自然言語処理における逆強化学習・模倣学習の適用

坪井祐太*　牧野貴樹**†

＊日本アイ・ビー・エム株式会社 東京基礎研究所
＊＊東京大学 生産技術研究所
＊IBM Research - Tokyo
＊＊IIS, the University of Tokyo
＊E-mail: yutat@jp.ibm.com
＊＊E-mail: mak@sat.t.u-tokyo.ac.jp

キーワード：自然言語処理 (natural language processing), 逆強化学習 (inverse reinforcement learning), 模倣学習 (imitation learning), 構造化予測 (structured prediction)
JL 002/02/4202–0086 ©2002 SICE

1. まえがき

自然言語処理（Natural Language Processing; NLP）は，人間が日常的に使っている自然言語をコンピュータで処理する技術の総称である．特に，インターネットの発展により膨大なテキストデータが利用可能になったことで産業界からも注目を集めており，企業の研究活動も活発である．また，2011 年にクイズ番組でコンピュータが人間のチャンピオン 2 人を負かすこととなったのは，NLP の一つの応用である質問応答技術の発展によるものである [1]．

複雑な自然言語を扱うために，NLP はいくつもの処理をつなげたパイプライン処理として構成されることが一般的である．多くの NLP 応用で共通して利用される基盤処理として単語分割・品詞タグ付け・固有表現抽出・構文解析などがあり，機械翻訳・自動要約・質問応答などの応用ではこれらの解析結果を入力として用いてより高度な処理を行う．本稿では，NLP の基盤処理の実現においてどのように強化学習，特に逆強化学習（Inverse Reinforcement Learning）・模倣学習 (Imitation Learning) が活用されているかについて紹介する．これらは，いわば人間が読む作業を模倣する決定過程を設計・学習する手法である．

タスクを理解するための例として，英文 “Fruit flies like a banana.”（訳：果物バエ–ショウジョウバエ–はバナナを好む）に対する品詞タグ付けと構文解析（係り受け構造解析および句構造解析）の出力を図 1 に示す．品詞タグ付けは各単語の品詞を予測するタスク，係り受け構造解析は単語間の係り受け関係を示す木構造を予測するタスク，句構造解析は名詞句（NP）や動詞句（VP）などの再帰構造を予測するタスクである．品詞によって文の内容を示す単語（名詞や動詞など）と機能語（冠詞など）を区別することができ，また構文解析によって長文でも述語項の関係（AがBをCする）などを抽出することが容易になる．

これらのタスクでは，コーパスと呼ばれる正解付きのテキストデータを用いて予測モデルを学習する統計的なアプローチが主流である．初期には教師付き学習が適用されてきた [2]．教師付き学習では説明変数の値（周囲の単語など）から目的変数の値（品詞など）を予測する分類器を構成するために，正解となる目的変数の値が付与されたデータ（正解データ）を使って分類器のパラメータを推定する．ただし，教師付き学習では問題設定としては事例間は独立で分布は一定であることを仮定していた（independently identically distributed; i.i.d.）．自然言語を解析する際には文や文書全体の整合性が重要になることが多いため，目的変数間の依存関係を考慮した**構造化予測**（Structured Prediction）と呼ばれる問題として定式化されることが多い．図 1 の例文では Fruit が形容詞であることと flies が名詞であること，flies が名詞であることと like が動詞であることは密接に関係しており，一方で文脈が変われば flies は動詞である可能性もある．図 1 と同様に単語列 flies like が出現する文 “Time flies like an arrow.”（訳：光陰矢のごとし）を例にとると，Time が名詞，flies が動詞，like は前置詞である．このように，例えば品詞タグ付けでは各単語の品詞間に依存関係がある．また，構文解析も局所的な判断だけではなく全体的な最適化が必要なタスクである．なお，依存関係を表現するために単純に目的変数も特徴量として既存の教師付き学習を適用してしまうと，ある点（事例）での目的変数の予測が他の点（事例）での特徴量に影響を及ぼすため i.i.d. の仮定が崩れてしまうことに注意が必要である．

