1O2-OS-18-10

順序学習に基づく逆強化学習による対話制御

Dialog Control via Preference-learning based Inverse Reinforcement Learning

杉山弘晃 Hiroaki Sugiyama　目黒豊美 Toyomi Meguro　南泰浩 Yasuhiro Minami

NTT コミュニケーション科学基礎研究所
NTT Communication Science Laboratories

Dialog systems that realize dialog control with reinforcement learning (RL) have recently been proposed. However, RL has an open problem: it requires a reward function that is difficult to set appropriately. To automatically set the appropriate reward function, we propose preference-learning based inverse reinforcement learning (PIRL) that estimates a reward function from dialog sequences and their pairwise preferences, which are calculated with annotated ratings to the sequences. Inverse reinforcement learning (IRL) finds a reward function with which a system generates similar sequences as the training ones. This indicates that current IRLs suppose that the sequences are equally appropriate for a given task; thus, it cannot utilize the ratings. In contrast, our PIRL can utilize the pairwise preferences of the ratings to estimate the reward function. We examine the advantages of PIRL by comparing competitive algorithms that have been widely used to realize dialog control. Our experiments show that our PIRL outperforms the other algorithms and has the potential to be an evaluation simulator of dialog control.

1. はじめに

対話システム研究における重要なトピックの一つに対話制御がある．対話制御とはあるユーザ状態における適切なシステム行動を決定する問題であり，従来は人手で定義されたルールを用いて対話制御を実現していた．しかしこのルールの数が増大すると，ルール間の一貫性を保つことが難しくなる．

この問題を解決するため，強化学習を対話制御へ適用する研究が進められている [Williams 07, Meguro 10]．強化学習とは，人手で設計された報酬関数に基づき，将来獲得できる報酬の期待値が最大となるシステム行動を自動的に決定する手法である．報酬関数が対話の目的を適切に表現している場合，強化学習は対話の目的に対して最適なシステム行動を自動的に決定することができる．この報酬関数を設計する難しさは，対話の種類によって異なる．目的が明確なゴールで表現される対話（ゴール指向対話）では，ゴールに高い報酬を，それ以外に負の報酬を設定すればよい場合が多く，報酬関数の設計は比較的容易である [Williams 07]．しかし，明確なゴールが存在しない対話（非ゴール指向対話）では，対話の目的を適切に表現する報酬関数の設計は難しい．例えばカウンセリングを行う対話システムを構築する場合，明確なゴールが存在しないため，どのシステム行動やユーザ状態に高い報酬を設定すればよいかは自明ではない．そのため，このように直接報酬関数を設計する方法は，ゴール指向対話の対話制御に限られてきた．

こうした問題に対して，非ゴール指向対話での報酬関数を適切に設定するため，対話の目的に対する適切さを表す評価値を人手で対話コーパスに付与し，その値から報酬関数を求める方法が提案されている [Meguro 10, Williams 05]．この方法は対話目的を表現する報酬関数を比較的容易に設定できるものの，評価値と対話の目的に対する適切さの関係が評価者間で異なるという曖昧性が残る．すなわち，適切さに関する相対的な評価基準が評価者間で一致している場合であっても，辛口の評価者は 3 点，甘口の評価者は 6 点のように，異なる評価値が一つの対話に付与されてしまう．この曖昧性のため，評価者間で報酬関数が一致せず，適切なシステム行動がそれぞれ異なるという問題が発生する．

一方，こうした評価値によらず報酬関数を自動的に設定する手法として，近年逆強化学習 [Ng 00, Abbeel 04] が提案されており，対話制御への応用が進められている [Chandramohan 11, Boularias 10]．逆強化学習とは，行動列（ユーザ状態とシステム行動のペアの系列）を入力として，その行動列に含まれるペアの関係で，あるユーザ状態に対するシステム行動を生成する報酬関数を推定する手法である．すなわち，対話の目的に対して最適な行動列を訓練行動列とすれば，逆強化学習は最適なシステム行動を生成する報酬関数を推定できる．この報酬関数の推定に評価値を必要としない点が逆強化学習の特徴である．これは行動列のみから報酬関数を推定できるという利点である一方，評価値を扱うことができないという欠点でもある．すなわち逆強化学習では，行動列が全て等しく最適であると仮定している．行動列に付与された評価が成功・不成功のように明確に分かれている場合には，不成功と評価された行動列を取り除くことで報酬関数の最適性を維持できる．しかし評価が明確に分けられず連続的に表現される場合，どこで切り分けても報酬関数は最適ではなくなる．例えば，最適ではないが許容範囲という程度の適切さを持つ行動列が訓練行動列に含まれている場合，最適性を維持するためにはこれらを取り除く必要がある．その一方，これらの行動列を取り除くと，許容範囲行動列も不適切行動列も一様に訓練行動列に含まれなくなり，ある行動列が許容範囲か完全に不適切かを区別できない報酬関数となる．

