助言付きビデオゲーム環境における助言の意味とゲームプレイの同時学習

○宮尾愛平　岡夏樹　田中一晶（京都工芸繊維大学）

Simultaneous Learning of the Meaning of Advice and Game Play in a Video Game Environment with Advice

∗N. Miyao, N. Oka and K.Tanaka (Kyoto Institute of Technology)

Abstract– "The use of multiple dialogue acts" is an understudied area in the study of language evolution. The purpose of this study is to verify the effectiveness of advising, which consists of multiple dialogue acts, in a situation where language and action are learned simultaneously, as the first step in a study that aims to elucidate the language evolution process constructively. In this study, we set up a game learning task with advice, in which the use of action instruction and evaluation is considered to be advantageous. In this task, a teacher agent advises a student agent who is learning game operations and word meanings. In this study, we propose a student model that learns the meaning of two types of advice, action instructions and evaluations, and uses them for game learning. The student model was implemented using deep reinforcement learning and used action instructions as a part of the state representation and evaluation as a reward or punishment. In this study, we used the Atari 2600 Breakout to evaluate the effect of advice on the learning speed of the game. Experimental results demonstrated that using two types of advice, action instructions and evaluations, facilitated game learning.

Key Words: deep reinforcement learning, language learning, meaning estimation, advice, video game

1　はじめに

1.1　背景・目的

我々人間は，相互に他者の意図を理解しようと日常的に意図推定を行っており，これによって，協調的行動などの社会的コミュニケーションが実現される．人間は他者の行動や発話，ジェスチャー等の情報を観察し，自己の持っている経験と照らし合わせることで他者の意図を推定し，円滑なコミュニケーションを図っている．相手の発話の対話行為（質問，依頼，申し出，陳述のような発語内行為のレベルでの発話意図）を推測できることは，意図共有の第一歩として重要となる．

言語進化研究において，従来のシミュレーション研究や実験室実験で使用されたタスクのほとんどは，1 種類の対話行為（陳述）だけでタスクが達成できるもの [1] か，複数の対話行為（指示 or 評価）が使われうるタスクであるが，意思疎通が成立するのは発話の種類が少数に収斂した場合という結果 [2] であった．従って，「複数の対話行為の使い分け」については，まだ研究が手薄な分野となっていると言える．そこで，本研究では，複数の対話行為の使い分けが有利なタスクを考案し，シミュレーション実験による言語進化プロセスの構成論的な解明を目指す．

1.2　構想

複数の対話行為が有利になるタスクとして，「助言付きゲーム学習」タスク（Fig. 1 参照）を考える．このタスクでは，生徒と教師の 2 体のエージェントが存在する．生徒はゲームのプレイヤとして操作を学習し，教師は生徒にゲーム操作の助言を行う．教師には事前にゲーム操作を学習させておき，ある程度のスコアを獲得できるようにしておく．このとき，助言には「この操作を行ったほうが良い」のような指示と「今の操作は良かった」のような評価の 2 種類の意図（対話行為）が含まれ得る．教師と生徒との間にコミュニケーション手段が確立されていないとすると，タスクの実行に伴って 2 体のエージェントは，高いスコアが獲得できるように，自らメッセージ意味付けを行うはずである．このタスクでは，その過程で「指示」と「評価」の 2 種類の助言，つまり，複数の意図を表現する言語が発生（創発）することを期待する．また，学習が終了した生徒はゲーム能力と言語能力を引き継いだ状態で，次の教師として新しい生徒の助言を行う．これによって，教師が学習した言語が生徒に伝わり，言語獲得が行われる．さらに，世代交代が繰り返されることによる言語の複雑化も取り扱うことができるため，より現実世界に近い言語進化現象の観察が期待できる．

本稿では，言語進化研究の前段階として，助言付きゲーム学習タスクを簡単化する．具体的には，教師の学習は考えず，教師が既に言語能力を有している状態を想定する．つまり，生徒は，常に同じ意味を持つメッセージを受け取るため，生徒が助言を利用してゲーム操作を効率的に学習できるのかという問題になる．本稿では，簡単化した助言付きゲーム学習タスクをシミュレーションすることで，ゲーム学習に対する助言の有効性を検証する．また，そのために，指示と評価の 2 種類の助言の意味を学習し，ゲーム学習に活用する生徒モデル提案する．

2　提案手法

2.1　タスクの形式化

提案タスクは，エピソード形式，離散時間，離散状態，離散行動の強化学習タスクとして表現する．単一のエージェントと環境の相互作用で表される通常の強化学習において，エージェントは，各タイムステップ t で環境から状態 s_t を観測し，自身の持つ方策 $\pi(a_t|s_t)$ に基づいた行動 a_t を選択する．a_t によって環境が変化すると，エージェントは，次の状態 s_{t+1} と環境報酬 r^e_{t+1} を観測する．本研究では，教師を環境の一部とし