変数の依存関係がある構造化予測問題では特に，与えら

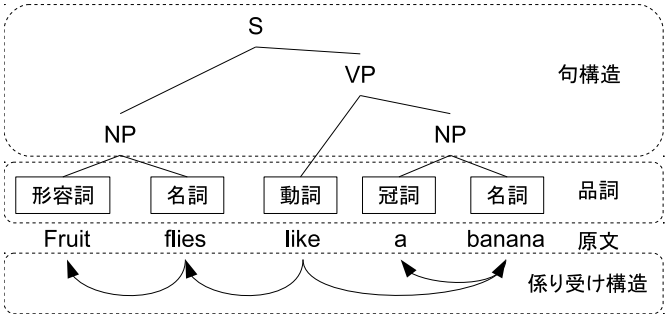

図 1　自然言語処理における構造化予測の例
“Fruit flies like a banana.” を入力とした場合の，品詞タグ付けと構文解析（係り受け構造解析および句構造解析）の出力．品詞タグ付けは品詞（配列構造），係り受け構造解析は係り受け関係 (各単語がノードとなる木構造)，句構造解析では句構造 (句の文法的役割をノードとする木構造) を出力とする．

† 総合科学技術会議により制度設計された最先端研究開発支援プログラム (FIRST 合原最先端数理モデルプロジェクト) により，日本学術振興会を通して助成を受けた．

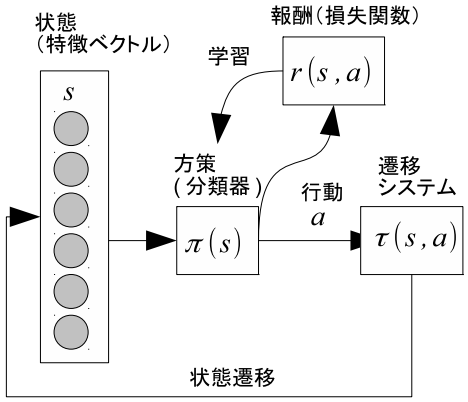

図 2　構造化予測に用いられる再帰的分類器
分類器によって状態遷移を繰り返すことで構造を構築する．各状態を特徴ベクトルとして表し，構造を構築するための次の行動を選択する．訓練時には，分類器は環境が与える報酬に基づいて更新する．

れたスコア関数の下で，入力に対して大域的に最適となる構造を決定することの難しさが問題となる．この困難を解決するアプローチとして，構造を逐次的に構築する遷移システムを設計し，構造操作を分類器で予測する手法がある（図 2）．こうすることで，個々の遷移の決定時に他の変数の値を参照することが自然に実現できる．さらに，構造操作を行動と考え，スコアを報酬とみなして強化学習を適用して行動選択させることで，逐次的な構築であるにもかかわらず，大域的最適となるような（強化学習で言えば，報酬和の期待値が最大化される）決定が実現できる．

NLP の代表的な構造化予測問題には，以下のような強化学習としての特徴がある．

1. コーパス上で正解構造は観測できるが，それを出力する行動列は観測できない
2. 人工的な遷移システムであるため，環境は既知で状態遷移は決定的だが，報酬（スコア関数）は設計が必要
3. 文や文書の単位で決定過程が分割されたエピソード的な決定過程
4. 単語やその連接を特徴とした広大な状態空間（数十万から数億の特徴次元）
5. 状態ごとに許容行動集合が異なる（図 3 に例を掲載）

これらの特徴のなかでも特徴 1 が既存の正解データを活用するためには重要であり，また正解構造から方策を学習するには適切な遷移システムを設計する必要がある．なお遷移システムによって仮定できる正解情報が異なる．文献 3) では仮定する正解情報を**最適訓練行動列**（Optimal Learning Trajectory）と**最適訓練方策**（Optimal Learning Policy）の 2 つに分類している．最適訓練行動列はある入力に対して正解構造を出力することのできる行動列である．一般的には正解出力が得られるように遷移システムを設計するため，既存の遷移システムに関しては正解データに対する最適訓練行動列を仮定できることが多い．ただし，最適訓練行動列が複数存在することがある．一方，最適訓練方策を仮定する場合はあらゆる状態において最適行動を知ることができる．最適訓練行動列と対比すると，正解データを出力する時には観測されない状態に対しても最適行動を知ることができるということである．たとえば，誤った行動を取った後には正解データを出力する行動列では観測されない状態に到達することがある．その状態であってもその後の誤りを最小限にする行動を正解情報として受け取ることを仮定する．これはより強い仮定であり，タスクや遷移システムによっては現実的な時間では最適訓練行動を得ることができない．

最適訓練行動列が得られる場合はエキスパートの行動列から報酬関数を推定する逆強化学習を用いることが可能になり，最適訓練方策が得られるタスクではエキスパートの行動列の近傍のみを効率的に探索する模倣学習を用いることが可能になる．次節以降では，逆強化学習と模倣学習の NLP への適用事例を紹介する．2 節では NLP の代表的な構造化予測タスクに対して提案されている遷移システムを紹介する．3 節では逆強化学習，4 節では模倣学習アルゴリズムを解説する．最後に，5 節でまとめと今後の展望を議論する．

