対話制御の方策再利用における行動関連確率の利用

The Use of Action-Relation Probability in Policy Reuse for Dialog Management

グエン トゥン [*1*4] 吉野 幸一郎 [*2*1*3] サクティ サクリアニ [*1*3] 中村 哲 [*1*3]
Tung The Nguyen Koichiro Yoshino Sakriani Sakti Satoshi Nakamura

[*1]奈良先端科学技術大学院大学 [*2]理化学研究所 ロボティクスプロジェクト
Nara Institute of Science and Technology RIKEN Robotics Project

[*3]理化学研究所 革新知能統合研究センター
RIKEN AIP Center

Reusing policies in a new domain, which is trained on the existing domain, is an important problem of dialogue management research based on reinforcement learning. This work defines action-relation probabilities between the action spaces of the new and the target domains using mixture density networks for the reuse of policies. Experimental results showed that the proposed modeling of action-relation probabilities based on component matching using regression realized the effective policy reuse.

1. はじめに

強化学習を用いた対話制御 [Williams 08] は、特にタスクゴールが定まった対話システムの行動制御に有効であることが知られている。対話システムが対話エピソード内で経験してきた状態系列にもとづいて行動を決定する関数は**方策 (policy)** と呼ばれる。対話制御の課題の一つは、より高い報酬期待値を得られる方策を学習することである。

この中でも、既に学習されたタスク・ドメイン（既存ドメイン）の方策から、ほとんど学習データが存在しない新しいタスク・ドメイン（転移先ドメイン）における良い方策を導くことは、方策転移と呼ばれる重要な課題である [Chen 18]。転移強化学習は、対話制御の分野でも長らく研究されてきた [Gasic 13]。これは、対話システム活用が求められる場面が多様で、開発されるアプリケーション全てに対して十分な学習データを想定することが現実的ではないためである。転移強化学習の中でも、既に学習した方策を利用する手法は**方策再利用 (policy reuse)** と呼ばれる。方策再利用は、他のパラメータ適応などの手法 [Mendez 19] と比較して、少ない学習資源でタスク・ドメイン適応が可能という利点がある。

対話システムの方策転移においては、対話システムのタスク・ドメイン間で、タスクゴールやシステムの行動空間についてある程度の類似を仮定することができる。こうした場合、似た文脈では似た行動を行うことで、既存ドメインでの方策を用いた場合でも転移先ドメインにおける適切な行動選択ができることが期待される。例えば、レストラン案内ドメインにおける予約の行動と、ホテル案内ドメインにおける予約の行動は、類似する文脈上で呼び出されることが多い。こうした行動同士の類似性を考えることで、既存ドメインにおける方策の一部を、転移先ドメインの方策に引き継ぐことができる。我々はこの点に着目し、対話システムの行動同士の関連確率を定義して用いる Dialog Policy Reuse Algorithm (DPRA) を提案する。

連絡先: Tung The Nguyen, tung.nguyen at is.honda-ri.com
吉野 幸一郎, koichiro.yoshino at riken.jp,
Sakriani Sakti, ssakti at is.naist.jp,
中村 哲, s-nakamura at is.naist.jp
[*4] 現在 HRI-JP 勤務