本研究は，上記の 2 つの手法を融合し，各々の問題点を解決することを目的とする．これを実現するため，評価値が行動列ごとに付与された付与された対話コーパスから，評価値の順序関係に基づき報酬関数を推定する，順序学習に基づく逆強化学習（PIRL）を提案する．順序学習とは教師あり学習の一分野であり，与えられた訓練順序関係から順序モデルを学習する手法である．すなわち PIRL は，付与された評価値の順序関係と同様の順序関係で訓練行動列を評価できるような報酬関数を推定する．なお PIRL では，付与された評価値の順序関係のみを用い，値そのものは用いない．これは前述した評価値と適切さの相対関係が評価者間で異なる一方，順序関係はこの曖昧性に対して頑健であり，評価者間で一貫すると期待できるためで

連絡先: sugiyama.hiroaki@lab.ntt.co.jp

ある．また，本研究で提案する PIRL では，用いる評価値は行動ごとではなく一連の行動列単位で付与され，対話コーパスにあらかじめ付与された値のみを用いることとする．

類似の問題を扱った研究に，Silva らの研究と Cheng らの研究がある．Silva らは，PIRL と同様にペアごとの順序関係を利用する IRL を提案している [Freire da Silva 06]．しかし Silva らの手法では，学習のイテレーションごとにその時点での報酬関数を基に生成した行動列を，評価者が毎回評価する必要がある．Cheng らは順序学習に基づく強化学習を提案しており，報酬関数を経由せず直接システム行動を推定するアプローチを採っている [Cheng 11]．しかし Cheng らの手法は，行動列ごとではなくシステム行動のペアごとに順序評価を行う必要がある．すなわちどちらの手法も，評価者が評価を付与する回数が極めて膨大になり，シミュレーション以外では利用することが難しい．一方本研究で提案する PIRL では，各行動列に対し一度だけ評価値を付与すればよく，上述の手法に比べ実対話でも導入しやすい．

本研究では，対話コーパス中のユーザ発話に対するシステム行動をシミュレートし，それを評価するオフライン評価実験を通して提案手法の有効性を検証する．

2. 順序学習に基づく逆強化学習

本節では，はじめに強化学習を用いた対話制御について定義する．それを踏まえて，提案する順序学習に基づく逆強化学習を説明する．

2.1 強化学習を用いた対話制御

強化学習や逆強化学習は，一般にマルコフ決定過程（MDP）の枠組みで表現される．MDP は $\{\mathcal{S}, \mathcal{A}, \mathcal{T}, \gamma, R\}$ によって定義される．ここで $\mathcal{S}$ はユーザ状態，$\mathcal{A}$ はシステム行動，$\mathcal{T}: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{S}$ は遷移関数，$\gamma \in (0, 1]$ は将来得られる報酬の割引率，$R: \mathcal{S} \times \mathcal{A} \rightarrow \mathcal{R}$ は報酬関数を表す．加えて本研究では，行動列 ζ をユーザ状態とシステム行動のペアの系列 $\zeta_i = \{s_{i,t}, a_{i,t} : 0 \leq t < T\}$ として定義する．

MDP において強化学習は，将来に渡って獲得される報酬の累積値の期待値を最大化するようなポリシー $\pi: \mathcal{S} \rightarrow \mathcal{A}$ を決定する問題として定式化される．本研究では π はユーザ状態 s に対するシステム行動 a の適切さを表す行動価値関数 $Q(s, a)$ を用いて $\pi(s) = \arg\max_a Q(s, a)$ として定義する．この場合強化学習は $Q(s, a)$ の推定問題に帰着される．