2. 自然言語処理タスクでの決定過程

本節では NLP タスクに状態・行動を割り当てて決定過程として扱う代表的な手法を例とともに示す．なお，以下のタスクはすべて文の単位でのエピソード的決定過程となる．

言語は表面上は 1 次元の構造をしており，基本的には一方向に読み進むように書かれている．そのため，長さ T の入力列 $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_T)$ に対してラベル列 $(y_1, y_2, \cdots, y_T)$ を予測する系列ラベリング問題として定式化されるタスクが多い．例えば，品詞タグ付けや固有表現抽出などは，文の各単語に対応する品詞，固有表現の種類（人名・地名など）といったラベルの列を予測する系列ラベリング問題として定式化される．文頭から順に決定的に予測する場合には，点 t における行動 a_t は時刻 t の単語に対応するラベルそのものとなり，状態 s_t は入力と $t-1$ までの予測ラベル列，$(\boldsymbol{x}, \hat{\boldsymbol{y}}_{1:t-1})$ を特徴ベクトルで表現する．ただし，$\boldsymbol{y}_{1:t-1} = (y_1, y_2, \cdots, y_{t-1})$ かつ $\hat{y}_t$ は点 t での方策が予測したラベルである．行動は各単語のラベルそのものであるから，系列ラベリング問題では最適訓練行動列は正解データのラベル列を直接利用できる．また，最適訓練方策は状態とは独立に正解データだけを参照して点 t の正解ラベルを返すルールとして構成できる [5]．

一方，構文解析は言語の再帰的な構造を解析する．構文解析の正解データは木構造であり，木を構築する行動列は陽に記述されていない点が大きく異なる．図 3 は係り受け構造解析における代表的な遷移システム（Arc-eager 法 [4]）

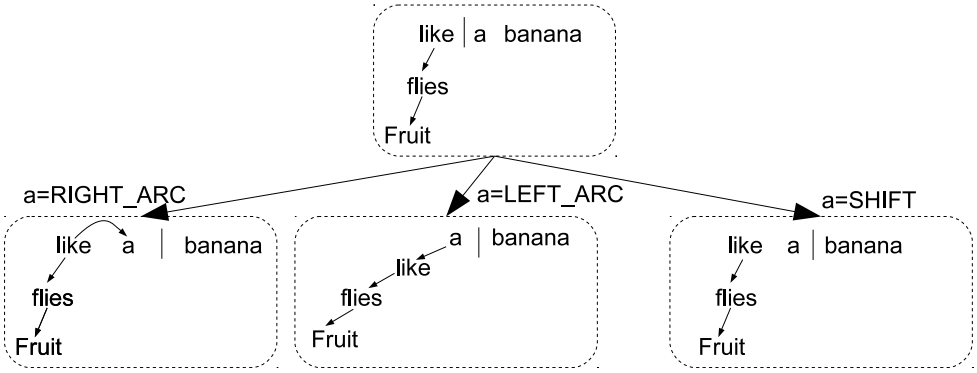

図 3　係り受け構造解析（Arc-eager 法 [4]）における決定過程の例
縦棒は判断点を示す．図上部は “like” まで部分的に解析した時の状態，図下部は矢印上のラベルが示す行動によって遷移した状態．各行動による状態遷移は次の通り：**RIGHT_ARC**) “like” から “a” の向きに依存関係を張り判断点を進める，**LEFT_ARC**）“a” から “like” の向きに依存関係を張り判断点を進める，**SHIFT**） “a” と “like” の間に依存関係を張らずに判断点を進める，**REDUCE**） “like” を判断点から外し，“like” の親を判断点におく．図 1 の係り受け構造を作るには **SHIFT** を選ぶのが正解．なお，図上部の状態では “like” の親が存在せず **REDUCE** 行動は選択できない．