2. 対話制御における方策転移

本節では、研究の前提となる基本的な強化学習の考え方と、これを用いた対話制御について説明する。また、方策転移の問題設定についても説明する。

2.1 強化学習を用いた対話制御

強化学習を用いた対話制御では、ある時刻 t における観測状態とシステムの行動、与えられる報酬をそれぞれ $s_t \in S$、$a_t \in A$、$r_t \in R$ とするマルコフ決定過程を想定する [Bellman 57]。状態、行動、報酬はそれぞれ対話システムが現在得ている情報（言語理解のどのスロットがどの値で埋まっているかなど）、それに対応して取りうる行動（未知のスロット値に対する問合せなど）、これらが行われた場合の報酬（対話成功判定や時間経過のペナルティなど）に対応する。状態遷移 $s_{t+1} \leftarrow \{s_t, a_t\}$ に与えられる報酬が r_t であり、対話終了時 T に至るまでの報酬の合計が $R_t = \sum_{k=1}^{T-k} r_{t+k}$ として与えられる。方策 π を状態 s_t が与えられた時に行動 a_i を出力する関数、あるいは確率モデルとして与えたとき、報酬の期待値は $E(R_t|s_t, \pi)$ と記述できる。この報酬期待値を最大化するような方策を求めることが強化学習の目的である。本研究では、各状態における各行動の選択確率 $\pi(s, a)$ を学習する。

2.2 方策転移の課題

方策転移とは、あるドメインで学習した方策を別のドメインに転移させることを指す。対話制御を含む強化学習を用いる問題の多くでは、用意できる学習データの量に対して、状態と行動の組み合わせ空間が非常に膨大である [Litman 00]。そこで、既に学習したドメインの方策を異なるドメインに転移することが試みられてきた。特に深層強化学習における多くの方策転移の手法では、既存ドメインで学習された方策を、転移先ドメインにおける方策の重み初期値として適応を行う [Mendez 19]。しかしこうした方式は、既存ドメインと転移先ドメイン間の知識を仮定しているものが多い。また、適応のための学習データがある程度必要となる。そこで本研究では、既存ドメインと転移先ドメインにおける行動同士の関連について確率モデルを定義し、既存ドメインで学習された方策を転移先ドメインでそのまま利用する（**再利用する**）ことを目標とする。この際、既存ドメインと転移先ドメインとの関係は明示的

に与えないモデルの構築を目指す。

3. 行動関連確率に基づく方策再利用

既存ドメインにおける状態と行動を $s \in S_{src}, a \in A_{src}$、転移先ドメインにおける状態と行動を $s \in S_{tgt}, b \in A_{tgt}$ としたとき、方策再利用の問題は次式のように定義できる。

$$\pi(s,b) = \sum_{a \in A_{src}} P(b|a,s)\pi(s,a) \tag{1}$$

ただし、ここで状態 s は既存ドメインと転移先ドメインの状態集合の和集合である $S_{src} \cup S_{tgt}$ から与える。ここで、$P(b|a,s)$ を求めることができれば、転移先ドメインで既存ドメインにおける方策 $\pi(s,a)$ を再利用することができる。本論文ではこの確率を行動関連確率と呼ぶ。

3.1 混合密度ネットワークによるモデル化

ここで、既存ドメインにおけるマルコフ決定過程上の状態遷移確率に転移先ドメインにおける行動 b を導入すると、

$$P(s'|a,s) = \sum_{b \in A_{tgt}} P(b|a,s)P(s'|b,a,s), \tag{2}$$

と式変形できる。ただし、s' は s の次の時刻における状態を表す。この式から、既存ドメインにおける状態遷移は、行動関連確率で重みづけされた混合分布モデルによって表現できることがわかる。そこで本研究では、この状態遷移を混合密度ネットワーク [Bishop 94] でモデル化することを考える。式 (2) を各行動 a_i、b_j が与えられた場合に対応させ、構成要素を混合分布モデルで置き換えると、

$$\begin{aligned} P(s'|a_i,s) &= P_i(s'|s) \\ &= \sum_{j=1}^{|A_{tgt}|} P_{ij}(s) \cdot p_{ij}(s'|s) \\ &= \sum_{j=1}^{|A_{tgt}|} w_{ij}(s) \cdot \mathcal{N}(s'; \mu_{ij}(s), \sigma_{ij}^2(s)), \end{aligned} \tag{3}$$

