H_004

強化学習とベイズ推定による言語的コミュニケーション学習モデル

A learning model of language communication by reinforcement learning and Bayes estimation

荒木 隆史† 坪根 正† 和田 安弘†
Takafumi Araki Tadashi Tsubone Yasuhiro Wada

1. はじめに

ヒトは自己の意図・要求を他者に伝える際に，言葉，手話，文字などの様々な言語的手段を用いる．例えば，"オレンジジュースを飲みたい"という意図を言葉 (日本語) で他者に伝える際には，「私はオレンジジュースが飲みたい」などと発話する．このとき，単に「ジュース」や「欲しい」，「飲む」といった単語のみの提示であったり，「ジュースがオレンジ飲みたい私は」といった順序では他者に上手く意図を伝えることが困難となる．「私はオレンジジュースが飲みたい」と正しい順序で発話することで正確に意図が伝わる．つまり，他者に"オレンジジュースを飲みたい"という意図を伝えるためには，「私」「オレンジ」「ジュース」「飲む」という単語 (シンボル) の獲得と，これらの単語を用いて，意図を表現するような順序の単語列を生成することが必要である．先行研究において，書字のような複雑な運動パターンを獲得するモデルが提案されている [1][2]．これらのモデルは手話で考えると，単語となる運動パターンの獲得の可能性を示している．

本報告では，要求 (意図) を表現する単語列の生成に焦点を当てる．ヒトは要求を表現できる単語の順序を，生まれつき知っているのではなく，また他者から文法を習うわけでもないと考えられる．既に獲得している単語を並べて発話してみて，その単語列が要求を表現できていれば要求が満たされ (上記の例では他者からオレンジジュースがもらえる)，表現できなければ要求を満たすことはできない．我々は要求に対して，このような正しい単語列の生成と誤った単語列の生成を繰り返し行うことで，ヒトが徐々に自己の要求を表現するような単語の順序を学習していくと仮定する．つまり，強化学習的な枠組みによって単語の順序を学習していくと考える．

ヒトは初め，要求を1語で表現するが，徐々に2語，3語，$\cdots$ と長い単語列を生成するようになると考えられる．本論文では，これにヒントを得て，意図を表現できる単語の順序を学習するモデルを提案する．提案モデルは，強化学習アルゴリズムとベイズ推定アルゴリズムを組み合わせた枠組みで構築した．本モデルは，N 個の単語から成る単語列 (N 語文) において，要求を表現するような単語の順序を，環境との相互作用を通して学習するモデルである．

2. 単語列学習モデル

モデルは強化学習の1つである *Actor*・*Critic* モデル [3][4] で構成され (図1)，*Actor*(図1の上破線枠) と

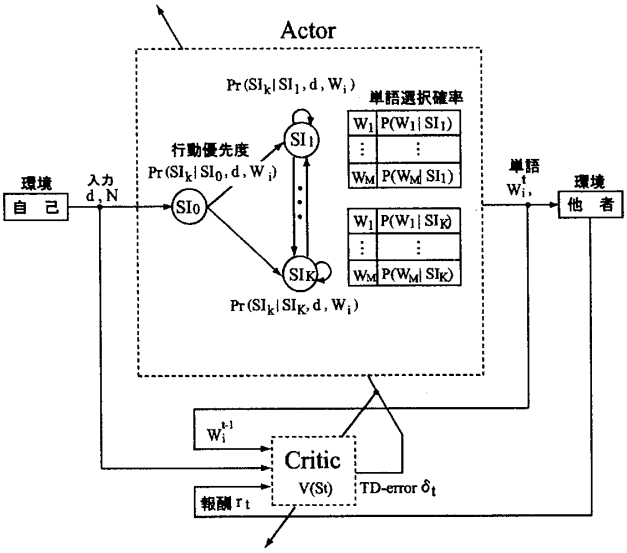


図 1: 学習モデル

Critic(図1の下破線枠) は2つの環境と相互作用する．環境の1つは学習者自身の中で，要求 (例えば"オレンジジュースが飲みたい"など) が発生する機構と仮定する (以降，この環境を自己と呼ぶ)．一方の環境は，自己が要求 (意図) を伝えようとしているヒトを想定する (以下，この環境を他者と呼ぶ)．本モデルは，1時刻で1つの単語を出力し，N 語からなる単語列の生成を1エピソードとする．