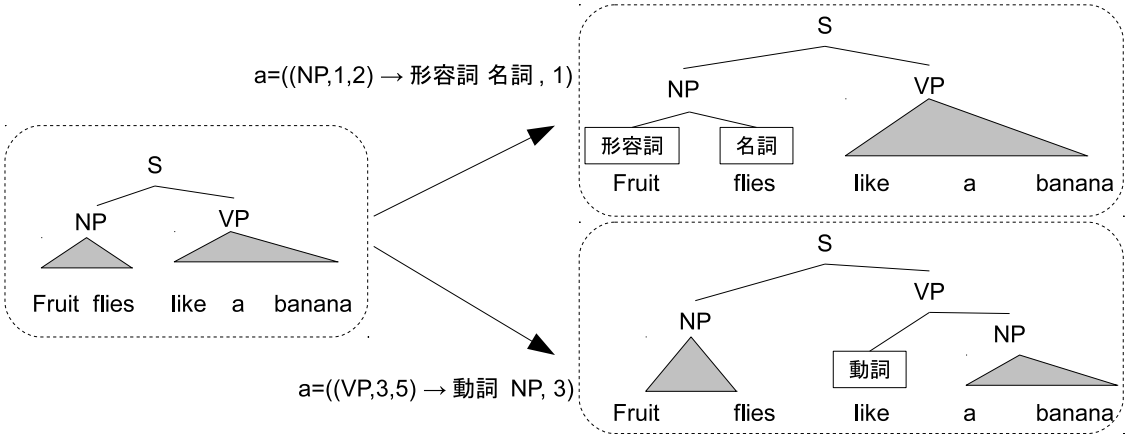

図 4　句構造解析における最適行動の疑似曖昧性
左図は元の状態，右は 2 つの状態遷移後の状態を示す．どちらの行動を先に行っても図 1 の正しい句構造を構成することができる．なお，$((\mathbf{O}, \boldsymbol{b}, \boldsymbol{e}) \rightarrow \mathbf{P}\ \mathbf{Q}, \boldsymbol{s})$ は位置 $\boldsymbol{b}$ から $\boldsymbol{e}$ までの非終端記号 $\boldsymbol{O}$ を位置 $\boldsymbol{s}$ で $\boldsymbol{P}$ と $\boldsymbol{Q}$ に分割する文脈自由文法の生成ルール．

の例である．部分的に構築された木構造および係り受け関係を判断する点以降の単語列を状態とし，行動は判断点の前後の単語の係り受け関係と次に判断点に置くべき単語を同時に決定する．文頭から状態遷移を繰り返すことで一文の単語数 W に対して高々$2W-1$ 回の状態遷移で係り受け構造木を作ることができ，高速な構文解析器となる．最適訓練行動列は正解の木構造から構成できる．また，Arc-eager 法に限り最適訓練方策が線形時間のアルゴリズムとして得られることが示されている [6]．なお，系列ラベリング問題と異なり，複数の行動列が同じ木構造を構築し得るため最適行動は一意に定まらない（疑似曖昧性）．

また，句構造解析では部分的に構築された木構造および単語列が状態となり，行動は文法規則である [7]．最適訓練行動列は正解の木構造から多項式時間で構成できるが，最適訓練方策については筆者らが知る限りでは提案されたものはない．また，係り受け構造解析と同様に最適行動は一意に定まらない．図 4 に句構造解析において異なる行動列が同じ句構造を構成する例を示す．図左の状態において，NP の展開（行動 $a = ((\mathrm{NP}, 1, 2) \rightarrow$ 形容詞 名詞$, 1)$）と VP の展開（$a = ((\mathrm{VP}, 3, 5) \rightarrow$ 動詞 NP$, 3)$）のどちらを先にしても最終的には図 1 の句構造を構成することができる．

なお，疑似曖昧性への対処方法としては，行動に順序付けをして最適行動を一意に正規化する方法[7]，ランダム選択 [5]，訓練中の方策で選択する方法 [6] などが使われている．

3. 逆強化学習による自然言語処理

以上のように，決定過程で構造化予測問題を記述できることはわかったが，強化学習の手法を適用するためには，この決定過程に対して報酬関数を設計する必要がある．しかし，この報酬関数の設計は自明ではない．単純には，各

判断点での最適訓練行動との違いを報酬に対応付けることが考えられるが，部分構造の間の依存性が表現できないため，構造全体の最適性が表現できない．そこで，正解データから得られる最適訓練行動列をもとにして，逆強化学習によって報酬関数を設計する手法が研究されている．

方策を現在の状態 $s \in S$ から次の行動 $a \in A$ への写像 $\pi : S \to A$ と定義する．また，長さ T の状態列・行動列をそれぞれ $\boldsymbol{s}, \boldsymbol{a}$ と書き，NLP への適用では環境が既知のため決定的な遷移システム $\tau(s,a) : S \times A \to S$ を仮定する．多くの逆強化学習アルゴリズムでは報酬関数を線形関数 $r_{\boldsymbol{\theta}}(s,a)$ で近似する：

$$r_{\boldsymbol{\theta}}(s,a) = \boldsymbol{\theta}^\top \boldsymbol{\phi}(s,a).$$