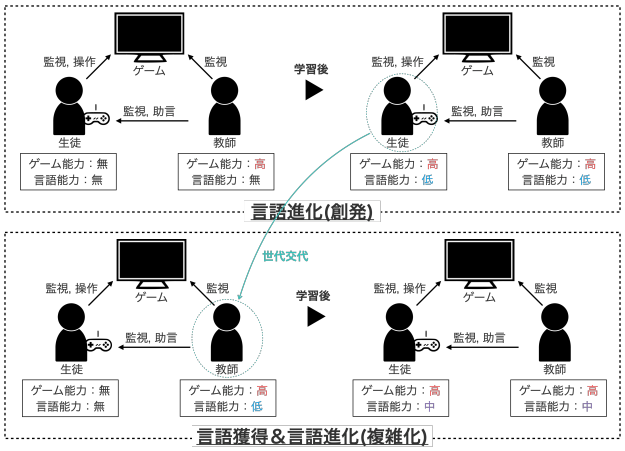

Fig. 1: Game learning task with advice.

て捉えることで，エージェント（生徒）が環境からメッセージ（助言）を観測する拡張を 2 通り行う．

1 つ目は，a_t の選択前後で 1 回ずつメッセージを受け取ることが可能な設定である．a_t 前のメッセージ m_t^b は s_t に対する行動の指示を示し，行動選択後のメッセージ m_t^a は a_t に対する評価を示す．生徒にとって，メッセージの役割が既知となっているため，この設定を「対話行為が既知の設定」と呼ぶこととする．

2 つ目は，a_t の選択前に 1 回だけメッセージを受け取ることが可能な設定である．対話行為が未知の設定では，a_t 前のメッセージ m_t が a_{t-1} に対する評価なのか，s_t に対する行動の指示なのかが分からない．生徒にとって，メッセージの役割が未知となっているため，この設定を「対話行為が未知の設定」と呼ぶこととする．

2.2 教師モデル

教師は，One-hot ベクトルのメッセージを生成する．ここで，指示を示すメッセージを行動指示語，評価を示すメッセージを行動評価語と呼ぶことにする．行動指示語は，教師が選択した行動を示す．行動評価語は，教師が選択した行動と生徒が選択した行動が一致していたのかを示す．一致していた場合，生徒が正しい行動を選択したこと意味する「GOOD 評価」，そうでない場合，生徒が間違った行動を選択したこと意味するを「BAD 評価」を示す．対話行為が既知の設定の場合，行動指示語は s_t に対して，行動評価語は a_t に対して生成される．対話行為が未知の設定の場合，行動指示語は s_t に対して，行動評価語は a_{t-1} に対して生成される．助言を生成には，予めゲーム操作の学習済みモデルを作成し，その出力を利用する．また，学習済みモデルの作成には，代表的な深層強化学習手法の 1 つである A2C アルゴリズム [3] を使用する．

2.3 生徒モデル

生徒は，操作学習器と報酬学習器の 2 種類のモジュールで構成されている．以下では，それぞれのモジュールについての説明を行う．

2.3.1 操作学習器

操作学習器は，s_t と m_t から a_t を決定する[1]．学習アルゴリズムは，A2C[3] を採用し，ニューラルネットワークを用いて，方策 $\pi(a_t|s_t, m_t)$ と状態価値関数 $V(s_t, m_t)$ を表現する．生徒は，割引累積和 $R = \sum_{t=0}^{\infty} \gamma(r_{t+1}^e + r_{t+1}^i)$ の期待値を最大化する．ここで，環境報酬 r_{t+1}^e，評価報酬 r_{t+1}^i，割引率 γ である．

今回は，複雑なゲーム画面のピクセル情報を状態として扱うため，単純なメッセージに従うように方策が表現できれば，効率良く行動を学習することが期待できる．従って，操作学習器では行動指示語を活用することを想定した設計となっている．