自己は要求 d とその要求を表現するための単語列を構成する単語の数 N を *Actor* と *Critic* へ入力する．d，N が入力されると，*Actor* は時刻 $t(=1,\cdots,N)$ で，K 個の内部状態のいずれか1つへ遷移し，遷移先の内部状態で，モデルが既に獲得している M 個の単語の中から1つを確率的に選択し出力する．この内部状態の遷移順と単語選択確率が，要求に対してどのような順序でどの単語を選択するのかに影響を与える．他者は t で *Actor* が出力した単語 W_i^t に応じて報酬 r_t を *Critic* へ与える．ここで $i(=1,\cdots,M)$ は単語の種類を表すインデックスである．*Critic* は自己から入力された d，N と他者から与えられた報酬 r_t，さらに $t-1$ で *Actor* が出力した単語 W_i^{t-1} を基に *Actor* がとった行動を TD 誤差の形式で評価する．

2.1 *Actor* の構造

Actor は K 個の内部状態 SI を持つネットワークで構成される．各内部状態は互いに，状態遷移の行動の強

†長岡技術科学大学，Nagaoka University of Technology

弱を表す行動優先度 P_r で結合される．また，各内部状態に付属するテーブルは，単語 W_i の状態別の選択確率 $P(W_i|SI_k)$ である．この行動優先度と単語選択確率を学習によって更新することで，要求に応じてどの単語をどのような順序で出力すべきか学習していく．

2.2 *Critic* の構造

Critic は報酬の期待値を表す状態の関数 $V(S_t)$ として表現される．状態 S_t は自己から入力された要求 d と *Actor* が $t-1$ で出力した単語 W_i^{t-1} から1つ定まる $(S_t = (d, W_i^{t-1}))$．*Critic* はこの報酬の期待値 $V(S_t)$ を学習によって更新する．

2.3 モデルの動作

環境 (自己) から入力 d，N が与えられると，現在の内部状態から，*Actor* は式 (1) に基づいて行動優先度 P_r が最大値を示す内部状態へ遷移する．ただし，最大値を示す P_r が複数存在した場合は，それらの中からランダムに選択される．また，P_r の初期値は0とする．

$$SI_{j'} = \arg\max_{SI_k} Pr(SI_k|SI_j, d, W_i^{t-1}) \quad (k = 1, \cdots, K) \tag{1}$$

ここで，SI_j は遷移元の内部状態，$SI_{j'}$ は遷移先の内部状態，W_i^{t-1} は1つ前の時刻で *Actor* が出力した単語を示す．

続いて遷移先の内部状態 $SI_{j'}$ で，単語選択確率 $P(W_i|SI_{j'})$ に従って単語を1つ選択，出力する．ただし，各内部状態における単語選択確率の初期値は，等確率とする (式 (2))．

$$P(W_i|SI_k) = \frac{1}{M} \qquad (i = 1, \cdots, M) \tag{2}$$

ここで，M はモデルが既に獲得している単語数である．

Critic は，時刻 t ごとに *Actor* が出力した単語に応じて，報酬 r_t を他者から受け取る．報酬 r_t は，N 語目の単語を出力したとき ($t = N$ のとき)，生成された N 語の単語列が要求を表現するものであれば1，そうでなければ0とした．また，単語列を生成する過程 ($t \neq N$ のとき) で与えられる報酬は，常に0とした．

