対話制御の方策再利用における行動関連確率の利用

グエン トゥン, 吉野 幸一郎, サクティ サクリアニ, 中村 哲

奈良先端科学技術大学院大学

理化学研究所ガーディアンロボットプロジェクト (GRP)

理化学研究所革新知能統合研究センター (AIP)

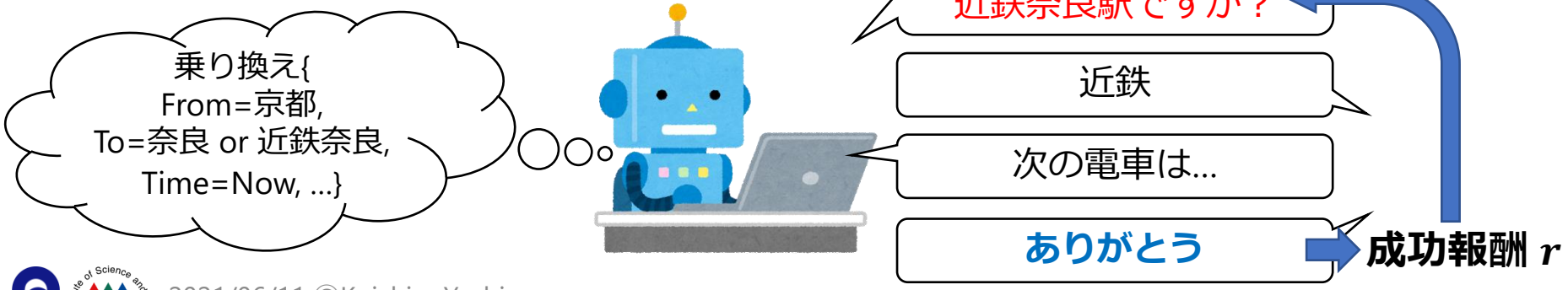
※第一著者は現在 Honda Research Institute Japan 勤務

対話システムの行動決定問題

◆対話システムは将来の報酬を考えて行動する必要

- 音声認識・言語理解に由来する曖昧性に対しての聞き返し
  - 聞き返しが正解かどうかはその時点ではわからない
- 話題選択やモジュール選択の意思決定についても同様

強化学習を利用した
報酬期待値の最大化



強化学習を用いた対話制御

◆対話のプロセスをマルコフ決定過程で定式化

- 時刻 t における、観測状態: s_t, システムの行動: a_t, 報酬: r_t
- 対話終了時 T に至るまでの報酬の合計 $R_t = \sum_{k=1}^{T-k} r_{t+k}$ を最大化
- 状態は状態遷移確率に沿って変化: $p(s_{t+1}|a_t, s_t)$
- 各観測状態における行動の選択確率（方策）: $\pi(s, a) = p(a|s)$
- システムの振る舞いが方策 π に従う場合
  報酬の期待値は $E(R_t|s_t, \pi)$ → これを最大化する π を求める

