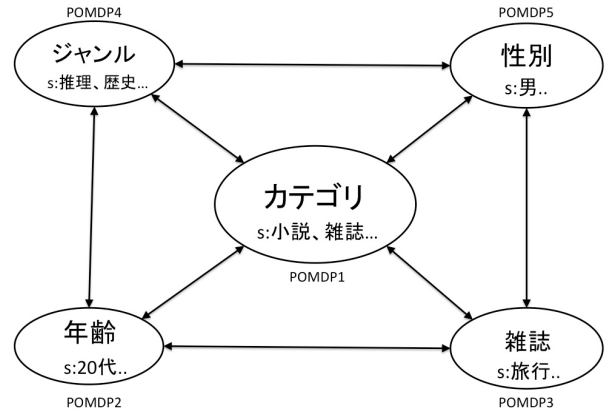


Fig. 4 P-POMDP の適用例

3.3 POMDP 間の遷移の学習

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 “カテゴリ”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 冊, “年齢”を含む本が Book1 の 1 冊, “性別”を含む本が Book2 の 1 冊である. 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 $p(a|s)$ 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 [6]. また, $p(a|s) = 0$ 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, $p(a|s) > 0$ の値を取る各状態 s と各行動 a に対しては, $r(s, a) = 10 * p(a|s)$ 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. 得られた報酬を基に、Point-based Value Iteration (PBVI) [7] を用いて方策の学習を行い、システムは POMDP 間の遷移を選択する.

4 ニューラルネットワーク

4.1 word2vec

近年、Recurrent Neural Network 等のニューラルネットワークを利用した対話システムが多く研究されている [8]. 今回、ニューラルネットワークによる対話システムを

Table 2 遷移表 $p(a|s)$ の例

現在の POMDP	状態 s	行動 a : 次の POMDP に遷移			
		カテゴリ	ジャンル	年齢	性別
カテゴリ	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代	0.5	0.5	-	0
性別	男	0.5	0.5	0	-

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
		カテゴリ	ジャンル	年齢	性別	
カテゴリ	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代	5	5	-	-10	-1
性別	男	5	5	-10	-	-1

構築する際に、Table 1 のようなラベルデータを学習データとして用いるために、word2vec を利用した手法を提案する [9]。word2vec は、テキスト中の各単語をベクトル化し、ニューラルネットを用いて各単語ベクトルの重みを学習し、単語間の意味的な類似度を表現するための手法である。本研究では、word2vec のネットワーク構造として、Fig.5 の Skip-gram モデルを用いる。Skip-gram モデルは入力層、中間層、出力層から構成され、テキスト中のある単語 w_t を入力とし、その周辺の単語 $w_{t-k}, \dots, w_{t-1}, w_{t+1}, \dots, w_{t+k}$ を出力とする。入力層と中間層をつなぐ重み行列は、各単語に対する重みベクトルの集合であり、word2vec が最終的に生成する単語分布表現となる。単語を分布表現とすることで、ある単語を入力したときに、その単語と近いベクトル構造を持つ単語、つまり意味的に関連性の高い単語を複数個出力することができる。

4.2 word2vec による対話制御

学習データとして Table 1 のようなラベルデータを用いることで、word2vec による対話システムを実現する。まず、このラベルデータを Table 4 のようなテキストに変換する。例えば、Book1 を検索する過程において、ユーザから“小説”という value が得られた後、システムは次に“ジャンル”，もしくは“年齢”をユーザに尋ねる行動が適切であると考えられる。このような過程を連続した単語として表現することで、Table 4 のテキスト（小説 ジャンル 年齢）が得られる。そして、このテキストデータを Skip-gram モデルの学習データとして用いる。各テキストにおける、先頭の value に対応する単語を入力、それ以外の slot に対応する単語を出力として学習を行う。これにより、

Table 4 テキストデータ例

Book 1			Book 2		
小説	ジャンル	年齢	小説	ジャンル	性別
推理	カテゴリ	年齢	歴史	カテゴリ	性別
20代	カテゴリ	ジャンル	男性	カテゴリ	ジャンル

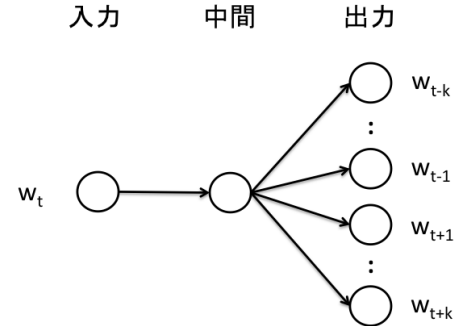


Fig. 5 Skip-gram モデルの構造

ユーザの発話である value を入力すると、次のシステムの質問に対応する slot を出力することができる。こうして得られたモデルに対し、“小説”と入力した例を Table 5 に示す。Cosine distance は 2 単語間の意味空間上の距離（類似度）を表す。この結果を対話システムに適用すると、システムはユーザから“小説”という発話を観測した後、0.824:0.757:0.713 の割合で“本の著者を尋ねる”，“本のジャンルを尋ねる”，“ユーザの年齢を尋ねる”という行動を選択する。このようにして、word2vec を用いた対話システムを構築する。