ただし，$\boldsymbol{\phi}(s,a) : S \times A \to \Re^d$ は d 次元特徴ベクトルへの写像，$\boldsymbol{\theta} : \Re^d$ は d 次元パラメータである(注1)．本来は，報酬関数と方策は独立であるが，逆強化学習の NLP への適用では報酬関数を用いて $\pi_{\boldsymbol{\theta}}(s) = \mathrm{argmax}_{a \in A}\, r_{\boldsymbol{\theta}}(s,a)$，または報酬が最も大きくなる行動列 $\mathrm{argmax}_{\boldsymbol{a}} \sum_t r_{\boldsymbol{\theta}}(s_t, a_t)$ を予測することが多い [7]．

逆強化学習手法は，最適訓練行動列が与えられていることを仮定し，報酬関数 $r_{\boldsymbol{\theta}}(s,a)$ の下での行動列と最適訓練行動列とで定義される目的関数を最小化する．以下では，逆強化学習手法の代表的なものとして Maximum Margin Planning（MMP）[8] と最大エントロピー逆強化学習（ME-IRL）[9] を紹介する(注2)．エピソード i に対する最適訓練行動列 $(\boldsymbol{s}^{(i)}, \boldsymbol{a}^{(i)})$ が与えられているとき，MMP は以下の目的関数を最小化する報酬関数を推定する．

$$\sum_i \left(\sum_{(\hat{s},\hat{a}) \in (\hat{\boldsymbol{s}}^{(i)}, \hat{\boldsymbol{a}}^{(i)})} r_{\boldsymbol{\theta}}(\hat{s},\hat{a}) + l(\boldsymbol{a}, \boldsymbol{a}^{(i)}) - \sum_{(s,a) \in (\boldsymbol{s}^{(i)}, \boldsymbol{a}^{(i)})} r_{\boldsymbol{\theta}}(s,a) \right)$$

ただし，$(\hat{\boldsymbol{s}}^{(i)}, \hat{\boldsymbol{a}}^{(i)})$ はエピソード i に対する報酬の総和 $\sum_{t=1}^{T} r_{\boldsymbol{\theta}}(s_t, a_t) + l(\boldsymbol{a}, \boldsymbol{a}^{(i)})$ を最大化する状態列と行動列である．直感的には MMP では最適訓練行動列の報酬と他の行動列の報酬の差 (マージン) が最大になるような報酬関数を求めていることになる．なお，$l(\boldsymbol{a}, \tilde{\boldsymbol{a}}) : A^T \times A^{\tilde{T}} \to \Re_+$ は行動列と行動列の距離を返す関数で，$l(\boldsymbol{a}, \tilde{\boldsymbol{a}})$ に合わせてマージンは調整される．

また，最大エントロピー逆強化学習（ME-IRL）では対数線形モデルに報酬関数を使用したモデルを提案している:

$$P(\boldsymbol{a}|\boldsymbol{\theta}) = \frac{\exp\left(\boldsymbol{\theta}^\top \sum_{t=1}^{T} \boldsymbol{\phi}(s_t, a_t)\right)}{Z(\boldsymbol{\theta})}.$$

(注1) ここでは文献 7) に倣い s, a を入力とする報酬関数を想定するが，τ の逆写像 $\tau^{-1}(s) : S \to S \times A$ が存在することを仮定すると，$r(s_t, a_t)$ は時刻 $t+1$ の状態 s_{t+1} への即時報酬と等価である．

(注2) 説明では目的関数から省いているが，バリアンスを減らすために L2 正則化などを併用する

ただし，$Z(\boldsymbol{\theta})$ は分配関数である．ME-IRL は最適訓練行動列に対する次式の負の対数尤度を最小化する:

$$-\sum_i \log P(\boldsymbol{a}^{(i)}|\boldsymbol{\theta}). \tag{1}$$

式 1 の勾配は最適訓練行動列と報酬関数下での特徴ベクトルの差であるが，凸関数であるので最小化時にはその差（勾配）はゼロとなるため，以下の定理により逆強化学習として望ましい性質を持つ．

Theorem 1 方策 π の下での特徴ベクトルの期待頻度を $\mu_\pi = \mathbb{E}[\sum_t^T \boldsymbol{\phi}(s,a)|\pi]$ とし，π と π^* の特徴ベクトルの期待頻度の異なりが ϵ 以下，つまり $||\mu_\pi - \mu_{\pi^*}||_2 \leq \epsilon$ であるならば，線形の報酬関数を想定すると次式の報酬の期待値の差も ϵ 以下である [10]:

$$\left| \mathbb{E}\left[\sum_t^H r(s,a)|\pi\right] - \mathbb{E}\left[\sum_t^H r(s,a)|\pi^*\right] \right| \leq \epsilon.$$