◆強化学習で方策を導出

- 学習データやシミュレータに基づくいくつかの手法
  - Q-learning, Deep Q-network, 方策勾配法, Actor-Critic, …

いずれの方法も一定量の学習データが必要

制御の対話ドメイン適応

◆新しい対話ドメインで制御を構築したい

問題: 構築のために一定量の学習データが必要

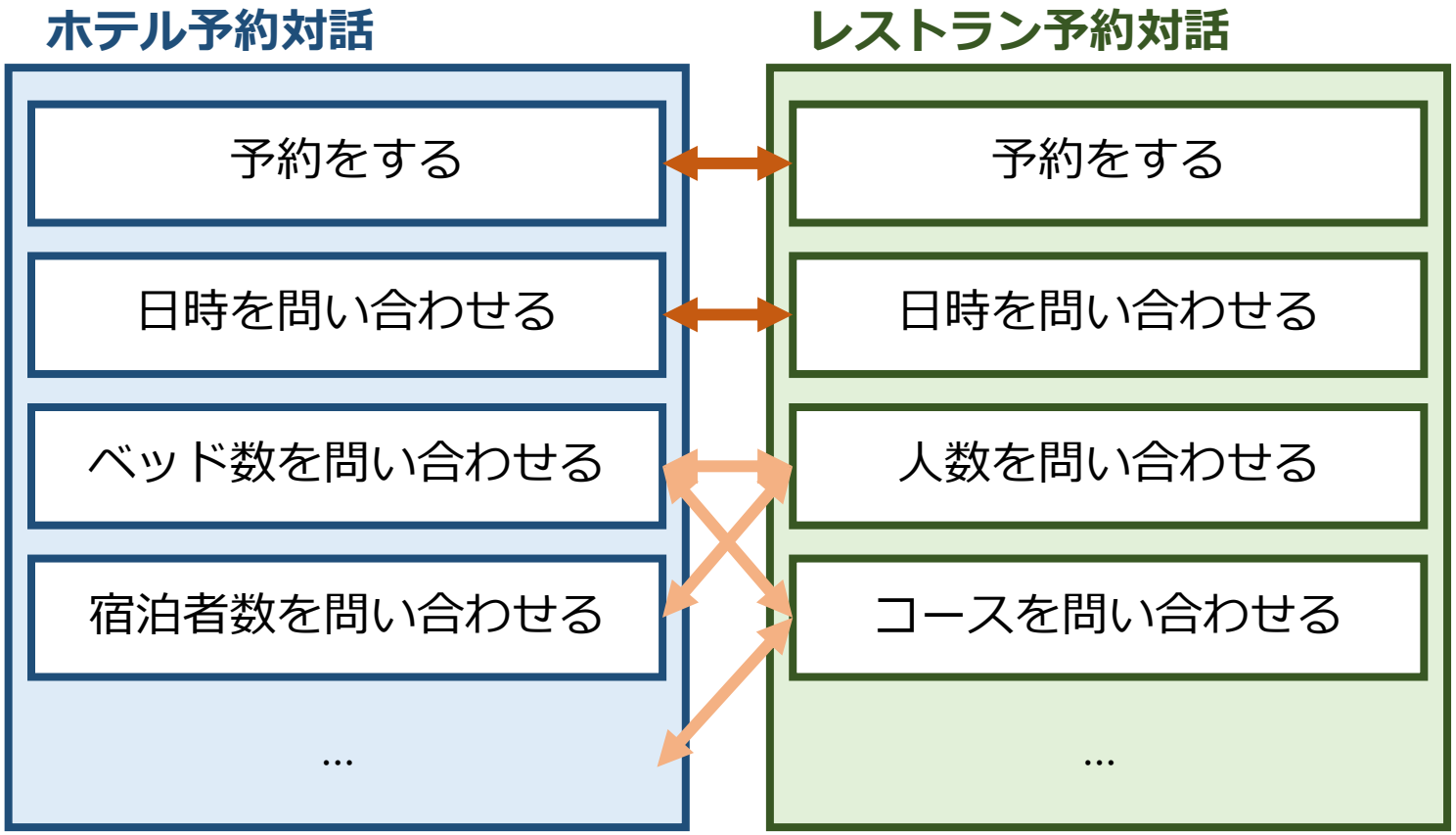
解法: 学習した既存ドメインの方策やデータをうまく使う

◆既存ドメインの方策の利用方法

- パラメータ適応:
  既存の方策を初期値として少量の学習データで適応
- **方策再利用:**
  既存の方策をそのまま使う方法を考える
  → うまく使えれば少ない学習資源でドメイン適応が可能
  （Dialogue Policy Reuse Algorithm; DPRA）