2.3.2 報酬学習器

報酬学習器は，メッセージから r_{t+1}^i を生成する．r_{t+1}^i は，メッセージの「評価推定指標（メッセージの GOOD 評価らしさ，BAD 評価らしさを測る指標）」を算出し，評価推定指標を評価報酬に変換することで得る．評価推定指標 p_{V_n} は，以下の手順で算出する．準備として，メッセージの語彙数（One-hot ベクトルの次元数）|V| と同じ数だけ大きさ L のメモリを用意し，メモリと語彙を 1 つずつ対応させておく．生徒は，各タイムステップで $\pi(a_t|s_t, m_t)$ を m_{t+1} と対応するメモリに保存していく[1]．ただし，保存数がメモリの大きさを超えた場合，最も古いデータを削除し，新しいデータを保存する．語 V_n に対応するメモリでは，保存された行動選択確率の平均値 π_{V_n} を算出する．算出した全メモリの平均値を正規化することで，V_n の評価推定指標 p_{V_n} を得る．p_{V_n} は，1 に近い値ほど GOOD 評価らしいことを示し，0 に近い値ほど BAD 評価らしことを示す．

これまでの説明で明らかになっている通り，報酬学習器では行動評価語を活用することを想定した設計となっている．行動評価語が適切に報酬として使用されれば，環境報酬のみの場合と比較して，操作学習器のパラメータをより素早く更新できることが期待できる．評価推指標を評価報酬に変換する方法は，後述する．

3 関連研究

言葉の意味獲得に関する先行研究の多く [4, 5, 6] は，物や動作などの参照的な意味を取り扱っている．言葉は，聞き手に影響を与えるような機能的な意味も持っており，この両方の意味を獲得できることが重要である [7] が，機能的な意味の獲得の研究は限られている．鈴木ら [8] は，行動の事後指示（「右に行くべきだった」のような我々の研究とは異なる役割を持った指示）と評価の 2 種類の言葉が混在する状況で意味学習を行った．言葉の意味を学習するエージェントは，同時にタスクを強化学習で学習しており，エージェントが学習した価値に基づいて言葉の意味を決定している．しかし，意味を学習した助言がエージェントの行動学習には活用されていない．我々の研究と同じく，指示や評価を計算機によって生成しており，1. 様々な言い回しが考慮されていない，2. 誤った助言が与えられない，3. 助言を与えない選択肢が考慮されていない，4. 助言の生起頻度に偏りがない，といったように，実際の助言の性質とは離れている．一方，岡ら [9] は，自然な発話を含むインタラクションデータから指示と評価の 2 種類の言葉の意味学習を行う方法を提案し，上記の 4 つの問題を改善している．しかし，助言の意味学習のみにとどまっており，意味を学習した助言の活用はなされていない．一方，Najar ら [10] は，タスクの学習と助言の意味の同時学習に取り組んでおり，意味を学

[1] 対話行為が既知の設定の場合，m_t を m_t^b，m_{t+1} を m_t^a に置き換える．

習した助言は，行動の学習に活用されている．しかし，指示と評価の 2 種類の言葉を取り扱っているが，学習の対象には，指示しか含まれていない．我々の研究は，行動と指示と評価の 2 種類の対話行為が含まれる助言の意味を学習し，さらに，助言を行動の学習に活用するエージェントの設計に取り組んでいる．

4　シミュレーション実験

エージェントが学習するゲームは，ALE[2]で提供されているブロック崩しゲームの「Atari2600 Breakout」とした．また，教師の事前学習モデルは，「OpenAI Baselines」[11] を使用し，40M タイムステップの学習を行うことで作成した．学習した重みを使用し，30 エピソード実行したときの平均スコアは 388.3 ± 20.89 である．このとき，方策は離散確率分布で表現されているため，最も確率の高い行動を選択した．助言も同じく，最も確率の高い行動を選択し，生成した．以下では，実施した 2 つの実験の方法と結果を述べる．

4.1　実験 1：複数の対話行為からなる助言の有効性の検証

生徒モデルにおける操作学習器の学習は，報酬学習器が生成する評価報酬に大きく依存する．従って，報酬学習器の性能自体がゲームの操作学習に大きな影響を与える．そこで最初に行う実験では，行動評価語の意味が正しく推定できる理想的な報酬学習器を仮定する．これは，教師から評価報酬が直接与えられることを意味する．この実験では，指示と評価の両方の助言を使用することが，ゲーム操作を高速に学習するための有用な手段であることを確認する．

4.1.1　方法

対話行為が既知の設定では，2（メッセージの内容）× 4（評価報酬の大きさ）の 8 条件を設定し，40M タイムステップの学習を行う．メッセージの内容は，zero 条件（4 次元の零ベクトル），instruct 条件（行動指示語を表す 4 次元の One-hot ベクトル）とした．評価報酬の大きさは，{GOOD 評価,BAD 評価 }={+0,-0}, {+0.01,-0.01}, {+0.1,-0.1} or {+1,-1} とした．