5 評価実験

5.1 実験設定

従来の POMDP(C-POMDP), H-POMDP と P-POMDP, そして word2vec を用いた 4 つの対話システムを構築した。C-POMDP は状態数が 202, 行動数が 59 の大きな単体の POMDP である。H-POMDP は、この C-POMDP を 57 個の小さな POMDP に分割し、それらを階層的に構築したもので、各 POMDP の平均状態数は 3.6, 平均行動数は 2.5 である。P-POMDP は、分割された 57 個の POMDP を並列に配置したもので、各 POMDP の平均状態数は 3.6, 平均行動数は 3.3 である。word2vec は、202 種類の value と 57 種類の slot の和である 259 種類の入出力データが用いられ、中間層のノード数は 12 次元とした。これら POMDP の報酬と word2vec の重み行列は、書店のウェブサイト上にある、200 冊の本のデータから学習される。システムはユーザに質問を行い、それに対する回答をもとに value を取得し、それら value を全て含む本が見つければ、その本を提示する。また、5 回目の対話でその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 value をより多く含む本

Table 5 word2vec に“小説”と入力した場合の出力

Word	Cosine distance
著者	0.824537
ジャンル	0.757847
年齢	0.713555

Table 6 実験結果

	成功率	学習時間 (s)			
		平均	最大	最少	合計
C-POMDP	72%	-	-	-	8659.32
H-POMDP	51%	0.084	4.496	0.004	4.804
P-POMDP	69%	0.205	11.33	0.004	11.713
word2vec	64%	-	-	-	0.609

を提示する。各対話システムについて、それぞれのタスク成功率と学習時間を比較した。タスク成功率は、テスト用の 100 冊の本データから計算される。ユーザ役がそのテストデータのラベルデータをもとにシステムの問いに答え、システムがテストデータと同種の本を提示できればタスク成功とする。学習時間は、PBVI を用いた各 POMDP の方策の計算にかかる時間と word2vec の重み行列の計算にかかる時間（計算環境 = CPU : Intel(R)Core(TM)i7-2600 3.4GHz, 実装メモリ (RAM) : 12.0GB）を表し、H-POMDP と P-POMDP は 57 個の POMDP の集合なので、各 POMDP の平均、最大、最少、合計の学習時間を測定する。C-POMDP と word2vec については合計の学習時間のみ表示する。

5.2 実験結果

実験の結果を Table 6 に示す。C-POMDP は多くの状態を含むため、多様なユーザの発話に対応でき、タスク成功率が最も高かった。しかし、状態数が多いと、計算量も指数的に増加するため、学習時間が最も長かった。H-POMDP は学習時間が少ないが、遷移が階層的に制限されているためにタスク成功率が最も低くなった。これらに対し、P-POMDP は C-POMDP に近いタスク成功率を示し、さらに、H-POMDP と同程度の計算量を示した。これは、並列構造が遷移の制限を緩和し、システムがより最適な行動を選択できるため、タスク成功率が上昇したからである。word2vec はタスク成功率は P-POMDP より低いですが、学習にはほぼ時間がかからなかった。これは、各 POMDP が累積報酬を最大化するために数ステップ先の対話を予測するのにに対し、word2vec は数ステップ先の対話を予測をせずに、その時点での適切な行動を選択するためである。また、状態空間が分割されている H-POMDP に比べ、word2vec はより多くのユーザ発話に対応できたため、H-POMDP よりタスク成功率が高くなった。

6 おわりに

今回、強化学習の一種である POMDP による対話制御とニューラルネットによる対話制御を比較するために、word2vec を用いた対話システムの構築を提案した。単語のラベルデータを用いて、ニューラルネットによる対話システムを構築する場合、word2vec を適用することで十分な精度の行動選択が可能であることが、タスク成功率より分かった。今後は強化学習とニューラルネット組み合わせた対話システムを構築し、従来のシステムとの比較を行う。

参考文献

- [1] Matthew Henderson, Blaise Thomson and Steve Young, “Deep Neural Network Approach for the Dialog State Tracking Challenge” Proceedings of the SIGdial, 2013.
- [2] Jason D. Williams and Steve Young, “Partially observable Markov decision processes for spoken dialog systems,” Computer Speech and Language, 21, 393-422, 2007.
- [3] 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.
- [4] 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.
- [5] 山田 耀司, 滝口 哲也, 有木 康雄, “状態空間の分割と状態遷移の学習に基づく Parallel POMDP の評価,” 電子情報通信学会信学技報, 115(253), 39-43, 2015.
- [6] 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.
- [7] 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.
- [8] Matthew Henderson, Blaise Thomson and Steve Young, “Word-based Dialog State Tracking with Recurrent Neural Networks” Proceedings of the SIGdial, 2014.
- [9] 岩井 美樹, 二宮 崇, “word2vec に基づく述語項構造の分布表現獲得,” 言語処理学会, 2015.