方策再利用（DPRA）のアイディア

◆元ドメインと転移先ドメインの行動の類似性を利用



方策における行動関連確率

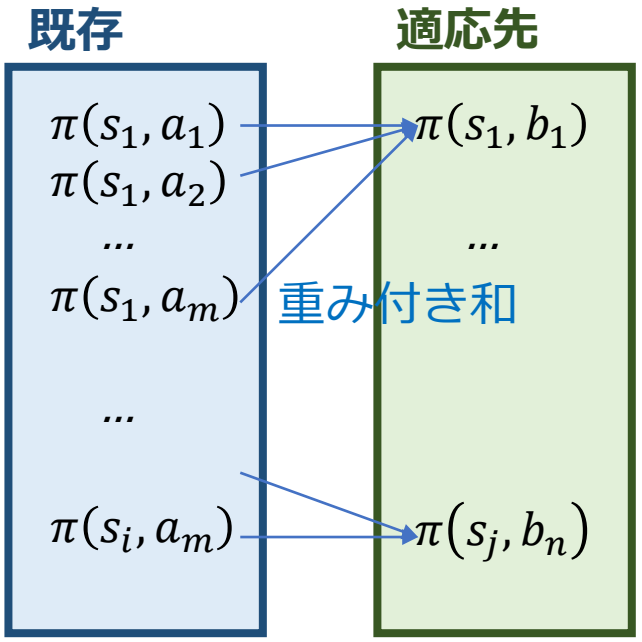
◆ $\pi(s,b) = p(b|s) = \sum_a p(b|a,s)p(a|s) = \sum_a p(b|a,s)\pi(s,a)$