対話行為が未知の設定では，メッセージの内容と与え方を変化させた，次の 5 条件を設定し，40M タイムステップの学習を行う：no-advice 条件（零ベクトルのみを与える）, instruct 条件（行動指示語を表す 4 次元の One-hot ベクトルのみを与える），evaluate 条件（行動評価語を表す 2 次元の One-hot ベクトルのみを与える），both-alternate 条件（行動指示語もしくは行動評価語を表す 6 次元の One-hot ベクトルを指示と評価が交互になるように与える），both-random 条件（行動指示語もしくは行動評価語を表す 6 次元の One-hot ベクトルを指示と評価がランダムに切り替わるように与える）．今回は理想的な報酬学習器を想定しているため，行動評価語が与えられた場合，{good,bad} = {+0.1,-0.1} の評価報酬を与える．評価報酬の大きさは，後述する対話行為が既知の設定での実験結果において，教師の平均スコアまで最も早く学習することを確認した instruct&{+0.1,-0.1} 条件の値を採用した．

[2]Arcade Learning Environment：Atari2600 ゲームを使用した AI を開発するための簡単なインターフェースを提供するプラットフォーム

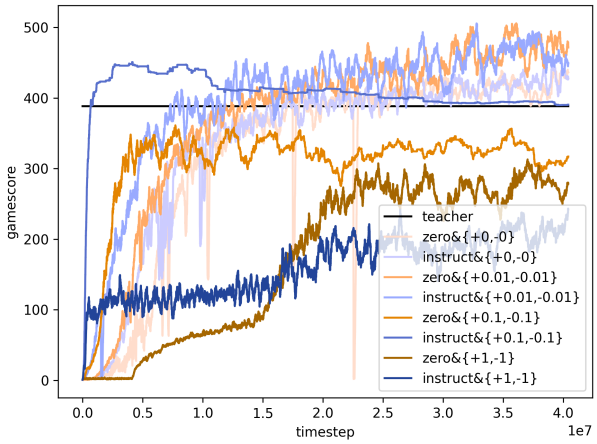

Fig. 2: Results of Experiment 1. Transitions of the game score in each condition when the dialogue act is known.

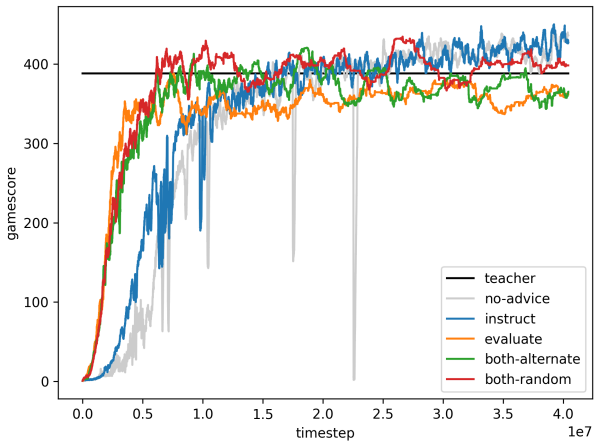

Fig. 3: Results of Experiment 1. Transitions of the game score in each condition when the dialogue act is unknown.

4.1.2　結果

Fig. 2 は，対話行為が既知の設定における生徒の学習曲線を示す．また，Fig. 3 は，対話行為が未知の設定における生徒の学習曲線を示す．どちらも縦軸が獲得スコア，横軸が学習時間であり，教師の平均スコアを示す teacher（黒線）がプロットされている．zero&{+0,-0} 条件と no-advice 条件，instruct&{+0,-0} 条件と instruct 条件は，それぞれ，同じ曲線である．生徒の獲得スコアが教師の平均スコアを超える速度を学習速度と定義すると，どちらの設定においても，指示と評価の両方を使い分けた設定が最も早く学習できている．

4.2　実験 2：生徒モデルの評価

次の実験では，提案した報酬学習器を用いた生徒の学習を行う．この実験では，本研究で提案する生徒モデルが，指示と評価の 2 種類の助言を使い分け，ゲーム操作学習を促進できるかを確認する．

4.2.1　方法

対話行為が既知の設定では，生徒の行動選択前に行動指示語（4 次元の One-hot ベクトル），行動選択後に

行動評価語（2 次元の One-hot ベクトル）を与え，40M タイムステップの学習を行う．

対話行為が未知の設定では，行動指示語もしくは行動評価語を示すメッセージ（6 次元の One-hot ベクトル）を与え，40M タイムステップの学習を行う．ただし，教師は，指示と評価をランダムに切り替えるとした．

