逆強化学習による文章における人間らしさの推定

岸川大航　大関洋平
東京大学
{d-kishikawa,oseki}@g.ecc.u-tokyo.ac.jp

概要

文章からの作者の個性の推定は、言語モデルにおける人間らしい文章生成、あるいは文章の定量的分析などに応用可能であるため、重要な研究課題であると考えられる。本研究では、言語生成過程をマルコフ決定過程によってモデル化し、報酬を推定する逆強化学習を用いることによって、報酬の形で文章の個性を推定することを提案する。計算機実験として、夏目漱石の文章とそれ以外の作家の文章を提案手法によって学習させ、文章が識別可能であること、推定報酬を用いた文章表現に対する定量的な評価が可能になることを示す。

1　はじめに

ある文章において、作者らしさ（作者の個性）をどのように定量的に評価するかは、自然言語処理において重要な研究課題である。文章における作者の個性を定量化することができれば、たとえばChatGPT に代表される言語モデルに対し、非人間（単純なマルコフ連鎖、あるいは言語モデルなど）の文章に対する人間の文章の「人間らしさ」を定量化することで、より人間らしい文章を生成することが可能となり、「日本語文章における日本人らしさ」を定量化することによる日本語の非母語話者に対する語学教育などにも応用可能である。したがって、文章の個性を推定する手法の実用化が求められているといえる。

本研究では、文章生成が人間の意思決定の連続に基づくものであるとの考えから、文章生成過程をマルコフ決定過程によって定式化し、与えられた手本となる行動例（デモンストレーション）から報酬を推定する、逆強化学習を文章に対して適用し、報酬の形で文章の個性を推定することを提案する。推定された報酬を用いて、文単位で作者らしさを定量的に評価できるほか、各文において、具体的にどの表現が高い、あるいは低い報酬値となるかを可視化できるため、作者らしい文章表現を特定することも可能である。

2　準備

2.1　マルコフ決定過程 (MDP)

マルコフ決定過程 (Markov Decision Process; MDP)[1] は、強化学習（Reinforcement Learning; RL）の基礎となる確率過程である。MDP は、状態 $s \in S$、行動 $a \in A$、状態 s で行動 a をとった際に次状態 s' に遷移する状態遷移確率 $p(s'|s,a)$、状態 s で行動 a をとった際の報酬 $r(s,a)$、割引率 $\gamma(0 \leq \gamma \leq 1)$、そして初期状態分布 $p_0(s)$ から構成される。MDP における最適方策 π^* は、累積割引報酬を最大化する方策であり、式（1）によって表される。

$$\pi^*(a_t|s_t) = \mathrm{argmax}_\pi \mathbb{E}_{s_0 \sim p_0}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t,a_t)|s_t, a_t \sim \pi(a_t|s_t)\right] \quad (1)$$

MDP 上において最適方策を学習する手法が強化学習（Reinforcement Learning; RL）であり、意思決定主体のことをエージェントと呼ぶ。エージェントは環境内に存在し、MDP に従いながら、試行錯誤によって最適方策を獲得する。

2.2　逆強化学習 (IRL)

RL が、与えられた報酬関数に対して最適方策を獲得する手法であるのに対し、逆強化学習（Inverse Reinforcement Learning; IRL）[2] はその逆であり、最適方策をとるエージェント、エキスパートが存在し、エキスパートが累積割引報酬を最大化するという仮定のもとで、最適方策、または最適方策に従って、環境内で記録された状態から構成される時系列（軌跡（trajectory）と呼ぶ）のデータセット、デモンストレーションを所与として、その背後に存在する報酬関数を推定する手法である。

具体的には、IRL は以下の過程に従って報酬を推

定する。

1. エキスパートデモンストレーションと、現在のエージェントの軌跡の乖離度を、何らかのダイバージェンスによって比較し、報酬関数を推定する。
2. 報酬関数に従って、エージェントに最適方策を獲得させる。
3. エージェントを環境内で行動させ、軌跡を記録する。
4. 1. に戻る。

この過程を繰り返し、最終的にダイバージェンスが 0 となった場合、エージェントは最適方策を獲得し、その時推定中の報酬関数がエキスパートが持つ報酬関数となる。IRL は、報酬関数を設計することが難しいロボットの複雑な行動の学習や、本来の報酬関数が未知である人間や生物の意図分析 [3, 4] に用いられている。中でも、オフライン IRL 手法は 2. が不要であるため、高速に学習可能である。