Critic は，入力値 d, N と *Actor* が出力した単語 W_i^{t-1}，他者から与えられる報酬 r_t が入力され，式 (3) の TD 誤差 δ_t の形式で *Actor* の行動を評価する．

$$\delta_t = r_t + \gamma V(S_{t+1}) - V(S_t) \tag{3}$$

ここで，$V(S_t)$ は状態 S_t における報酬の期待値，γ は割引率を示す．また，$V(S_t)$ の初期値は0とする．

2.4 モデルの更新

Actor では，内部状態遷移の行動優先度 P_r，単語選択確率 $P(W|SI)$ の両者を更新する．P_r は *Critic* が出力した TD 誤差 δ_t を基にモデルの更新を行い，$P(W|SI)$ の更新にはベイズ推定アルゴリズムを用いる．

P_r は1エピソードにおいて，実際の内部状態遷移に関連したもののみを式 (4) より更新する．

$$Pr(SI_k|SI_j, d, W_i^{t-1}) = Pr(SI_k|SI_j, d, W_i^{t-1}) + \alpha\delta_t \tag{4}$$

ここで，α は学習率を表す．

次に $P(W|SI)$ の更新をベイズ推定アルゴリズムを用いて行う．ベイズ推定は，あるデータを観測する前に観測者の持っている知識 (事前分布) が，データ観測後に変化した結果 (事後分布) を推定するものである．本モデルでは，単語選択確率 $P(W|SI)$ を事前分布，N 語目の単語出力に対する報酬 $R = \{0, 1\}$ をデータとし，1エピソード中に遷移した各内部状態における単語選択確率の事後分布を，ベイズの定理を用いて式 (5) で算出する．

$$P(\hat{W}|R) = \frac{P(R|\hat{W})P(\hat{W})}{\sum_{i=1}^{M} P(R|\hat{W}_i)P(\hat{W}_i)} \tag{5}$$

ここで，$\hat{W}$ は時刻 t で内部状態 SI において選択された単語を表し，$P(\hat{W})$ は $\hat{W}$ の選択確率である．また，$P(R|\hat{W})$ は尤度で，$\hat{W}$ を選択したときに $t = N$ で成功の報酬 $R = 1$ が与えられる確率を表す．尤度 $P(R|\hat{W})$ は式 (6) より算出される．

$$P(R = 1|\hat{W}) = \frac{E_{success,d} + \exp(-\frac{E_{success,d}}{\tau})}{E_{\hat{W},d} + M \cdot \exp(-\frac{E_{success,d}}{\tau})} \tag{6}$$

ここで，$E_{\hat{W},d}$ は過去のエピソードにおいて，要求 d が入力されたとき，$\hat{W}$ が選択されたエピソード数を表す．$E_{success,d}$ は $E_{\hat{W},d}$ の内，成功報酬 $R = 1$ が与えられたエピソード数である．また，M はモデルが既に獲得している単語数を表す．分母，分子それぞれの第二項はノイズを表し，τ はノイズをどの程度考慮するか決定するパラメータである．すなわち，τ を大きくすると学習後期までノイズが残り，小さくすると学習初期でノイズがなくなる．