MMP も ME-IRL も最適訓練行動列の報酬和が最大になるような報酬関数，つまり構造全体の整合性を重視した報酬関数になっており構造化予測問題に適している．文献 7) も指摘するように，実はこれらの逆強化学習手法は，NLP で扱う問題設定では以前から適用されてきた構造化予測のためのアルゴリズムとほぼ同じである．たとえば，(指数個の候補の中から）最大報酬行動列 $\hat{\boldsymbol{a}}$ や分配関数 Z を効率的に計算できる遷移システムの場合には，MMP はマージン最大化構造化予測 [11] と，ME-IRL は条件付き確率場 [12] と同等である．NLP では，変数間の依存関係が分解可能であることを仮定して動的計画法を用いる，組み合わせ最適化問題を線形計画問題に緩和して解く，などの近似を用いて最大報酬行動列や分配関数を計算することでマージン最大化構造化予測や条件付き確率場が用いられてきた．なお，文全体の整合性を考慮した構造化予測を実現するために，学習時だけでなく解析時（テスト時）も報酬和が最大になる行動列を予測することが一般的である．

これらの方法は文全体の整合性を満たす方策が得られるが，変数群の局所的な条件付き独立性を仮定する必要があり，また多項式時間であっても計算量が非常に大きいなどの課題点があった．次節では最適訓練方策を仮定することで局所的な条件付き独立性を仮定する必要のない手法を紹介する．

4. 模倣学習による自然言語処理

模倣学習(注3)はエキスパートの行動列を参照して方策を学習する手法の総称であり，前節で述べた逆強化学習もその意味では模倣学習の一種といえる．本節では特に報酬関数の推定は行わずに，エキスパートの行動列を参照して状

(注3) 徒弟学習（Apprenticeship Learning）とも呼ばれる．

態・行動空間を効率的に探索する手法の一つである Dataset Aggregation（DAgger）アルゴリズム [13] を紹介する．

DAgger は方策反復（Policy Iteration）と呼ばれるアルゴリズムの一つで，各反復で最適訓練方策とそれまでに学習済みのすべての方策が訪れた状態を訓練データとして方策を学習する．DAgger の最初の反復では，既存の教師付き学習と同じように最適訓練行動列によって観測された状態を訓練データとして方策を学習する．次の反復以降では，最適訓練行動列によって観測された状態に加えて学習した方策が訪れた状態も訓練データに加えて学習する．その際，方策は最適訓練行動列とは違う行動をとりえるので，最適訓練行動列によっては観測されない未知の状態に対しても正しい行動を示すことができる最適訓練方策が必要になる．DAgger をアルゴリズム 1 に示す．ただし，β_i は状態遷移を行う際に使用する最適訓練方策と訓練している方策との混合方策の混合率で $N \to \infty$ のとき $\frac{1}{N}\sum_k \beta_k \to 0$ を満たす数列である．混合率は $\beta_1 = 1, \beta_k = 0 (k > 1)$，つまり初回は最適訓練方策 π^* を用い，2 回目以降は現在の方策を用いる方法が経験的に良いことが報告されている [13), 14)]．

最適訓練行動列だけを使って方策を学習してしまうと方策は誤った行動の後に行動を選択することを想定していない．そのため，一か所の行動の誤りがそのあとの行動に影響し続けて間違ってしまう**誤り伝搬**の問題が指摘されていた．“Fruit flies like a banana.”（図 1）の品詞タグ付けを例にとると，Fruit の品詞を名詞と誤って予測した場合には，英語の正解データ上では名詞-動詞の並びが多く次の flies の品詞を動詞として予測するのが自然なため，予測履歴の特徴を重視したモデルでは誤りが連続してしまい，続く “like/動詞 a/冠詞 banana/名詞” と整合性を取ることができない．一方，DAgger は方策が実際に訪れるであろう状態をサンプリングしその下での経験誤差を最小化するため，方策が選ぶ行動が最適訓練行動列と異なる，つまり過去の行動に誤りがあった場合にもそれ以降の行動は誤らないことが期待できる．先の例では，方策が Fruit の品詞を名詞として誤ったときであっても flies の品詞は名詞であることを最適訓練方策により教示することになり，誤りに影響されにくいように名詞-動詞の並びの特徴を重視しない方策が得られる．理論的にも，決定数（予測数）を T としたとき，最適訓練行動列だけを使って方策を教師付き学習したとき誤差の上界は $O(T^2)$ であるが， DAgger の誤差は $O(T)$ で抑えられることが示されている [13)]．