2.2 順序学習に基づく逆強化学習

非ゴール指向対話コーパスには，各行動列に評価値が付与されているものが多い [Meguro 10, Williams 05]．従来の逆強化学習では，各行動列は等しく最適であると仮定していたため [Abbeel 04, Chandramohan 11, Boularias 10]，これらの評価値を利用することができなかった．本研究では訓練行動列とそれらに付与された評価値（訓練評価値）を利用して報酬関数を推定する手法として，順序学習に基づく逆強化学習（PIRL）を提案する．

PIRL の基本的なアイディアは，訓練行動列のペア $\{\zeta_i, \zeta_j\}$ ごとに，それらの訓練評価値 $\{e_i^*, e_j^*\}$ の順序関係 $o_{i,j}^* = \frac{e_i^* - e_j^*}{|e_i^* - e_j^*|} = \{-1, 0, 1\}$ を考え，それと同じ順序関係で推定評価値 $e_i^{\boldsymbol{\theta}}, e_j^{\boldsymbol{\theta}}$ を出力するような報酬関数 $\boldsymbol{\theta}$ を，順序学習を用いて推定するというものである．本研究ではこの順序学習を，$o_{i,j}^* \neq 0$ $(e_i^* \neq e_j^*)$ の関係を持つペアに対する二値分類問題と考え，

$$L(\boldsymbol{\theta}) = P(\boldsymbol{o}^*|\boldsymbol{\zeta}, \boldsymbol{\theta}) = \sum_{i,j:i<j, o_{i,j}^* \neq 0} \frac{(1 + o_{i,j}^{\boldsymbol{\theta}})^{\frac{1+o_{i,j}^*}{2}} \cdot (1 - o_{i,j}^{\boldsymbol{\theta}})^{\frac{1-o_{i,j}^*}{2}}}{2M} \quad (1)$$

を最大化する問題として定式化する．$L(\boldsymbol{\theta})$ は報酬関数 $\boldsymbol{\theta}$ における推定順序関係 $o_{i,j}^{\boldsymbol{\theta}} = \frac{e_i^{\boldsymbol{\theta}} - e_j^{\boldsymbol{\theta}}}{|e_i^{\boldsymbol{\theta}} - e_j^{\boldsymbol{\theta}}|}$ が $o_{i,j}^*$ に一致するペアの割合を表す．ただし，M は $o_{i,j}^* \neq 0$ を満たす訓練行動列ペア数とする．また $e_i^{\boldsymbol{\theta}}$ を，$\boldsymbol{\theta}$ における推定評価値として，後述する行動価値関数 $Q^{\boldsymbol{\theta}}(s_{i,t}, a_{i,t})$ に基づき

$$e_i^{\boldsymbol{\theta}} = \sum_{s_{i,t}, a_{i,t} \in \zeta_i} Q^{\boldsymbol{\theta}}(s_{i,t}, a_{i,t}), \quad (2)$$

と定義する．

本研究で提案する PIRL は，報酬関数を逐次的に更新し推定する．この逐次更新は，各イテレーションごとに報酬関数 $\boldsymbol{\theta}^n$ を用いて推定順序関係 $\boldsymbol{o}^{\boldsymbol{\theta}^n}$ を計算し，それらと訓練順序関係 $\boldsymbol{o}^*$ が異なるペア $\{i, j : o_{i,j}^* \neq o_{i,j}^{\boldsymbol{\theta}^n}\}$ ごとに計算された勾配 $\frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n}$ に従って行われる．詳細をアルゴリズム 1 に示す．

input ：訓練行動列 $\boldsymbol{\zeta}$ およびその順序関係 $\boldsymbol{o}^*$
output: 報酬関数 $\boldsymbol{\theta}$

0. 報酬関数 $\boldsymbol{\theta}^0$ を初期化.