式 (5) で算出した事後分布を，次のエピソードにおける事前分布とすることで単語選択確率が更新される (式 (7))．

$$P(\hat{W}) = P(\hat{W}|R) \tag{7}$$

Critic では，1エピソード中における各状態 S_t での報酬の期待値 $V(S_t)$ を式 (8) で更新する．

$$V(S_t) = V(S_t) + \beta\delta_t \tag{8}$$

ここで，β は学習率を表す．

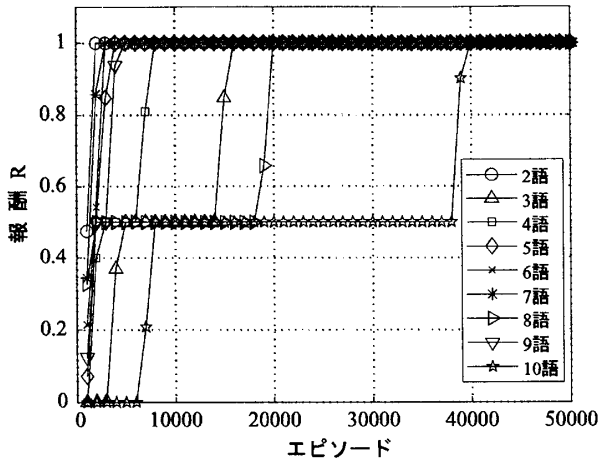


図 2: シミュレーション 1 における報酬の推移

3. シミュレーション

提案モデルが単語の順序を正しく学習できるか確認するため，3 種類のシミュレーションを行った．シミュレーション 1 は 2 種類の要求 d_1，d_2 に対しての $N = 2$(2 語文) から $N = 10$(10 語文) の単語列の学習，シミュレーション 2 は 3 種類の要求 d_1，d_2，d_3 に対する $N = 2$，$N = 3$ の単語列の学習である．シミュレーション 3 は，シミュレーション 2 の条件で学習済みのモデルが，新しく単語を獲得したときの，新たな要求 d_4 に対する単語列の学習である．各シミュレーションにおいてパラメータは，行動優先度の学習率 α を 0.2，報酬の期待値の学習率 β を 0.1，TD 誤差の割引率 γ を 0.9，尤度のノイズ項における τ を 0.3，ネットワークの内部状態数 K を 100 とした．また今回は問題を簡単にするため，単語の代わりに文字 (アルファベット) を用いた．すなわち，1 つのアルファベットがある 1 つの単語を表すと仮定する．

3.1 シミュレーション 1 : 2 語文から 10 語文の学習

2 語文，3 語文，$\cdots$，10 語文の学習を試みる．ここでは，モデルが既に獲得している単語を $a, b, \cdots, t$ のアルファベット 20 個とし，モデルへ 2 種類の要求 d_1，d_2 をランダムに与えた (d_1, d_2 は要求の種類を表し，数値としてモデルへ入力する)．また成功報酬が与えられる単語列は，d_1 に対しては ab(2 語)，abc(3 語)，$\cdots$，$abcdefghij$(10 語) とし，d_2 に対しては kl，klm，$\cdots$，$klmnopqrst$ とした．学習回数は 2 語から 10 語までそれぞれ 50000 エピソードずつで，最初に 2 語を 50000 エピソード，次に 3 語を 50000 エピソード，$\cdots$ と徐々に長い単語列を学習させた．

学習の経過を観測するため，N 語目の単語出力に対する報酬 R の推移を図 2 に示す．ここで，各プロットは 1000 エピソードごとの報酬 R の平均値を示している．学習開始時は低い報酬値を示すことが多いが，その後学習を重ねていくと学習後期では 2 語から 10 語いずれの長さの単語列も常に 1 を示し続けている．これは，学習初期では入力された要求を表現できない単語列を生成することが多いが，学習を重ねることで，要求を正しく表現するようにモデルが単語の順序を学習したことを表しているといえる．