文献 15) では NLP タスクにおいて逆強化学習に基づく手法と比較を行っており， DAgger の実用上の利点として次の 2 点を確認している．

1. 決定的解析では DAgger の方が逆強化学習に基づく手法よりも構造予測精度が高い
2. 現実的な計算時間で局所的でない構造情報を特徴として使用できる

Algorithm 1 DAgger

初期化: $D \leftarrow \emptyset, \pi_1$ は任意の方策
for $k = 1, 2, \cdots, K$ **do**
  $\pi_k \leftarrow \beta_k \pi^* + (1 - \beta_k)\hat{\pi}_k$
  π_k を実行し $D_k = \{(\boldsymbol{\phi}(s_{\pi_k}), \pi^*(s_{\pi_k}))\}$ を収集
  データを集約: $D \leftarrow D \cup D_k$
  D を用いて $\hat{\pi}_{k+1}$ を学習
end for
検証用データで性能の良い $\hat{\pi}_k$ を選択

利点 1 における決定的解析とは各判断点で最もスコアの高い行動を選ぶことによって最大報酬行動列を近似する高速な解析方法（貪欲法）である．Web 規模のテキストデータを処理するには解析が速いことは実応用で重要であるため，決定的解析手法は好まれて用いられている．文献 15) では英語係り受け構造解析の実験において，逆強化学習に基づく手法では解析速度を上げるために決定的解析を採用すると構造予測精度が大きく下がってしまい，DAgger による決定的解析の方が構造予測精度が高いことを示している．また，利点 2 に関しては，逆強化学習に基づく手法では計算量の問題から $1, 2$ 単語前の局所的な予測ラベルだけを特徴量に用いるのが系列ラベリング問題においては一般的であった．しかし，DAgger による決定的解析では現実的な処理時間でより長い 4 単語前までの予測ラベル列を使うことができることを示している．このように，逆強化学習に基づく手法と比較すると，模倣学習は NLP タスクでの決定的解析において優位性を持っていることが示されている．

5. おわりに

本稿では自然言語処理タスクと強化学習の関連性，および既存の訓練データを活用する方法として逆強化学習と模倣学習を紹介した．逆強化学習はこれまで NLP に適用されてきた構造化予測アルゴリズムと同様に文全体の整合性を満たす行動を選択する方策を学習でき，模倣学習手法として紹介した DAgger は決定的な解析であっても誤りの連鎖を引き起こしにくい方策を学習できることを解説した．最後に今後の研究の方向性について議論したい．

最初に，多くの NLP タスクでは最適訓練行動列または最適訓練方策を得ることができる遷移システムを設計することは自明ではなく，個々のタスク毎に設計する必要がある．機械翻訳・自動要約・質問応答・対話システムのような自然言語を出力するタスクに適用することは，人間の言語生成を模倣するシステムを設計・学習することを意味しており言語学的にも興味深い研究課題である．別の方向性としては，文献 16), 17) ではタスクを解くための決定過程ではなく解析速度と精度のトレードオフや動的に特徴選択を行うための決定過程を強化学習を用いて学習している．このように解析システムの様々な面において逐次決定過程としての定式化と最適化の可能性がある．

次に，逆強化学習が特に有効であるのは，方策よりも報酬関数の推定が容易であるタスクであると予想する．2 節で例を示したが，遷移システムによっては複数の行動列が同一の解析結果を生むような疑似曖昧性が存在する．これまで NLP に適用されてきた逆強化学習では，報酬関数は状態と行動を引数としてきたが本来は報酬関数は状態に対して定義されるのが一般的である．その場合，疑似曖昧性が高い遷移システムでは行動を選択するための方策よりも各状態を評価する報酬関数の方が簡潔に表現できることが期待できる．また，歴史的には構造化予測に対するアルゴリズムが逆強化学習手法に大きな影響を与えてきた．近年の NLP タスクにおいては，局所的な構造から段階的に大域的な構造の予測を行う手法が関心を集めており [18),19)]，逆強化学習への展開も興味深い方向性である．

最後に，本稿では正解データがあるタスクのみを紹介したが，興味深い発展として教師なし構造化予測問題に方策反復を用いた研究がある [20)]．教師なし学習では正解データが与えられていないことを想定し，オートエンコーダのように予測した構造（隠れ状態）を経由して元の文を再構成した時の誤差（reconstruction error）が小さくなるように方策を学習する．文献 20) では教師なし係り受け構造解析に適用し，他の教師なし手法に比べて少ない反復で収束することを示している．NLP においても教師なし学習は長年の課題であり，教師なし構造化予測問題へのアプローチとして今後の発展が期待される．