のように記述することができる。つまり、$p_{ij}(s'|s)$ は平均を $\mu_{ij}(s)$、分散を $\sigma_{ij}^2(s)$ とするガウス分布から得られる状態遷移確率 $p(s'|a_i,b_j,s)$ となる。この混合密度ネットワークにおけるパラメータ $w_{ij}(s)$ は、学習データから得られる遷移から、

$$p(s'|s) = \sum_{m=1}^{M} w_m(s) \cdot \mathcal{N}(s'; \mu_m(s), \sigma_m^2(s)), \tag{4}$$

$$L = -\log(\prod_{i=1}^{N} p(\sum_{m=1}^{M} w_m(s_i) \cdot \mathcal{N}(s_i'; \mu_m(s_i), \sigma_m^2(s_i)))) \tag{5}$$

の対数尤度を最大化することで得られる。

3.2 コンポーネントマッチングによるモデル適応

ただし、既存ドメインのデータのみでは転移先ドメインにおける行動 b_j に関する知識が与えられないため、状態遷移確率 $p(s'|a_i,b_j,s)$ を正しく求めることが難しい。そこで、少量の転移先ドメインにおける状態遷移のサンプル $s' \leftarrow \{s, b_j\}$ が得られる状況で、コンポーネントマッチングを行う場合を考える。つまり、$\mathcal{N}(s'; \mu_{ij}(s), \sigma_{ij}^2(s))$ をデータから得られた

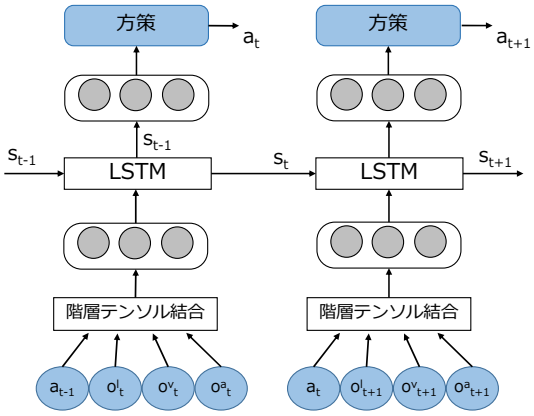

図 1: 実験に用いた end-to-end 対話制御

$p(s'|b_j,s)$ に対応させる。このコンポーネントマッチングのため、本研究では 2 種類の手法を試行する。

一つ目の手法では、

$$p(s'|a_i,b_j,s) = p(s'|b_j,s) \forall a_i \in A_{src}, b_j \in A_{tgt} \tag{6}$$

を仮定した上で、転移先ドメインにおけるデータから単純に混合分布モデルのパラメータを対数尤度最大化する。つまり、

$$L = -\log(\prod_{i=1}^{N} \mathcal{N}(s'; \mu_{ij}(s), \sigma_{ij}^2(s)), \tag{7}$$

の対数尤度を最大化する。このモデルを DPRA-reg. と呼ぶ。

二つ目の手法では、$p(s'|b_j,s)$ もまた混合密度ネットワークによって表現されるという考えのもと、

$$P(s'|b,s) = \sum_{a \in A_{src}} P(a|b,s)P(s'|b,a,s), \tag{8}$$

のパラメータを求める。ここで、式 (2) で定義された元の混合密度ネットワークとの相違は重み $P(a|b,s)$ なので、式 (4) と同様にこの重みを以下の式の対数尤度最大化によって求める。このモデルを DPRA-MDN と呼ぶ。

4. 評価

提案した方策再利用の手法について、実際の対話制御の学習において評価を行った。この際、状態・行動空間が類似するタスク間で方策再利用を行った場合と、状態・行動空間が大きく異なるタスク間で方策再利用を行った場合の評価をそれぞれ行った。以下に詳細な実験設定と結果について示す。

4.1 実験設定

4.1.1 統一の実験設定

まず、いずれの場合も対話制御のモデル化には end-to-end の枠組み [Zhao 16] を用い、学習データから MAP 推定によって学習されたユーザシミュレータを用いて DDQN[Van Hasselt 16] で学習を行った。この手法では、図 1 に示すように、入力として発話中の言語特徴（o_t^l）、音響特徴（o_t^a）、顔画像から得られる画像特徴（o_t^v）の生データを入力として用い、出力として行動を生成する。この際、状態を表す中間層（s_t）は 128 次元に設定した。ネットワークの学習には Adam[Kingma 14] を用いた。学習率は初期値を $1e-3$ とし、1,000 エピソードごとに 10%ずつ減少させた。また、方策転移のベースライン