評価報酬は式 (1) を用いて生成する．この式では，全ての語の評価推定指標の平均値 $\frac{1}{|\mathrm{V}|}$ より評価推定指標が大きくなるほど，大きな正の評価報酬に変換し，小さくなるほど，小さな負の評価報酬に変換する．上記の設定の実験は，評価報酬を制限した上で一度実施している．その結果では，対話行為が既知の設定と対話行為が未知の設定のどちらの場合も，BAD 評価は-0.1〜-0.2，GOOD 評価は 0.1〜0.2 の評価報酬に変換されていた．この値は，実験 1 で最も早く生徒が学習できたときの評価報酬の値（0.1）に近く，この実験でも同程度の評価報酬の値に変換されるとすれば，生徒は高速にゲーム操作を学習することが可能なはずである．

$$r^i_{t+1} = \frac{p_{V_n} - \frac{1}{|\mathrm{V}|}}{\frac{1}{|\mathrm{V}|}} = p_{V_n}|\mathrm{V}| - 1 \qquad (1)$$

4.2.2 結果

Fig. 4 は，生徒の獲得スコアの遷移を示している．縦軸が獲得スコア，横軸が学習時間であり，教師の平均スコアを示す黒線（teacher），助言を与えない場合=実験 1 の zero&{+0,-0} 条件の生徒の獲得スコアの遷移を示す灰線（no-advice 条件）がプロットされている．濃青線（proposal 条件 1）は，実験 2 の対話行為が既知の設定で学習させた場合の生徒の獲得スコアの遷移を示す．また，濃橙線（proposal 条件 2）は，実験 2 の対話行為が未知の設定で学習させた場合の生徒の獲得スコアの遷移を示す．薄青線（ideal 条件 1）と薄橙線（ideal 条件 2）は，実験 2 の条件で報酬学習器を使用せず，GOOD 評価のときに+0.1，BAD 評価のときに-0.1 の評価報酬を与えたときの生徒の獲得スコアの遷移が示されている．ideal 条件は，実験 1 の対話行為が既知の設定における zero&{+0,-0} 条件，Fig. 4 における ideal 条件は，実験 1 の対話行為が未知の条件における both-random 条件と同じものがプロットされている．

Fig. 5 上は，Fig. 4 の proposal 条件の評価推定指標の遷移を示している．また，Fig. 5 下は，Fig. 4 の proposal 条件の評価推定指標の遷移を示している．ラベル「indicator”数字”」の数字は，どの語に対応しているのかを示す（0:BAD 評価，1:GOOD 評価，2〜5: 行動指示語）．border は，評価報酬が 0 となる値を示しており，評価推定指標が border より高い場合，正の評価報酬に変換され，border より低い場合，負の評価報酬に変換される．

どちらの設定においても，proposal 条件が no-advice 条件の学習速度を超えており，ideal 条件ともほとんど変わらない速度となっていることが確認できる．しかし，対話行為が未知の設定においては，学習の後半で GOOD 評価の意味の推定を間違えてしまっている．

5 考察

5.1 実験 1：複数の対話行為からなる助言の有効性の検証

対話行為が既知の設定

対話行為が既知の設定では，最も早い学習を実現した instruct&{+0.1,-0.1} 条件で，獲得スコアがピークに到達した後に，徐々に減少していた．原因としては，生徒が環境報酬よりも評価報酬の獲得を優先したことが考えられる．ブロック崩しゲームは，ブロックが崩せない状態が続いたとしても，ペナルティが与えられない．そのため，ブロックが崩せない期間は，生徒に評価報酬のみが与えられることになる．もし生徒に与えられる評価報酬の合計が正であるならば，生徒は，ゲームオーバーにさえならないように行動を選択していれば，将来得られる報酬を無限に増加させることができる．すると，環境報酬が無視できる状態となり，評価報酬の獲得を優先するように方策が変化する．その結果，instruct&{+0.1,-0.1} 条件のように，獲得スコアが減少し始めると考えられる．獲得スコアが教師の平均スコアを超え，ピークに到達してから減少し始める原因については，次の 2 点が考えられる．1 つ目は，ブロック崩しゲームの仕様上，ある程度ブロックの数が減らないとブロックが崩せない状況にならないということ（つまり，ゲーム開始直後は，ボールを打ち返すと，必ずブロックに当たってスコアが獲得できる．）である．2 つ目は，残りのブロックの数が少なくなるほど，的確にボールを飛ばさないとブロックを崩せなくなるため，環境報酬が獲得しづらくなり，評価報酬の獲得を優先しやすい状況になるということである．

instruct&{+0.1,-0.1} 条件が他の条件の学習曲線よりも値の変動が少なく，安定している理由については，生徒の方策がほとんど決定的になっていることが挙げられる．評価報酬によって，生徒の方策が過剰に強化されることで，教師と同じ行動の選択確率が非常に高くなる．また，教師は決定的に行動を選択するため，生徒の行動も同じく決定的となる．結果，獲得スコアの分散が下がり，獲得スコアの安定につながったと考えられる．では，教師と同じ行動を選択する生徒が，なぜ教師の平均スコアを超えることができるのかという