- a は既存ドメインの行動、 b は適応先ドメインの行動
- ただし観測情報の空間は同じ

◆類似する文脈で類似する行動を取る方策の重み付き和を求める

- 行動類似度と既存の方策を考慮
- 類似する行動に似た行動をとる

どうやって行動類似度を求めるか？



混合密度ネットワークと行動類似度

◆マルコフ決定過程上の状態遷移確率に適応先の行動を導入

$$p(s'|a,s) = \sum_{b} p(b|a,s)p(s'|b,a,s)$$

- ただし s' は s の次の時刻の観測状態

◆各行動 a_i, b_j が与えられた場合状態遷移確率は

$$p(s'|a_i,s) = p_i(s'|s) = \sum_{j=1}^{|B|} p_{ij}(s)p_{ij}(s'|s)$$

◆これを混合分布モデルで置き換え

$$\sum_{j=1}^{|B|} w_{ij}(s)N(s';\ \mu_{ij}(s), \sigma_{ij}^2(s))$$

目的関数

$$\sum_{j=1}^{|B|} w_{ij}(s) N(s';\ \mu_{ij}(s), \sigma_{ij}^2(s))$$

◆混合密度ネットワークにおけるパラメータ w_{ij} を求めたい

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

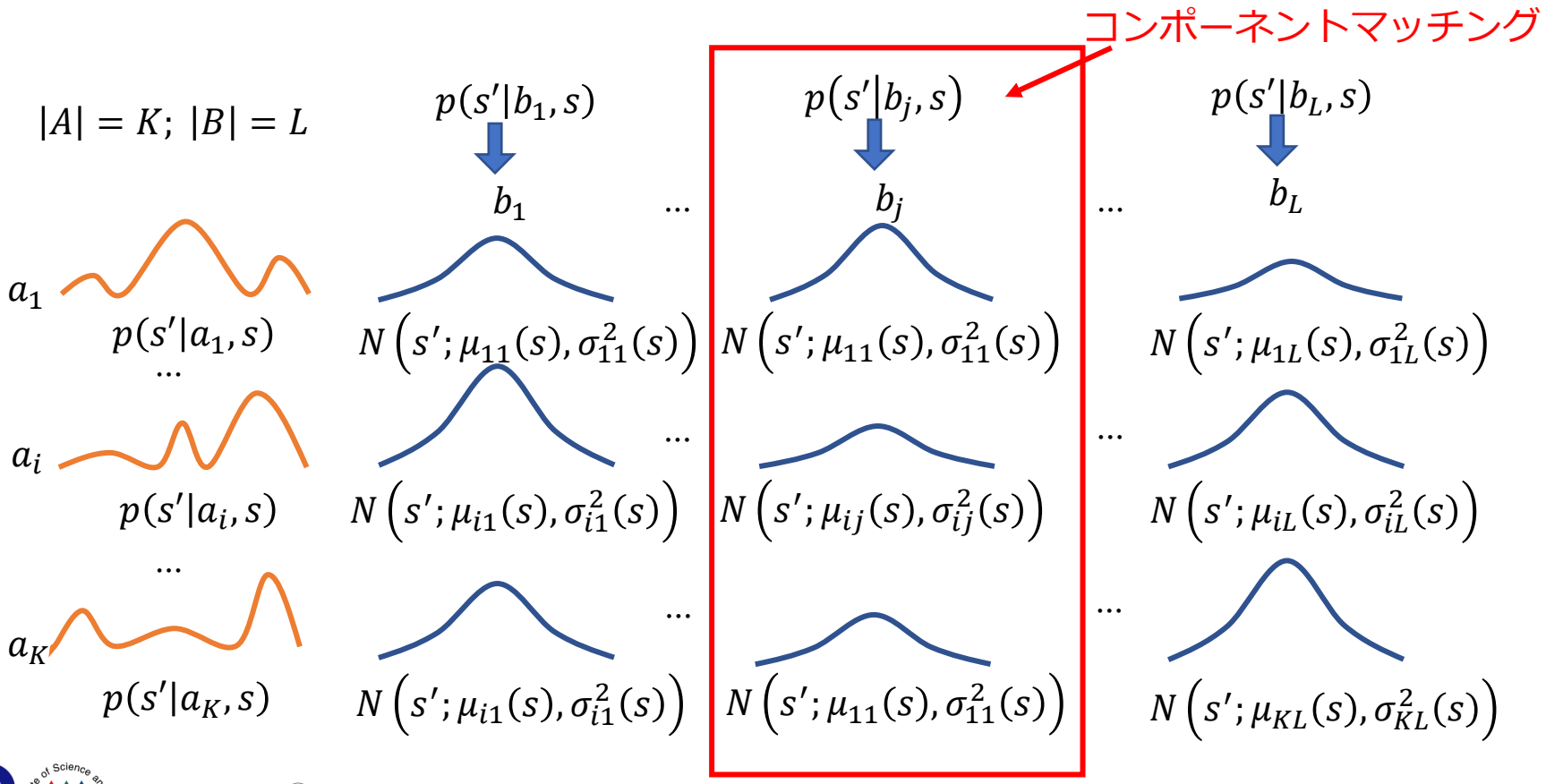