（2013 年 9 月 11 日受付）

参 考 文 献

1) スティーヴン・ベイカー：IBM 奇跡のワトソンプロジェクト: 人工知能はクイズ王の夢をみる, 早川書房 (2011)
2) A. McCallum, D. Freitag and F. Pereira: Maximum entropy Markov models for information extraction and segmentation, in *Proceedings of the 17th International Conference on Machine Learning*, 591/598 (2000)
3) F. Maes, L. Denoyer and P. Gallinari: Structured prediction with reinforcement learning, *Machine Learning Journal*, **77**–2-3, 271/301 (2009)
4) J. Nivre: An efficient algorithm for projective dependency parsing, in *Proceedings of the 8th International Workshop on Parsing Technologies*, 149/160 (2003)
5) H. Daumé III, J. Langford and D. Marcu: Searn in practice, Unpublished (available at `http://pub.hal3.name/`) (2006)
6) Y. Goldberg and J. Nivre: A dynamic oracle for arc-eager dependency parsing, in *Proceedings of the 24th International Conference on Computational Linguistics*, 959/976 (2012)
7) G. Neu and C. Szepesvári: Training parsers by inverse reinforcement learning, *Machine learning*, **77**–2-3, 303/337 (2009)
8) N. Ratliff, J.A.D. Bagnell and M. Zinkevich: Maximum margin planning, in *Proceedings of the 23rd International Conference on Machine Learning*, 729/736 (2006)
9) B.D. Ziebart, A. Maas, J.A. Bagnell and A.K. Dey: Maximum entropy inverse reinforcement learning, in *Proceedings of the 23rd national conference on Artificial intelligence*, AAAI Press, 1433/1438 (2008)
10) P. Abbeel and A.Y. Ng: Apprenticeship learning via inverse reinforcement learning, in *Proceedings of the 21st international conference on Machine learning*, New York, NY, USA, 1/8 (2004), ACM
11) B. Taskar, V. Chatalbashev, D. Koller and C. Guestrin: Learning structured prediction models: a large margin approach, in *Proceedings of the 22nd international conference on Machine learning*, 896/903 (2005)
12) J. Lafferty, A. McCallum and F. Pereira: Conditional random fields: Probabilistic models for segmenting and labeling sequence data, in *Proceedings of the 18th International Conference on Machine Learning*, 282/289 (2001)
13) S. Ross, G.J. Gordon and D. Bagnell: A reduction of imitation learning and structured prediction to no-regret online learning, in *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, 627/635 (2011)
14) A. Vlachos: An investigation of imitation learning algorithms for structured prediction, in *Proceedings of the 10th European Workshop on Reinforcement Learning*, (2012)
15) 坪井祐太：模倣学習による決定的解析での誤り伝播の回避, 言語処理学会第 19 回年次大会, (2013)
16) J. Jiang, A. Teichert, H. Daumé III and J. Eisner: Learned prioritization for trading off accuracy and speed, in *Advances in Neural Information Processing Systems 25*, 1340/1348 (2012)
17) H. He, H. Daumé III and J. Eisner: Imitation learning by coaching, in *Advances in Neural Information Processing Systems 25*, 3158/3166 (2012)
18) D. Weiss and B. Taskar: Structured prediction cascades, in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, (2010)
19) A.M. Rush and S. Petrov: Vine pruning for efficient multi-pass dependency parsing, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 498/507 (2012)
20) H. Daumé III: Unsupervised search-based structured prediction, in *Proceedings of the 26th Annual International Conference on Machine Learning*, 209/216 (2009)

［著 者 紹 介］

坪井 祐太

2002 年 奈良先端科学技術大学院大学 情報科学研究科 博士前期課程修了．同年日本アイ・ビー・エム (株) 入社．2009 年 奈良先端科学技術大学院大学 情報科学研究科 博士後期課程修了．博士（工学）．日本アイ・ビー・エム東京基礎研究所にてテキストマイニングの研究開発に従事．

牧野 貴樹

2002 年 東京大学 大学院理学系研究科 情報科学専攻 博士課程修了．博士 (理学)．2005 年–2011 年，東京大学総括プロジェクト機構領域創成/学術統合化プロジェクト研究部門学術統合化プロジェクト (ヒト) 特任助教．2011 年より東京大学生産技術研究所最先端数理モデル連携研究センター特任准教授．機械学習，コミュニケーション数理モデルの研究に従事．