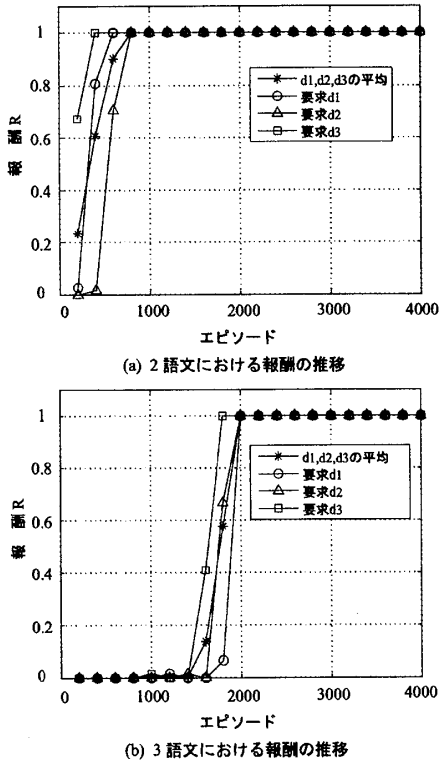


図 3: シミュレーション 2 における報酬の推移

よって，提案モデルは 2 種類の要求に対して，それぞれを正しく表現するような 2 語から 10 語文を学習することができると言える．

3.2 シミュレーション 2 : 3 種類の要求に対する学習

ここでは，3 種類の要求に対する 2 語文，3 語文の学習を試みる．モデルが既に獲得している単語を a，b，$\cdots$，i の 9 個とし，モデルへ 3 種類の要求 d_1，d_2，d_3 をランダムに与えた．また成功報酬が与えられる単語列は，d_1 に対しては bc，abc とし，d_2 に対しては ef，def，さらに d_3 に対しては hi，ghi とした．学習回数は初めに 2 語を 4000 エピソード学習した後で，3 語を 4000 エピソード学習させた．また，要求が入力される順序によって学習経過や結果が異なるため，要求の入力系列を変更して 30 回シミュレーションを行った．その結果，モデルが常に成功報酬を与えられるようになるまでに要した学習回数は，2 語文の学習において最小 296 回，最大 1421 回，平均 796.4 回であった．一方 3 語文の学習では最小 1794 回，最大 13301 回，平均 5975.9 回であった．

一例としてシミュレーションを 30 回行った内，2 語文と 3 語文を学習するのに要した合計エピソード数が最小のときの，2 語，3 語文それぞれにおける報酬 R の推移

を図3に示す．各プロットは200エピソードごとの平均値を示す．2語文については，学習開始時にはいずれの要求も低い報酬値を示すが，800エピソード学習を重ねると全ての要求で成功報酬1が与えられるようになった．また要求ごとの推移を見ると，要求 d_3，d_1，d_2 の順で1に収束していることがわかる．これはモデルが要求を1つずつ学習していったことを表している．一方，三語文では2000エピソード以降，どの要求に対しても常に正しい単語列を生成していることがわかる．

これらのことから，提案モデルは要求の種類が3つの場合においても，各要求を表現できる単語列を学習することができると言える．

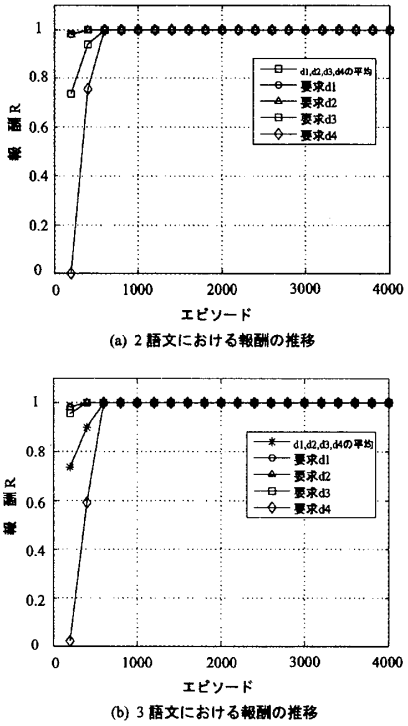
3.3 シミュレーション3：モデルが新しく単語を獲得した場合

前節で，3種類の要求に対して2，3語文を学習する場合，提案モデルが正しく動作することを確認した．ここでは，1度ある要求を表現する単語列を学習したモデルが，新たに単語を獲得した場合を想定し，それら新規な単語から成る単語列を学習することができるかシミュレーションする．すなわち，シミュレーション2で2，3語文を4000エピソードずつ学習したモデルが，新たに3つの単語 j，k，l を獲得したときを考える．