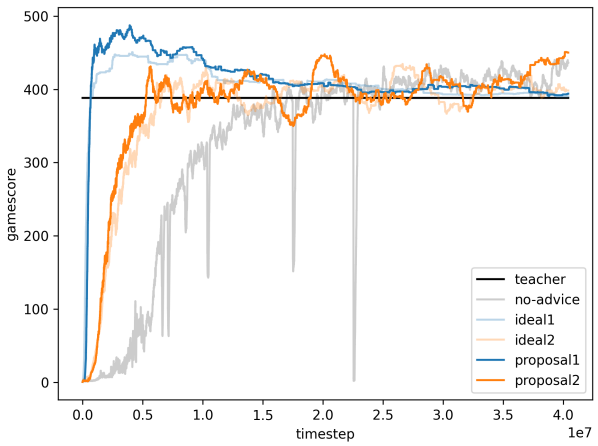

Fig. 4: Results of Experiment 2. Transitions of the game score in each condition.

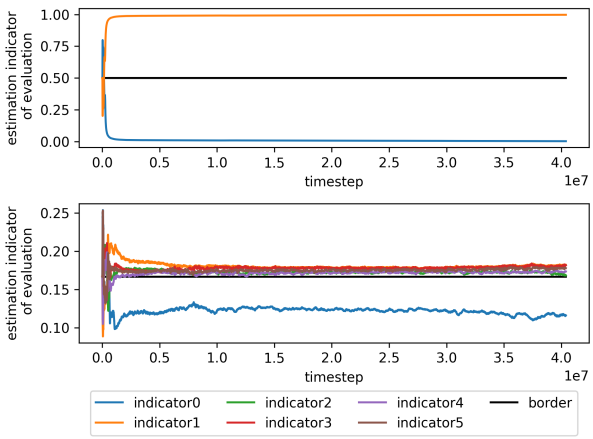

Fig. 5: Results of Experiment 2. Transitions of the estimation indicator of evaluation in each condition.

疑問が残る．この疑問に対しては，次の説明が考えられる．教師は決定的に行動を選択するため，一度スコアが獲得できない状況に遭遇すると，その状況が無限に繰り返されてしまう．しかし，そのループを抜け出すことができれば，教師はボールを跳ね返す位置にパドルを動かすことを学習しているはずなので，さらに獲得スコアを上昇させることができるはずである．生徒が未知の状態を体験した場合，生徒の方策は，環境報酬や評価報酬によって強化されていないため，エントロピーが高く，探索が行われやすい状況であると考えられる．よって，教師が無限ループに陥るような状況であったとしても，生徒が教師とは異なる行動を選択するため，無限ループを回避することができる．無限ループを回避した後の生徒の行動が教師の行動と同じだと，無限ループを回避する行動（教師とは異なる行動）も評価報酬によって強化される（なぜなら，強化学習では，将来得られる報酬を最大するような行動を強化するため）．これを繰り返すと，教師よりも生徒が高いスコアを獲得することが可能である．ただし，生徒も無限ループに陥ってしまうことで，次第に評価報酬を受ける生徒の行動が教師の行動に近づいていき，評価報酬を最大化する方策に変化することで，獲得スコアの減少していくと考えられる．

対話行為が未知の設定

対話行為が未知の設定においては，指示と評価の両方を与える both-random 条件と both-alternate 条件の学習速度が最も早いが，途中までは評価のみを与える evaluate 条件の方が僅かに学習速度が早いという結果であった．途中まで evaluate 条件の学習速度が最も早かった理由には，指示よりも評価の方がパラメータ更新に対する影響が大きことが挙げられる．instruct 条件と evaluate 条件を見比べれば分かる通り，獲得スコアの増加率は報酬を直接与える evaluate 条件の方が圧倒的に高い．both-random 条件と both-alternate 条件は，instruct 条件と evaluate 条件を組み合わせた条件であるため，学習速度が instruct 条件と evaluate 条件との間になると考えれば，納得できる結果と言える．

evaluate 条件では，生徒の獲得スコアが教師の平均スコアを超えられなかったのに対して，同じ大きさの評価報酬が与えられる both-random 条件と both-alternate 条件では，教師の平均スコアを超えていた．evaluate 条件では，環境報酬を無視して評価報酬を最大化するような方策を学習していると考えられるが，both-random 条件と both-alternate 条件では，評価報酬が与えられる頻度が evaluate 条件の半分であるため，評価報酬の獲得を優先するような方策が学習しにくい状況であり，今回のような結果となったと考えられる．