for *n* **to** *N* **do**
- 1. 現在の報酬関数 $\boldsymbol{\theta}^n$ に従って行動価値関数 $Q^{\boldsymbol{\theta}^n}(s, a)$ を計算 (3).
- 2. $Q^{\boldsymbol{\theta}^n}(s, a)$ に従って $\boldsymbol{\zeta}$ を評価し (2)， 推定順序関係 $\boldsymbol{o}^{\boldsymbol{\theta}^n}$ を計算.
- **foreach** $\{i, j | i < j,\ e_i^* \neq e_j^*\}$ **do**
  - **if** $o_{i,j}^* \neq o_{i,j}^{\boldsymbol{\theta}^n}$ **then**
    - 3. $\frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n}$ を計算 (4).
  - **end**
- **end**
- 4. L の収束を評価.
- 5. L-BFGS アルゴリズムを用いて $\frac{\partial L}{\partial \boldsymbol{\theta}^n}$ に従って $\boldsymbol{\theta}^n$ を更新.

end

Algorithm 1: 順序学習に基づく逆強化学習

このアルゴリズムは，訓練行動列 $\boldsymbol{\zeta}$ と訓練順序関係 $\boldsymbol{o}^*$ を入力とする．ステップ 1 では，現在の報酬関数 $\boldsymbol{\theta}^n$ に従って行動価値関数 $Q^{\boldsymbol{\theta}^n}(s, a)$ を計算する．PIRL では $\boldsymbol{\theta}^n$ の更新に勾配法を用いるため，$Q^{\boldsymbol{\theta}^n}(s, a)$ の微分値を用いる．このため，$Q^{\boldsymbol{\theta}^n}(s, a)$ を微分可能なように近似的に

$$Q^{\boldsymbol{\theta}^n}(s, a) = \sum_{s'} \{\theta^n(s, s') P_T(s'|s, a) + \gamma \max_{a'} \sum_{s''} P_T(s'|s, a) \theta^n(s', s'') P_T(s''|s', a')\}, \quad (3)$$

として定義する．これは，ユーザ状態 s における将来獲得する報酬の期待値を，2 時刻先までの期待値で近似したものである．ただし $\theta(s, s')$ はユーザ状態が s から s' に遷移した場合の報酬値を表し，$P_T(s'|s, a)$ はシステム行動 a によってユーザ状態が s から s' へ遷移する確率を表す．次に，(2) を用いて推定評価値 $\boldsymbol{e}^{\boldsymbol{\theta}^n}$ を計算し，推定順序関係 $o_{i,j}^{\boldsymbol{\theta}^n}$ を定める．ステップ 3 では，訓練順序関係と推定順序関係が異なるペア $\{\zeta_i, \zeta_j : o_{i,j}^* \neq o_{i,j}^{\boldsymbol{\theta}^n}\}$ ごとに，報酬関数の勾配を

$$\frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n} \propto o_{i,j}^* \{\sum_{s_{i,t}, a_{i,t} \in \zeta_i} \frac{\partial Q^{\boldsymbol{\theta}^n}(s_{i,t}, a_{i,t})}{\partial \boldsymbol{\theta}^n} - \sum_{s_{j,t}, a_{j,t} \in \zeta_j} \frac{\partial Q^{\boldsymbol{\theta}^n}(s_{j,t}, a_{j,t})}{\partial \boldsymbol{\theta}^n}\}. \quad (4)$$

として計算する．(3) より，$\frac{\partial Q(s,a)}{\partial \boldsymbol{\theta}}$ の $\theta(s_1, s_2)$ における値は，

$$\frac{\partial Q(s, a)}{\partial \theta(s_1, s_2)} = \delta_{s, s_1} P_T(s_2|s_1, a) + P_T(s_1|s, a) P_T(s_2|s_1, a'_{s_1}),$$

となる．ただし，$a'_s = \arg\max_a \sum_s \theta(s,s')P_T(s'|s,a)$，$\delta_{s,s_1}$ は $s = s_1$ のとき 1，それ以外の時に 0 を取るクロネッカーのデルタとする．

こうして得られた各ペアごとの勾配 (4) を足しあわせた $\frac{\partial L}{\partial \boldsymbol{\theta}^n} = \sum_{i,j:o^*_{i,j} \neq o^{\boldsymbol{\theta}^n}_{i,j}} \frac{\partial L_{i,j}}{\partial \boldsymbol{\theta}^n}$ を報酬関数の勾配として，L-BFGS アルゴリズム [Liu 89] を用いて逐次的に報酬関数を更新する．