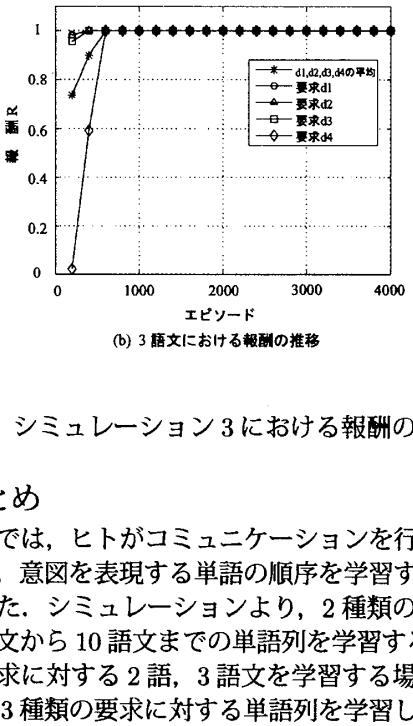
モデルが獲得している単語を a，b，$\cdots$，l の12個とする．ただし，各内部状態における単語選択確率 $P(W|SI)$ の初期値を，新しく獲得した j，k，l については1/12とし，残りの9個は総和が1になるように正規化した値とする．モデルへは4種類の要求 d_1，d_2，d_3，d_4 がランダムに入力される．また成功報酬が与えられる単語列は，d_1 に対しては bc(2語)，abc(3語) とし，d_2 に対しては ef，def，d_3 に対しては hi，ghi，さらに d_4 に対しては kl，jkl とした．ただし，モデルは d_1，d_2，d_3 を表現できる単語列はシミュレーション2で学習済みである．そして，要求の入力系列を変更して，30回シミュレーションを行った．その結果モデルが常に成功報酬を与えられるようになるまでに，2語文の学習において最小186回，最大4449回，平均877.8回のエピソード数を要した．3語文の学習では最小90回，最大23499回，平均3094.7回であった．

30回シミュレーションを行った中で，2語文と3語文を学習するのに要した合計エピソード数が最小のときの報酬 R の推移を図4に示す．各プロットは200エピソードごとの平均値である．2，3語文の両者において，既に学習済みの d_1，d_2，d_3 は学習開始時から高い報酬値を示していることがわかる．要求 d_4 に関しては学習初期では誤りの報酬0が与えられることが多いが，学習後期では2語文，3語文共に常に成功報酬1を示し続けている．

これらより，提案モデルは新しく単語を獲得したときに，新たな要求に対する単語列を学習することができると言える．



(a) 2語文における報酬の推移



(b) 3語文における報酬の推移

図4: シミュレーション3における報酬の推移

4. まとめ

本報告では，ヒトがコミュニケーションを行う際に必要となる，意図を表現する単語の順序を学習するモデルを提案した．シミュレーションより，2種類の要求に対する2語文から10語文までの単語列を学習する場合，3種類の要求に対する2語，3語文を学習する場合，さらには一度3種類の要求に対する単語列を学習したモデルが新たに単語を獲得した場合において，提案モデルが正しく動作することを確認した．

今回行ったシミュレーションでは，常に成功報酬が与えられるようになるまでにおよそ1000単位でのエピソード数が必要であった．今後の課題として学習速度の向上が挙げられる．

参考文献

[1] Y. Wada. and M. Kawato. : A theory for cursive handwriting based on the minimization principle. Biological Cybernetics, 73(1):3-13(1995)

[2] K. Tokunaga. , Y. Wada. : A model for movement pattern acquisition process, " , SICE Annual Conference, P-44, pp. 1323-1327(2004)

[3] Sutton, R. S. , Barto, A. G. : Reinforcement Learning:An Introduction. Cambridge, MA:A Bradford Book, MIT Press(1998)

[4] Barto, A. G. , Sutton, R. S. , and Anderson, C. W. : Neuronlike adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst. , Man, and Cybernetics, 13:834-846(1983)