$$L = -\log(\prod_{i=1}^{N} p(\sum_{m=1}^{M} w_m(s_i) N(s'; \mu_m(s), \sigma_m^2(s))))$$

適応先ドメインデータを用いた対数尤度最大化で求める

混合密度ネットワークと行動類似度

◆状態遷移確率を混合密度ネットワークで表す

- ●データにあてはまりのよいパラメータで重み付き和を取る



コンポーネントマッチング

◆少量の転移先ドメインデータを用いる

- 少量の $s' \leftarrow \{s, b_j\}$ からコンポーネントマッチングを行う
- つまり $N(s';\ \mu_{ij}(s), \sigma_{ij}^2(s))$ と $p(s'|b_j, s)$ の対応を取る

◆2つの方法

- $p(s'|b_j, a_i, s) = p(s'|b_j, s)$ を仮定してパラメータを求める
  - 直接的にパラメータを回帰で求める（識別モデル的手法）
  - **DPRA-reg.**
- $p(s'|b_j, s)$ の混合密度ネットワークから求める
  - 混合密度ネットワークを仮定し尤度最大化を行う（生成モデル的手法）
  - **DPRA-MDN**

評価における仮説

◆**提案法のメリット**

- 提案する方策再利用は**少量の学習データで高精度**に適応が可能
  - 単なるパラメータ適応手法との比較
- **計算コスト**の上でのメリット（計算時間の評価）
  - 深層強化学習のパラメータ適応は高コスト