3. 実験

本節では，Maximum Entropy IRL，Profit-sharing に基づく強化学習，提案手法の 3 アルゴリズムの比較を通して提案手法の有効性を検証する．Maximum Entropy IRL は最高水準の IRL アルゴリズムの一つである [Ziebart 08]．ただし評価値を扱うことはできないため，本研究では高い評価値を付与された行動列のみを訓練行動列とする．このアルゴリズムとの比較を通して，低評価データを取り除くことによるデータスパースネスの問題の影響を検証する．

Profit-sharing に基づく強化学習は，評価値が付与された行動列から行動価値関数を学習する一般的な方法である [Grefenstette 88]．このアルゴリズムとの比較を通して，評価値をそのまま報酬関数として用いる場合との差異を検証する．

3.1 対話データ

本研究では，我々の以前の研究で収集した非ゴール指向対話データを用いる [Meguro 10]．これは，対話相手の話を積極的に聞くことを目的とする聞き役対話システムの実現を目指した研究である．この中で我々は，実験参加者を 10 人の聞き役（男性 5，女性 5）と話し役（男性 18，女性 19）に分け，聞き役と話し役が 1 人ずつペアで参加する聞き役対話を 1260 対話収録した．収録された各発話文に 32 個の対話行為タグ（挨拶，質問など）のいずれかを付与し，さらに対話収録に参加していない 2 人の評価者が対話ごとに評価を付与した．このとき評価者は，もし自分が話し役だったとしたら話を聞いてもらったと感じたか，という基準で，7 段階のリッカート尺度で評価を付与した．

本研究では，ユーザ状態空間 $\mathcal{S}$ として話し役の対話行為タグを，システム行動空間 $\mathcal{A}$ として聞き役の対話行為タグを利用する．本対話データでは，一発話に複数文が含まれる場合があった．そのため，各アルゴリズムは複数ユーザ状態に対する行動価値関数を $Q'(\boldsymbol{s},a) = \frac{1}{|\boldsymbol{s}|}\sum_{s' \in \boldsymbol{s}} Q(s',a)$ として表現する．一方，システム行動は各発話に対して 1 つのみ生成できると仮定する．

3.2 評価指標

各アルゴリズムを用いたシステムが適切な行動列を生成できるかを調べる率直な方法として，システムが生成した行動列と訓練行動列に含まれない実際の行動列（テスト行動列）の一致率を測る方法が考えられる．しかし本研究で用いる対話コーパスに含まれる行動列では，人の実際の行動を記録しているため，あるユーザ状態に対して複数のシステム行動が出現する．このため，それらを正しく生成することは非常に困難である．そのうえ，あるユーザ状態に対する適切な行動は一意に定まらないため，異なるシステム行動であっても適切に近いシステム行動とはなりうる．これらより，単純な行動列の一致率はアルゴリズムの評価には不十分であると考えられる．

そのため本研究ではアルゴリズムを比較するために，順序関係の一致率，順序関係の相関係数，期待評価値という 3 つの評価指標を定義する．順序関係の一致率と順序関係の相関係数は PIRL が最大化している指標であり，評価値が高い/低い行動列を適切に分類できる性能を表す．すなわちこれらの値が高いシステムは，適切な行動列を生成できると期待できる．また，システムの出力した順序評価と評価者の順序評価の一致率が評価者間の値に近ければ，そのシステムを評価シミュレータとして利用できる可能性がある．特に順序関係を用いることで，評価値付与に伴う曖昧性に対して頑健となることが期待できる．これに加えて，直感的に理解しやすい指標として期待評価値を定義する．これは，実際の対話でアルゴリズムが得ると期待される評価値である．こうした値を得るためには，アルゴリズムが出力する行動列を評価者が直接見て評価する方法が考えられる．しかし，行動列は発話行為タグで表現されており言語情報を含まないため，そのままでは評価者であっても評価が難しい．そのため本研究では，アルゴリズムごとに最高・最低の推定評価値が付与された n 個の行動列に対して評価者が付与した評価値の平均値を，高・低期待評価値として計算する．最高の推定評価値が付与されている行動列は，テスト行動列の中で最もアルゴリズムが生成しやすい行動列であるため，実際の対話においても同様の行動列が得られると考えられる．そのためこうした行動列から計算される高期待評価値は，実際の対話においてアルゴリズムが得る評価値に近くなると期待できる．一方，最低推定評価値が付与された行動列から計算される低期待評価値も，アルゴリズムが不適切なシステム行動を抑制できるかを評価する上で重要な指標である．