3 文章の個性推定のための IRL

文章を作成するエージェントが、形態素 w を一つずつ出力しながら文章を生成する過程のことを文章生成過程と呼ぶことにする。文章生成過程は以下の過程からなる。

1. 時刻 t のエージェントは、時刻 t までの形態素からなる文章断片 $s_t = \{w_0, w_1, w_2, \cdots, w_t\}$ を現状態として観測し、自身の持つ方策（執筆方針にあたる）$\pi(a|s)$ に基づき、行動 a_t として次の形態素 w_{t+1} を出力する。
2. 次に環境は、次状態として $s_{t+1} = \{w_0, w_1, w_2, \cdots, w_t, w_{t+1}\}$ を、何らかの文章断片に対する評価関数 $C(x)$（文章断片 x が好ましいほど $C(x)$ は高い値を返すものとする）を報酬値 $r_t = C(s_t)$ として返す。
3. 最後に、報酬 r_t に基づいてエージェントは自身の持つ π を学習し、観測した次状態 s_{t+1} に基づき行動 a_{t+1} を出力する。
4. 1. に戻る。

以上の定式化によって文章生成過程をモデル化するとき、エキスパートデモンストレーションにあたるのは特定の作者の文章である。文章を形態素解析によって形態素単位に分割し、状態にあたる形態素列 s_t を、何らかの文章ベクトル化手法 $f(x)$ による変換関数 $v_t = f(s_t)$ によってベクトル化した系列を軌跡とする。すなわち、ある文章 $\sigma = \{w_0, w_1, \cdots w_T\}$ が $s_0 = \{w_0\}, s_1 = \{w_0, w_1\}, \cdots, s_T = \{w_0, w_1, \cdots w_T\}$ という文章断片に分解されるとき、軌跡 $\tau = \{v_0, v_1, \cdots, v_T\}$ となる（T は文章の終端時刻を示す）。

4 計算機実験

計算機実験として、夏目漱石の作品とそれ以外の作家の作品の文章データから、夏目漱石らしさが推定できるかどうかを検証する。今回の実験において用いるのは、エキスパートとして夏目漱石の 12 作品、ベースラインとして森鴎外・泉鏡花・太宰治・島崎藤村の各 3 作品、計 12 作品を、それぞれ用いる。データは青空文庫 [5] のものを使用した。

- エキスパート：夏目漱石 … 『こころ』『それから』『虞美人草』『吾輩は猫である』『坑夫』『行人』『三四郎』『草枕』『道草』『彼岸過迄』『坊っちゃん』『明暗』
- ベースライン：その他の作家 4 人
  - 森鴎外 … 『雁』『渋江抽斎』『青年』
  - 泉鏡花 … 『黒百合』『日本橋』『婦系図』
  - 太宰治 … 『斜陽』『人間失格』『惜別』
  - 島崎藤村 … 『家』『新生』『夜明け前』

これらのデータを結合・全文シャッフル・重複を除去した上で、train・dev・test データとしておおよそ 8:1:1 になるようにデータを分割し、train と dev を報酬の学習に使用した。また、報酬を出力するニューラルネットワーク（報酬モデル）を学習するための IRL アルゴリズムには LogReg-IRL[6] を用いた。

エキスパートおよびベースラインの test データにおけるスコアの分布をヴァイオリンプロットとして図 1 に示し、図 1 の X 軸を範囲 $[-25, 25]$ の範囲で拡大したプロットを図 2 に示す。また、各スコアの平均を表 1 に示す。ここでスコアとは、ある文章 σ について、各文章断片 w ごとの報酬値を足し合わせたもの、すなわち