◆**比較手法**

- 深層強化学習における方策パラメータ適応 [Mendez 19]: **ActEmb** (適応に5,000、25,000エピソード)
- 適応なし: **NoAdapt** (学習に5,000、25,000エピソード)

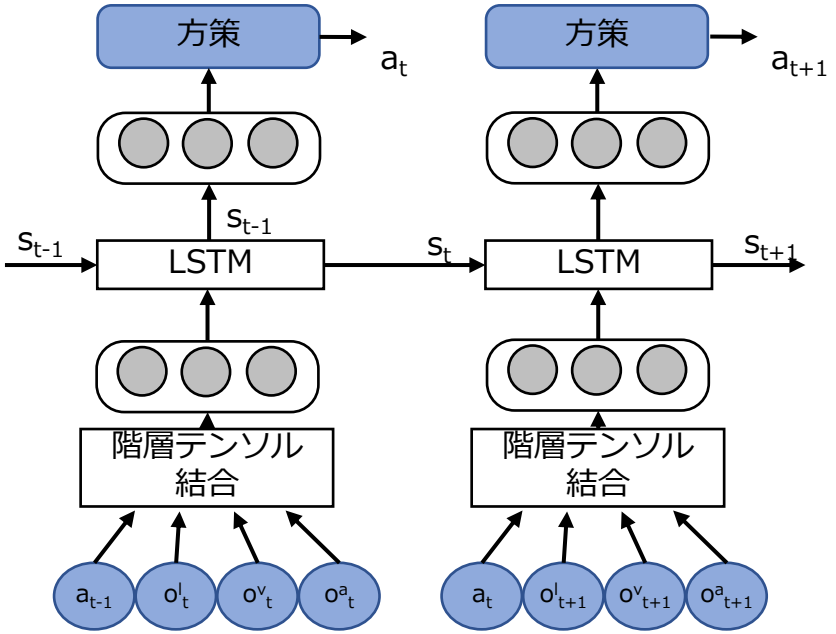
◆**評価指標**

- **行動選択精度**: 求めた方策に従った場合の行動決定の精度
- **計算速度**: GTX1080 で方策学習の収束にかかった時間

実験条件

◆対話制御

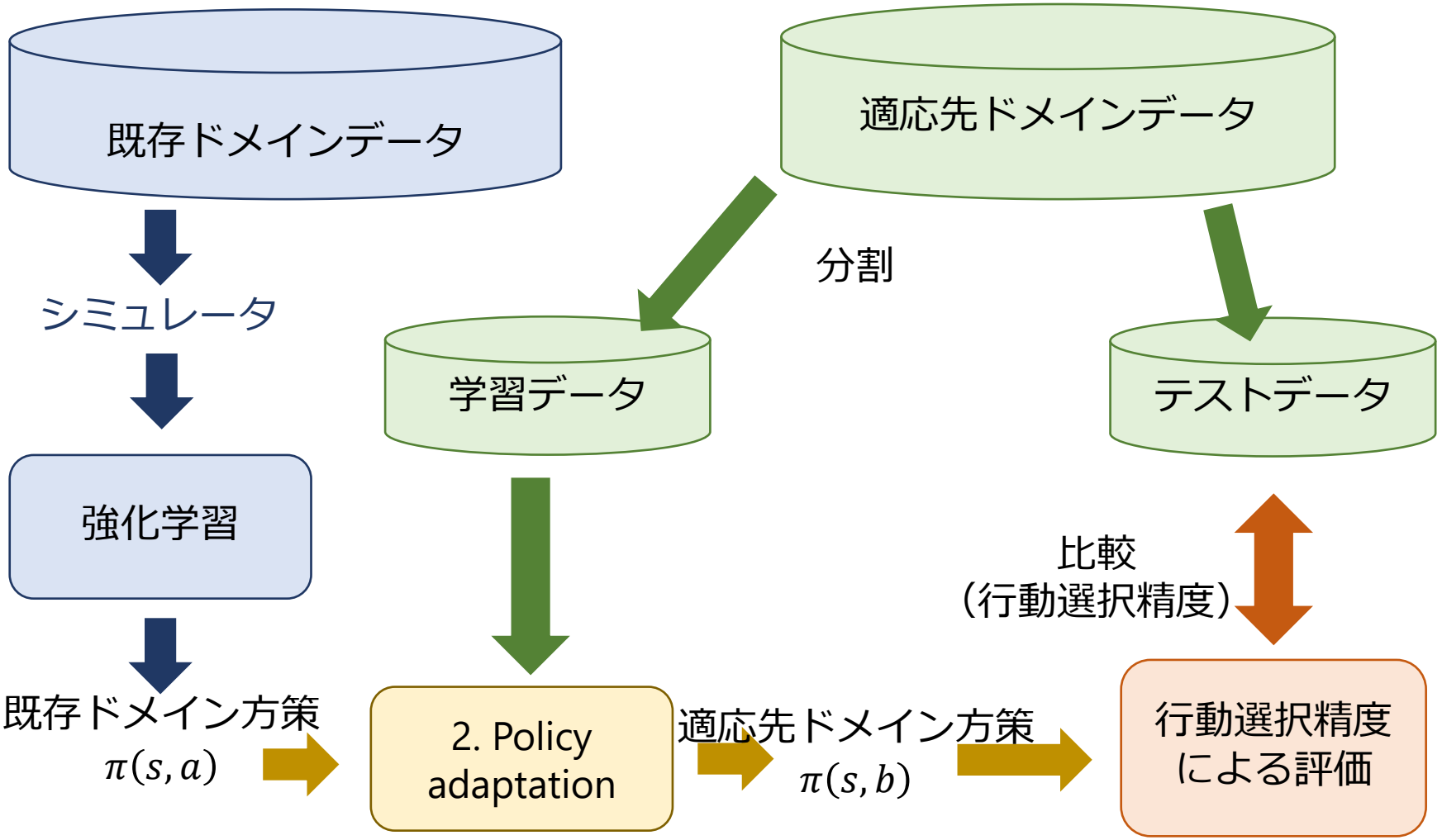
- Deep Q-network と
  階層テンソル結合に基づく
  end-to-end 対話制御



◆対話データ

- **元ドメインデータ**
  - 健康相談対話データ
    [Nguyen 20]
- **適応先ドメインデータ**
  - **類似:** 健康相談対話データ [Nguyen 20]
    ただし異なる行動クラス
  - **異なる:** 感情誘発対話データ [Lubis 18]