なお，評価者が付与した評価値の差が小さいペアは，評価値付与に伴う揺れによって順序関係が逆転する可能性がある．そのため本研究では評価者が付与した評価値が 2 以上離れているペアを対象に順序関係を推定，評価することとする．

3.3 実験設定

本対話データは人同士の対話を収録したものであるため，対話が破綻するような発話はほとんど含まれない．すなわち対話間の適切さの違いが小さく，評価者間の評価基準の僅かな差異が大きく評価値に反映されてしまいやすいため，正反対の評価値（2 と 6 など）が付与された対話が多く含まれていた．

こうした揺れを含めて，システムの出力する順序評価の評価者との一致率が評価者間の値に近ければ，評価シミュレータとして利用可能である．その一方，揺れを除去したデータに対し，どの程度適切に順序評価を推定できるかを調べることも，アルゴリズムの性能を知る上で重要である．そのため本実験では，使用するデータが異なる，全データ設定と選別データ設定の 2 通りの実験設定を考える．全データ設定では上述の対話データを全て用い，選別データ設定では二人の評価者間で評価値の差が 1 以下の行動列のみを用いる．

これら 2 つの実験設定の下で，収集した対話データ（行動列とその評価値）を訓練，調整，テストセットに分割する．全データ設定は，片方の評価者のデータから訓練，調整セットを作成し，もう一方の評価者のデータからテストセットを作る．ここで，データに含まれる行動列は評価者間で一致している（評価値のみが異なる）ため，訓練，調整セットとテストセットに重複がないように分割する．一方選別データ設定では，二人の評価者間で評価値の差が 1 以下の行動列のみを用い，各セットに分割する．なお，前述の通り収集した対話データは人同士が行なった対話であることから，対話が破綻するような行動列はほとんど含まれない．行動列にバリエーションを持たせるため，ランダムに生成した行動列に最低の評価値を付与したデータを訓練セットに加える．データの詳細を表 1 に示す．

3.4 結果

図 1 に各アルゴリズムと評価者との順序関係の一致率を示す．図 1 から，提案手法の PIRL がどちらの実験設定において

	全データ	選別データ
訓練行動列数	700 (ME:113)	500 (ME:88)
訓練ペア数	149515 (ME:2166)	75414 (ME:1043)
調整行動列数	300 (ME:300)	300 (ME:150)
調整ペア数	18177 (ME:18177)	16768 (ME:8416)
テスト行動列数	459	111
テストペア数	55310	2236

表 1: データセットの詳細．“ME:*” は Maximum Entropy IRL（評価値 4 以上）における数を表す．

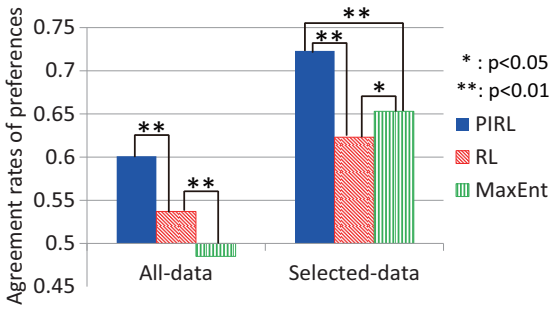

図 1: 各アルゴリズムと評価者との順序関係の一致率．評価者間は 0.632，ランダムに評価値を生成した場合のベースラインは 0.5 である．

も他手法を有意に上回っていることがわかる．選別データ設定の値が全データ設定に比べて高い理由は，評価者間で評価値が近くなるよう選別したことにより，評価値を付与する以前の適切さの評価がそもそも異なっているようなデータを取り除かれ，推定が容易になったためと考えられる．

	全データ			選別データ		
	Corr.	High	Low	Corr.	High	Low
PIRL	**0.200**	**4.66**	**3.46**	**0.363**	**5.06**	3.86
RL	0.069	4.46	4.13	0.218	4.46	**3.40**
MaxEnt	-0.001	3.80	3.60	0.285	4.66	3.93
評価者	0.211	4.90	2.86	0.817	6.80	1.33