5.2 実験 2：生徒モデルの評価

対話行為が既知の設定

対話行為が既知の設定では，学習開始直後，GOOD 評価は負の報酬，BAD 評価は正の報酬が与えられているが，学習が進むことで，GOOD 評価は正の報酬，BAD 評価は負の報酬となるように，修正することができていた．この理由については，次の説明が考えられる．学習開始直後では，獲得スコアが上昇するような行動をとると，罰が与えられるため，学習が不安定であると考えられる．生徒は，学習がより安定する方向に（安定して報酬が獲得できる状況，つまり，獲得スコアが上昇するような行動をとると，評価報酬が与えられる状況になるように）ネットワークの重みを変更しようとするため，GOOD 評価の評価推定指標が高く，BAD 評価の評価推定指標が低くなるように，方策が変化していく．結果，GOOD 評価は正の報酬，BAD 評価は負の報酬となるように，修正されたと考えられる．

proposal 条件が教師の平均スコアを超えている理由については，実験 1 の instruct&{+0.1,-0.1} 条件と同様の理由が考えられる．proposal 条件が ideal 条件の獲得スコアの最大値を上回っている理由については，ネットワークの初期の重みといったランダムな要素に影響を受けていると思われるが，他にも，proposal 条件の評価報酬が変化し続けるのに対して，ideal 条件の評価報酬が常に同じ値であることが挙げられる．評価報酬が変化するということは，助言の価値が変動することを意味し，これによって方策が不安定になるはずである．従って，評価報酬が変化する場合，教師が生徒よりも高いスコアを獲得した後の方策のエントロピーが高くなり，探索が行われやすくなる．結果，生徒と教師が同じ行動を選択する確率が低くなり，ブロックを崩せない状況のループに陥りにくくなったと考えられる．

対話行為が未知の設定

対話行為が未知の設定でも，同様の理由で，学習開始直後に，GOOD 評価は負の報酬，BAD 評価は正の報酬が与えられていたものが，GOOD 評価は正の報酬，BAD 評価は負の報酬となるように，修正ができていると考えられる．しかし，学習が進むにつれて，GOOD 評価の評価推定指標と行動指示語の評価推定指標が近い値をとるように変化していた．これは，評価報酬が与えられることで，教師と同じ行動の選択確率のみが強化され，生徒が同じ行動ばかりを選択するようになってしまった結果であると考えられる．そのような方策では，行動指示語は，行動選択確率の高い行動と共起しやすくなる．つまり，GOOD 評価と共起する行動の選択確率と行動指示語と共起する行動の選択確率が近づくことを意味する．逆に，生徒が教師と異なる行動を選択した場合は，BAD 評価が共起するため，教師と同じ行動の選択確率のみが強化された方策であるため，教師と異なる行動の選択確率は低い．従って，BAD 評

価の評価推定指標が非常に低い値となっていると考えられる．

6　今後の展望

実験では，提案した生徒モデルに次の 2 つの問題が見つかった：1. 生徒が環境報酬を無視して評価報酬を最大化する方策を学習する場合がある，2. GOOD 評価の推定が上手くできなくなる場合がある．1. の問題は，報酬シェーピングに関する先行研究で議論されており，改善方法が提案されている [12]．これらの方法を適用すれば，生徒が評価報酬に依存してしまう問題を改善することが可能なはずである．しかし，その場合は，各条件の学習曲線が適用する前と異なる振る舞いをすることが予想される．従って，指示と評価の両方を使い分けた方が有利になるかの検証が再度必要となるだろう．2. の問題は，3 通りの改善方法が考えられる．1 つ目は，報酬学習器が推定の対象とする語を評価のみに限定する方法である．対話行為が既知の場合の実験結果では，評価の推定に問題がなかったことから，対話行為が評価であることが識別できる手段を新しく用意することができれば，本研究で提案したアルゴリズムをそのまま使用することができる．2 つ目は，新しく GOOD 評価を推定するためのアルゴリズムを考案する方法である．GOOD 評価が推定できなくなる一方，BAD 評価に関しては，問題なく識別できているため，本研究で提案したアルゴリズムは，BAD 評価用の報酬学習器として使用することができる．3 つ目は，GOOD 評価の意味が失われること防ぐ方法である．最も簡単なものは，生徒の方策のエントロピーの平均などを取っておき，その値が一定値を下回った場合に，評価推定指標の更新を停止するという方法である．これにより，GOOD 評価と行動指示語の評価推定指標が近い値になることを防げるはずである．現状，最も具体的な案となっている 3 つ目の方法を適用することが，この問題が改善に向けた最初のステップとなるだろう．