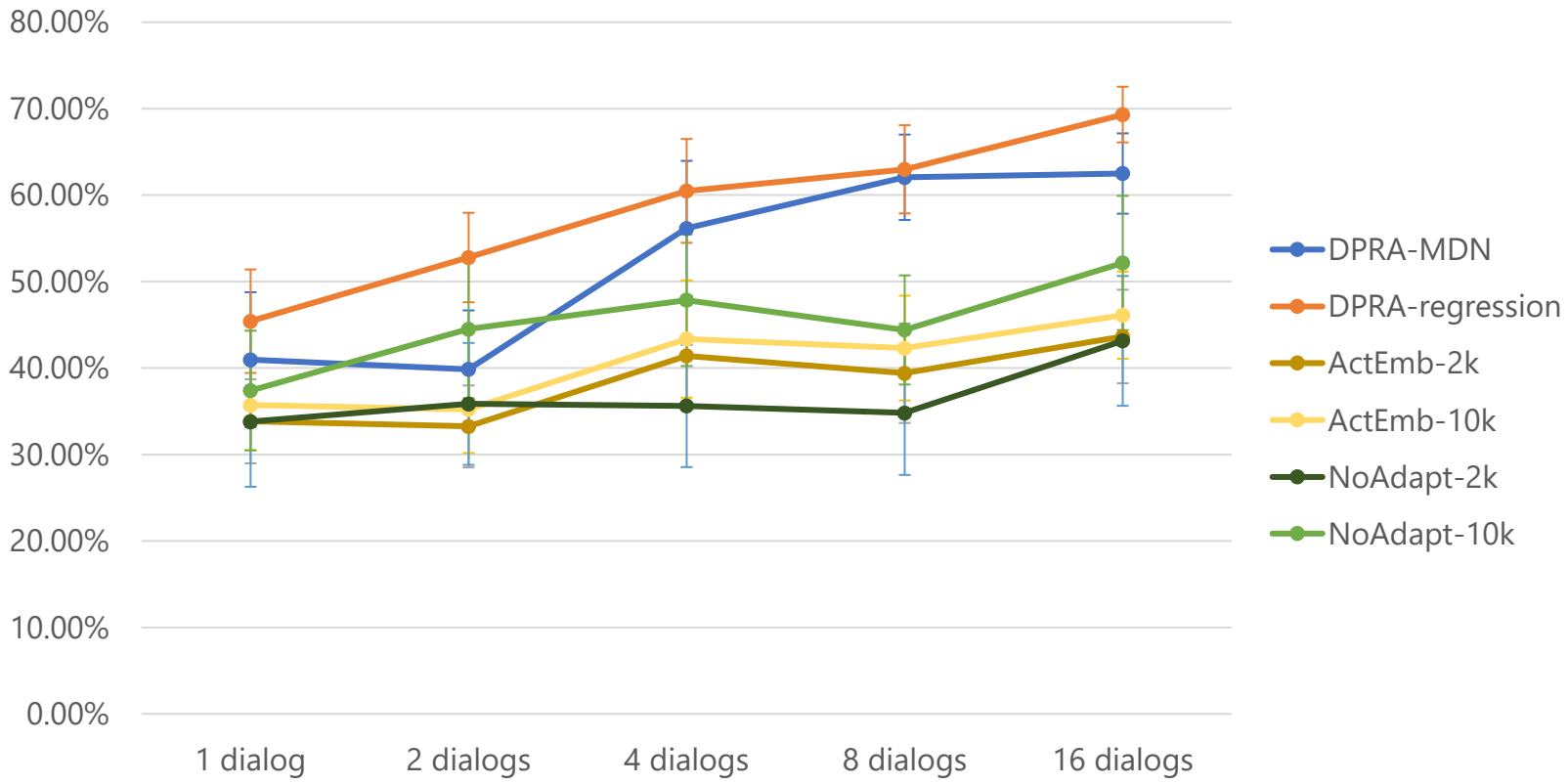
評価実験


既存ドメインデータ
適応先ドメインデータ
シミュレータ
分割
学習データ
テストデータ
強化学習
比較
（行動選択精度）
既存ドメイン方策
$\pi(s,a)$
2. Policy adaptation
適応先ドメイン方策
$\pi(s,b)$
行動選択精度による評価