手法として、行動出力層の直前までのパラメータを事前学習として、用いるモデルを利用した [Mendez 19]。

DPRA の各アルゴリズムでは、ニューラルネットワークを用いて各モデルの混合変数を近似予測した。このネットワークの隠れ層は 1 層 256 次元とし、学習率は $1e-4$、Adam を用い、学習エポック数は 10 とした。

4.1.2 類似タスクでの実験設定

類似する対話タスクとして、健康相談に関する対話データ [Nguyen 20] を用いた。この対話ドメインでは、対話システムはユーザに健康に関する提案を行い、ユーザに受理して貰うことを目的とする。この実験では、同じ対話ドメインの対話データを 51 対話の既存ドメイン対話と 24 対話の転移先ドメイン対話の 2 つに分割して利用した。なお、評価に際しては転移先ドメイン対話の一部を適応学習に用い、残りを評価データとして用いた。ただし、既存ドメインでは対話システム側が用いることができる行動クラスは Offer, Framing, End の 3 種類のみだが、転移先ドメインでは対話システムが用いることができる行動クラスが Offer-New, Offer-Change, Framing-Argue, Framing-Answer, End の 5 種類となるように変更を加えた。つまり、既存ドメインで定義されていた対話システムの行動クラスが、転移先ドメインでは細分化されて再定義される状況を考える。いずれも入力特徴量は、ユーザの顔画像、音声、書き起こしを用いた。学習においては、既存ドメインで 20,000 エピソード、ベースラインにおいては転移先ドメインで 2,000 エピソードのシミュレーションを行った。顔画像からは OpenFace ツールキット [Baltrušaitis 16] によって 20 種類のアクションユニットを抽出した。音声からは IS2009 emotion challenge standard feature-set [Schuller 09] に従い、OpenSMILE ツールキット [Eyben 10] で抽出した特徴を用いた。また、テキストは音声から得た書き起こしを用いた。報酬設計は元研究の通り、適切な行動の組み合わせで+10、それ以外の組み合わせで-10、適切なタイミングで対話を終了することで+100、それ以外のタイミングで対話を終了することで-100 とした。

4.1.3 異なるタスクでの実験設定

異なるタスク間の実験では、既存ドメインとして健康相談に関する対話データを、転移先ドメインとして感情誘発対話データ [Lubis 18] を用いた。この実験では、転移先ドメインにおける対話システムの目的は健康相談と異なり、ユーザの感情状態をポジティブに変更することである。この設定では、既存ドメインの対話データは健康相談データの 75 対話全てを用いた。転移先ドメインの対話データは感情誘発対話の 58 対話である。なお、評価に際しては転移先ドメイン対話の一部を適応学習に用い、残りを評価データとして用いた。また、健康相談対話で対話システム側が用いることができる行動クラスが 5 種類に対して、感情誘発対話では 9 通りの行動クラスが定義される。この中で健康相談対話と重複する行動は End のみである。学習においては、既存ドメインで 20,000 エピソード、ベースラインにおいては転移先ドメインで 5,000 エピソードのシミュレーションを行った。これは、感情誘発対話データの対話長が健康相談対話と比較して長く（1 対話あたり 23.6 分）、適応に時間が掛かると予想されたためである。本ドメインでは、ユーザの顔画像は用いず、ユーザの音声のみから OpenSMILE で特徴量抽出を行い入力として用いた。また、ユーザ発話の書き起こしを用いた。報酬設計は、適切な行動の組み合わせで+10、それ以外の組み合わせで-10、適切なタイミングで対話を終了することで+100、それ以外のタイミングで対話を終了することで-100 とした。