言語進化研究を目指す上では，改善すべき課題が山積みとなっている．我々の提案した生徒モデルは，状態とメッセージの両方を与えなければ行動が決定できない．我々の構想では，生徒が次の世代の教師となるモデルを考えており，生徒にも助言を生成する能力が必要となるため，メッセージが与えられない状況だとしても，生徒が行動を決定できる学習アルゴリズムが必要である．また，生徒は，教師の助言の意味を学習し，活用できても，自ら助言を生み出すことはできない．世代を超えた言語獲得現象を観察するためには，既存の言語が再利用できるような学習アルゴリズムが必要である．さらに，生徒に合わせて適切な助言を選択する教師の学習アルゴリズムが必要である．

7　まとめ

本研究では，複数の対話行為の使い分けに関する言語進化プロセスを構成論的に解明するという目的のもと，行動指示と評価の使い分けが有利となる「助言付きゲーム学習」タスクを提案した．また，言語進化研究の第一歩として，簡単化した助言付きゲーム学習タスクをシミュレーションするこで，言語と行動を同時に学習する場面における，複数の対話行為からなる助言の有効性を検証した．さらに，指示と評価の 2 種類の助言の意味を学習し，ゲーム操作学習に活用する生徒モデルを提案した．Atari2600 のブロック崩しゲームを使用し，助言がゲーム学習速度に与える影響を評価した結果，指示と評価の 2 種類の助言の使い分けがゲーム学習を促進することを確認した．しかし，生徒が評価報酬を最大化する方策を学習したり，評価の意味推定を失敗するなどの課題が残った．今後は，本稿で明らかとなった問題を修正しながら，言語進化現象に焦点を当てた研究に着手する予定である．

謝辞

本研究は JSPS 科研費 JP20H05004 の助成を受けたものである．

参考文献

1) S. Kirby, H. Cornish, and K. Smith, “Cumulative cultural evolution in the laboratory: An experimental approach to the origins of structure in human language,” Proceeding of the National Academy of Sciences, vol.105, no.31, pp.10681–10686, Sep. 2008.
2) 小松孝徳, 鈴木健太郎, 植田一博, 開一夫, 岡夏樹, “パラ言語情報を利用した相互適応的な意味獲得プロセスの実験的分析,” 認知科学, vol.10, no.1, pp.121–138, 2003.
3) V. Mnih, A.P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” International conference on machine learning, pp.1928–1937, 2016.
4) 岩橋直人, “ロボットによる言語獲得: 言語処理の新しいパラダイムを目指して,” 人工知能学会誌, vol.18, no.1, pp.49–58，2003.
5) D. Roy, “Learning from sights and sounds: a computational model,” Ph.D. Thesis, MIT Media Laboratory, Cambridge, 1999.
6) C. Yu, “A multimodal learning interface for grounding spoken language in sensory perceptions,” ACM Trans. Applied Perceptions, vol.1, no.1, pp.57–80, 2004.
7) D. Roy, “Semiotic schemas: a framework for grounding language in the Action and Perception,” Artificial Intelligence, vol.167, no.1–2, pp.170–205, 2005.
8) 鈴木健太郎, 植田一博, 開一夫, “自律的な行動学習を利用した評価教示の計算論的意味学習モデル,” 認知科学, vol.9, no.2, pp.200–212, 2002.
9) 岡夏樹, 増子雄哉, 林口円, 伊丹英樹, 川上茂雄, “Fisher の直接法を用いたインタラクションデータからの意味学習,” 知能と情報（日本知能情報ファジィ学会誌）, vol.20, no.4, pp.461–472, 2008.
10) A. Najar, O. Sigaud, and M. Chetouani, “Interactively shaping robot behaviour with unlabeled human instructions,” Autonomous Agents and Multi-Agent Systems, vol.34, pp.1–35, 2020.
11) P. Dhariwal, C. Hesse, O. Klimov, A. Nichol, M. Plappert, A. Radford, J. Schulman,S. Sidor, Y. Wu, and P. Zhokhov, “Openai baselines,” 2017.
12) B. Xiao, B. Ramasubramanian, A. Clark, H. Hajishirzi, L. Bushnell, and R. Poovendran, “Potential-based advice for stochastic policy learning,” 2019 IEEE 58th Conference on Decision and Control (CDC), pp.1842–1849, IEEE, 2019.