$$\mathrm{score}(\sigma) = \sum_{t=0}^{T} r_\theta(s_t)$$

である。図 1、図 2 の縦軸がスコアを示す。

図 1、図 2、表 1 より、分布は非常に近いものの、総じてエキスパートのスコア平均はベースラインのスコア平均を上回っており、分離できていることがわかる。

エキスパートおよびベースラインにおける実際の

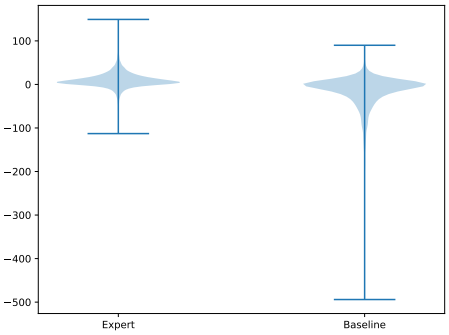

図 1　test データにおけるスコアのヴァイオリンプロット

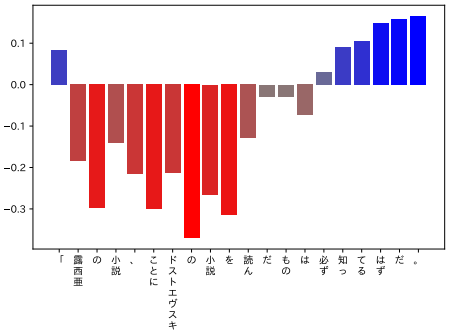

図 3　夏目漱石作品（test データ）における報酬値

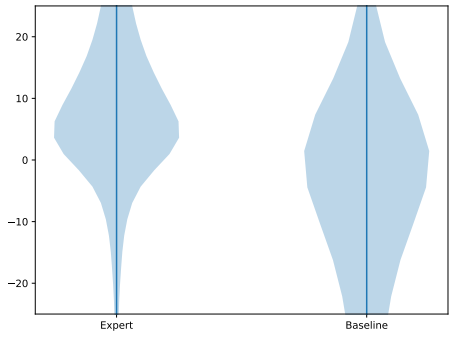

図 2　図 1 の X 軸を範囲 [−25, 25] において拡大した
ヴァイオリンプロット

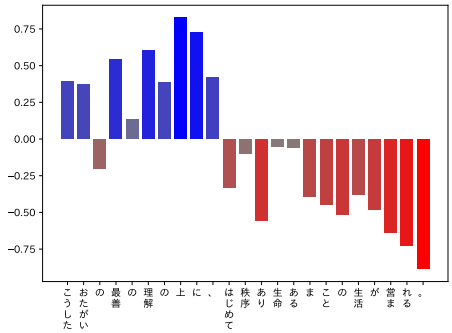

図 4　ベースライン作品（test データ）における報酬値

報酬値を可視化し、図 3、図 4 にそれぞれ示す。図の横軸は、最も左の形態素が文の最初の形態素であり、各形態素は文の冒頭からその形態素までの文章断片を示し、縦軸はその文章断片に対する報酬モデルの出力する報酬値である。まず図 3 であるが、これは『明暗』の一節である。文章の前半は負の報酬であるため、夏目漱石（すなわちエキスパート）らしくないと判定されているが、文末表現まで含めた場合は、報酬値が正であるため、全体として夏目漱石らしいと判定されている。一方、図 4 は島崎藤村の『新生』の一節である。こちらは対照的に、文の前半は夏目漱石らしいものの、文の後半が夏目漱石らしくないと判定され、全体として夏目漱石ではないと判定されている。これらの例では、文末表現が夏目漱石らしさを示す根拠として用いられているといえる。

4.1　未知のデータにおける検証

前節の test データは、train・dev・test データとしておおよそ 8:1:1 になるように分割されたデータであるため、基本的には学習に利用した学習に利用した文章と同一作品内での検証となり、スコアが特定の固有名詞（たとえば作品中の登場人物名）などの影響を大きく受ける可能性が考えられる。このような報酬モデルは、文章における個性の推定という点では適切ではなく、未知のデータにおいても個性を評価できるモデルの方が望ましいと考えられる。そこで、エキスパート・ベースラインともに、学習には利用していない、別の 12 作品のデータを利用し、前節で学習した報酬モデルの評価が一貫するかを検証する。利用した作品は次のとおりである。

- エキスパート：夏目漱石 ··· 『一夜』『永日小品』『幻影の盾』『手紙』『硝子戸の中』『創作家の態度』『文鳥』『変な音』『夢十夜』『門』『野分』『倫敦塔』
- ベースライン：その他の作家 4 人
  - 森鴎外 ··· 『かのように』『阿部一族』『山椒大夫』
  - 泉鏡花 ··· 『三枚続』『式部小路』『草迷宮』
  - 太宰治 ··· 『新ハムレット』『新釈諸国噺』『正義と微笑』
  - 島崎藤村 ··· 『桜の実の熟する時』『千曲川のスケッチ』『力餅』