表 1: 類似タスクにおける学習時間

手法	学習時間
DPRA-MDN	23s
DPRA-reg.	20s
Adapt-2k	320s

表 2: 異なるタスクにおける学習時間

手法	学習時間
DPRA-MDN	310s
DPRA-reg.	280s
Adapt-5k	7000s

4.2 実験結果

以下に類似タスク、異なるタスクそれぞれでの方策再利用の実験結果について示す。評価においては、学習に要した時間と、テストエピソードに対して学習された方策を適用した場合の行動選択の正確さを用いた。後者は、該当ターンまでの対話コンテキストが正しく与えられたとして、人手でアノテーションされた最適な行動を方策が選ぶことができた割合である。

4.2.1 計算時間での評価

まず、計算速度での評価を表 1、2 に示す。計算速度の面では、いずれも提案する DPRA に基づく方策再利用が、単なる方策のパラメータ適応を大きく上回る結果となった。DPRA の中では、単純に混合分布モデルのパラメータを更新するモデルの方が、転移先ドメインにおける混合密度ネットワークのパラメータを考慮するモデルよりも高速となった。なお、計算資源としてはいずれも同じ計算機 [*1] を用いた。

4.2.2 行動選択精度での評価

次に、学習された方策を用いた場合の適応先対話ドメインでの行動選択精度について表 3、4 に示す。なお、各実験はいずれも 50 回パラメータ学習を試行したモデルを用いた評価結果の平均を用いた。括弧内は 95%信頼区間の範囲である。また、横方向の対話数は転移先ドメインに対するパラメータ適応やパラメータ学習に利用可能な対話データ数である。これらは各回でランダムに決定し、残りを評価データとした。

まず類似タスクでの実験から、提案する方策再利用が、ベースラインとして用いた単なるニューラルネットワークのパラメータ適応よりも高い精度を実現できることがわかる。これは、行動同士の類似性をうまく捉えることで、既存ドメインで学習された方策を効率的に利用できたことが示唆される。適応に用いるデータ量が増えるに従って、いずれの手法の精度も向上した。この設定では、既存ドメインと転移先ドメインは異なる行動セットで同じタスクを解いていることも、学習が上手くいった理由として考えられる。

次に、異なるタスクでの実験では、方策再利用に基づく手法とパラメータ適応に基づく手法の差異は小さく、大きな差がないことがわかる。対話タスク自体が健康相談対話と感情誘発対話では大きく異なるため、既存ドメインで得られた方策やパラメータがあまり効果的に利用されなかった可能性がある。また、これらのタスクはいずれもマルチモーダル入力を用いるものであるが、方策再利用には入力の種類よりも、定義された対話システムの行動空間や対話ゴールにおける類似性が重要であることが示唆される。

∗1 GPU: GTX Titan X

表 3: 類似タスクにおける行動選択精度

#対話数	1	2	4	8	16
DPRA-MDN	40.96% (±7.79%)	39.86% (±6.81%)	56.16% (±7.82%)	62.05% (±4.94%)	62.05% (±4.64%)
DPRA-reg.	**45.41%** (±5.99%)	**52.79%** (±5.18%)	**60.48%** (±6.02%)	**62.98%** (±5.08%)	**69.30%** (±3.22%)
Adapt-2k (baseline)	33.83% (±4.88%)	33.25% (±4.73%)	41.41% (±5.92%)	39.41% (±5.75%)	43.64% (±5.41%)

表 4: 異なるタスクにおける行動選択精度

#対話数	1	2	4	8	16	32
DPRA-MDN	21.94% (±5.35%)	24.79% (±6.59%)	24.03% (±6.12%)	30.14% (±4.07%)	19.27% (±4.84%)	**31.09%** (±5.64%)
DPRA-reg.	23.79% (±7.40%)	**28.46%** (±7.07%)	23.33% (±5.45%)	29.59% (±7.31%)	**43.37%** (±2.19%)	26.59% (±5.27%)
Adapt-5k (baseline)	**31.49%** (±4.48%)	28.12% (±7.21%)	**37.53%** (±5.05%)	**32.91%** (±6.50%)	23.22% (±7.10%)	23.00% (±8.49%)