行動選択精度（類似ドメイン）

◆提案手法（DPRA）がより良い適応精度を実現

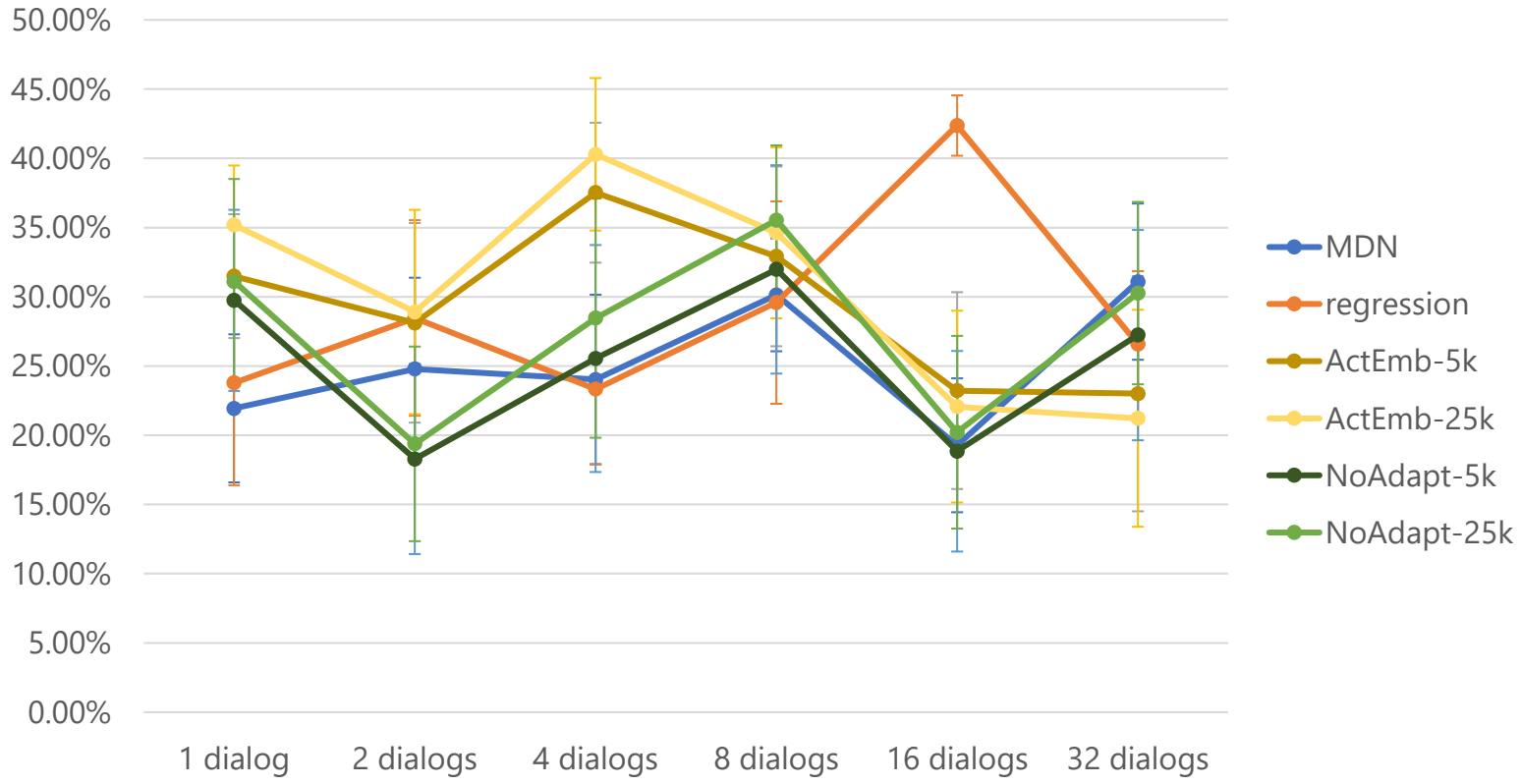
- コンポーネントマッチングは Reg. の方が MDN より良い



行動選択精度（異なるドメイン）

◆提案法（DPRA）と既存の適応法に大きな違いが見られず

- ●行動セットが大きく異なる場合は適応が難しい



計算時間での評価

◆モデルの学習に要した計算時間

- 計算に用いたのはいずれも GTX1080 (Pascal世代)

手法	学習時間
DPRA-MDN	310s
DPRA-reg.	280s
ActEmb-5k	7,000s

◆提案法は単なる深層学習のパラメータ適応より低コスト

- 学習されるネットワークはより単純

まとめ

◆対話制御の問題で行動関連確率を用いた方策再利用を提案

- 適応に必要なデータの低減
- 適応自体の計算コストの低減

◆実際の対話データを用いた評価

- 適応に必要なデータ・計算コストを低減可能
- ただしある程度のドメイン同士の類似が必要

◆今後の課題

- 適応が効果的に働くケースの検討
- 特に既存のタスク対話（MultiWOZなど）での評価