エキスパートおよびベースラインの test データに

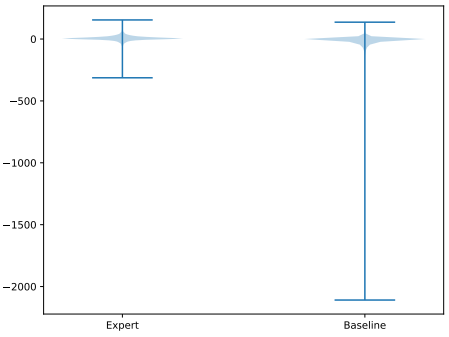

図 5　未知データにおけるスコアのヴァイオリン
プロット

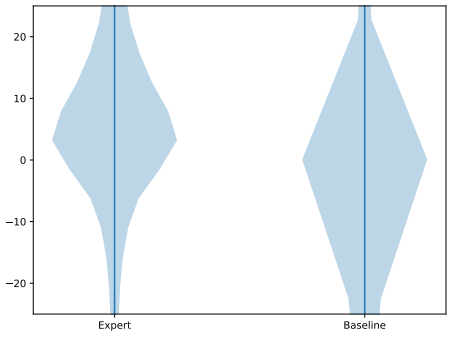

図 6　図 5 の X 軸を範囲 [−25, 25] において拡大した
ヴァイオリンプロット

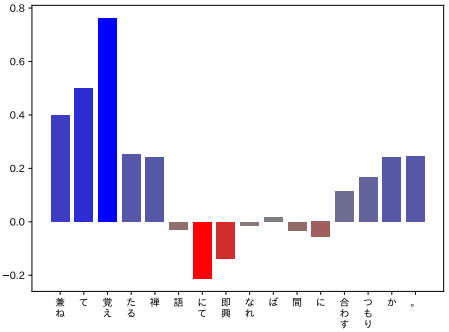

図 7　夏目漱石作品（未知データ）における報
酬値

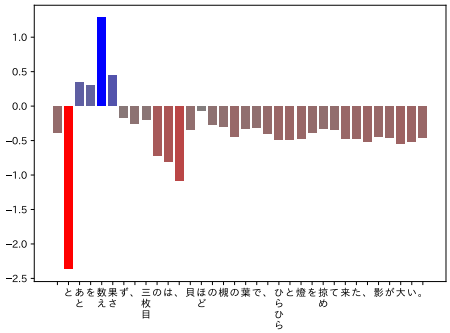

図 8　ベースライン作品（未知データ）における報酬値

おけるスコアの分布をヴァイオリンプロットとして図 1 に示し、図 5 の X 軸を範囲 [−25, 25] の範囲で拡大したプロットを図 6 に示す。また、各スコアの平均を、test データと同様に表 1 に示す。ヴァイオリンプロットは前節と同様である。

図 5、図 6、表 1 より、test データにおける結果よりは差が小さくなっているものの、総じてエキスパートのスコア平均がベースラインのスコア平均を上回っており、分離できていることがわかる。したがって、提案手法は未知のデータにおいても有効であるといえる。

報酬について、エキスパート、ベースラインをそれぞれ図 7、図 8 に示す。図 7（夏目漱石『一夜』の一節）では、文の冒頭部分、および文末表現まで入れた場合に正の報酬値が出力されており、夏目漱石らしいと判定されている。一方、ベースライン（泉鏡花『草迷宮』の一節）では、文の冒頭は正の報酬値が出ており、夏目漱石らしいと判定されているものの、文末までを含めると報酬値が負となり、夏目漱石ではないと判定されている。以上に述べた報酬の可視化による分析では、文末表現が報酬モデルの判断根拠であるという結論にいたったが、夏目漱石作品の計量分析において文末表現に着目した研究 [7] が存在することからも、この結論は妥当であるといえる。

5　まとめ

本論文では、文章の個性を推定する手法として、IRL を用いる手法を提案した。計算機実験として、夏目漱石とそれ以外の作家の文章から「夏目漱石らしさ」を推定することが可能かどうかを検証し、実際に両者が分離できること、報酬の可視化による判断根拠の説明が可能であることを確認した。今後は、人間および非人間の文章から、文章における人間らしさが推定できるかの検証を予定している。

表 1　夏目漱石作品および他作品におけるスコアの平均

	夏目漱石作品	他作品
test データ	9.39	-17.56
未知データ	5.93	-10.89

謝辞

本研究は、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Richard S Sutton and Andrew G Barto. **Reinforcement learning: An introduction**. MIT press, 2018.

[2] Andrew Y Ng, Stuart J Russell, et al. Algorithms for inverse reinforcement learning. In **Icml**, Vol. 1, p. 2, 2000.

[3] Shoichiro Yamaguchi, Honda Naoki, Muneki Ikeda, Yuki Tsukada, Shunji Nakano, Ikue Mori, and Shin Ishii. Identification of animal behavioral strategies by inverse reinforcement learning. **PLoS computational biology**, Vol. 14, No. 5, p. e1006122, 2018.

[4] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In **European conference on computer vision**, pp. 201–214. Springer, 2012.

[5] 青空文庫 Aozora Bunko. https://www.aozora.gr.jp/, 2024.

[6] Eiji Uchibe. Model-free deep inverse reinforcement learning by logistic regression. **Neural Processing Letters**, Vol. 47, No. 3, pp. 891–905, 2018.

[7] 土山玄. 文末表現の計量分析に基づく夏目漱石の小説の分類. 研究報告人文科学とコンピュータ（CH）, Vol. 2019-CH-120, No. 6, pp. 1–4, 2019.