これらの結果から、方策再利用を用いる場合は既存ドメインと転移先ドメインの類似性をある程度仮定することが重要であること、類似性がある程度ある場合は少ない転移先ドメインのデータしかない場合でも効率的な方策再利用が可能であることがわかった。

5. まとめ

本論文では、強化学習を用いた対話制御において、既存の対話タスク・ドメインで学習した方策を、異なる対話タスク・ドメインで利用するための方策再利用に取り組んだ。この方策再利用のため、対話ドメイン間の行動関連確率を混合密度ネットワークによってモデル化した。実験の結果、提案手法は既存のパラメータ適応手法よりも、効果的な既存ドメインの方策再利用・方策転移が可能であることがわかった。

今後は、より様々な対話ドメインでの検証を行うこと、特に一般的なタスク対話システムにおける検証を行う必要がある。

謝辞

本研究は JSPS 科研費 20H05567 および JP17H06101 の助成を受けたものです。

参考文献

[Baltrušaitis 16] Baltrušaitis, T., Robinson, P., and Morency, L.-P.: Openface: an open source facial behavior analysis toolkit, in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–10IEEE (2016)

[Bellman 57] Bellman, R.: A Markovian decision process, *Journal of mathematics and mechanics*, Vol. 6, No. 5, pp. 679–684 (1957)

[Bishop 94] Bishop, C. M.: Mixture density networks (1994)

[Chen 18] Chen, L., Chang, C., Chen, Z., Tan, B., Gašić, M., and Yu, K.: Policy adaptation for deep reinforcement learning-based dialogue management, in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6074–6078IEEE (2018)

[Eyben 10] Eyben, F., Wöllmer, M., and Schuller, B.: Opensmile: the munich versatile and fast open-source audio feature extractor, in *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462 (2010)

[Gasic 13] Gasic, M., Breslin, C., Henderson, M., Kim, D., Szummer, M., Thomson, B., Tsiakoulis, P., and Young, S.: POMDP-based dialogue manager adaptation to extended domains, in *Proceedings of the SIGDIAL 2013 Conference*, pp. 214–222 (2013)

[Kingma 14] Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, *arXiv preprint arXiv:1412.6980* (2014)

[Litman 00] Litman, D., Kearns, M. S., Singh, S., and Walker, M.: Automatic optimization of dialogue management, in *COLING 2000 Volume 1: The 18th International Conference on Computational Linguistics* (2000)

[Lubis 18] Lubis, N., Sakti, S., Yoshino, K., and Nakamura, S.: Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32 (2018)

[Mendez 19] Mendez, J. A., Geramifard, A., Ghavamzadeh, M., and Liu, B.: Reinforcement learning of multi-domain dialog policies via action embeddings, 3rd Workshop Conversational AI, Today ' s Pract. Tomorrow ' s Potential NIPS (2019)

[Nguyen 20] Nguyen, T. T., Yoshino, K., Sakti, S., Nakamura, S., et al.: Dialog management of healthcare consulting system by utilizing deceptive information, *Transactions of the Japanese Society for Artificial Intelligence*, Vol. 35, No. 1, pp. DSI–C_1 (2020)

[Schuller 09] Schuller, B., Steidl, S., and Batliner, A.: The interspeech 2009 emotion challenge, in *Tenth Annual Conference of the International Speech Communication Association* (2009)

[Van Hasselt 16] Van Hasselt, H., Guez, A., and Silver, D.: Deep reinforcement learning with double q-learning, in *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 30 (2016)

[Williams 08] Williams, J. D., Poupart, P., and Young, S.: Partially observable Markov decision processes with continuous observations for dialogue management, in *Recent Trends in Discourse and Dialogue*, pp. 191–217, Springer (2008)

[Zhao 16] Zhao, T. and Eskenazi, M.: Towards End-to-End Learning for Dialog State Tracking and Management using Deep Reinforcement Learning, in *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 1–10 (2016)