表 2: 順位相関係数（Corr.）および高/低期待評価値（High/Low）．評価者が付与した評価値の平均値は，全データ設定で 0.402，選別データ設定で 0.458．相関係数および高期待評価値は高い値が，低期待評価値は低いほうが優れている．なお，選別データ設定は評価者間で評価値が一致したもののみを集めたデータであるため，評価者とシステムとの違いが大きくなっている．

表 2 に順序関係の相関係数と，最高・最低推定評価値を付与された 15 行動列から計算した高・低期待評価値を示す．提案手法の PIRL は，他手法より高い相関係数と高い高期待評価値を示した．この傾向は，期待評価値の計算に用いる行動列数を 5〜30 の範囲で変更しても同様に見られた．このことから，PIRL が報酬関数の推定に有効であるといえる．また，順序評価の一致率と相関係数が全データ設定における評価者間に近い値であった．すなわち，PIRL と評価者との違いは評価者間の違いと同程度であるため，PIRL は評価シミュレータとしても応用可能である．

4. 結論

本研究では，評価値が付与された対話行動列から対話制御に用いる報酬関数を推定する，順序学習に基づく逆強化学習（PIRL）を提案し，オフライン評価実験を通して有効性を示した．従来の IRL では，行動列が単一の適切さを持つと仮定しており，多様な適切さを持つ行動列を扱えなかった．本研究で提案した PIRL は，IRL の適用範囲を広げる重要な拡張である．PIRL では，評価値から計算されるペアごとの順序情報に基づき，従来の IRL では取り除かれていた非最適なデータも用いて報酬関数を推定することができる．さらに実験を通して，PIRL が対話制御の評価シミュレータとして利用できる可能性を示した．

本研究ではオフライン評価実験のみによって提案手法の有効性を確認したため，今後実際のユーザと WOZ との対話を通したオンライン評価実験を行う予定である．また，本研究では対話行為タグでのみ対話を扱っていたため，発話の内容は扱うことができていなかった．この点を改善するため，IRL を用いた対話制御に使われている Hidden Topic Markov Model のようなトピックモデル [Boularias 10] を適用していくことを考えている．

参考文献

[Abbeel 04] Abbeel, P. and Ng, A. Y.: Apprenticeship learning via inverse reinforcement learning, in *ICML* (2004)

[Boularias 10] Boularias, A., Chinaei, H., and Chaib-draa, B.: Learning the Reward Model of Dialogue POMDPs from Data, in *NIPS 2010 Workshop on Machine Learning for Assistive Technologies (MLAT-2010)*, pp. 1–9 (2010)

[Chandramohan 11] Chandramohan, S., Geist, M., Lefevre, F., and Pietquin, O.: User Simulation in Dialogue Systems using Inverse Reinforcement Learning, in *Interspeech* (2011)

[Cheng 11] Cheng, W., Fürnkranz, J., Hüllermeier, E., and Park, S.: Preference-based policy iteration: leveraging preference learning for reinforcement learning, in *ECML-PKDD*, pp. 312–327 (2011)

[Freire da Silva 06] Freire da Silva, V., Reali Costa, A., and Lima, P.: Inverse reinforcement learning with evaluation, in *ICRA* (2006)

[Grefenstette 88] Grefenstette, J.: Credit assignment in rule discovery systems based on genetic algorithms, *Machine Learning*, Vol. 3, No. 2, pp. 225–245 (1988)

[Liu 89] Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Mathematical programming*, Vol. 45, pp. 503–528 (1989)

[Meguro 10] Meguro, T., Higashinaka, R., Minami, Y., and Dohsaka, K.: Controlling Listening-oriented Dialogue using Partially Observable Markov Decision Processes, in *Coling* (2010)

[Ng 00] Ng, A. and Russell, S.: Algorithms for inverse reinforcement learning, in *ICML*, pp. 663–670 (2000)

[Williams 05] Williams, J. D. and Young, S.: The SACTI-1 Corpus: Guide for Research Users, *Technical report, University Of Cambridge* (2005)

[Williams 07] Williams, J. D.: Applying POMDPs to dialog systems in the troubleshooting domain, in *Workshop on Bridging the Gap*, pp. 1–8, (2007)

[Ziebart 08] Ziebart, B., Maas, A., Bagnell, J., and Dey, A.: Maximum entropy inverse reinforcement learning, in *AAAI*, pp. 1433–1438 (2008)