1A2-02

並列学習を利用した対話戦略の獲得

Dialogue Strategies Acquisition using Parallel Learning

田口 亮 Ryo Taguchi　桂田 浩一 Kouichi Katsurada　新田 恒雄 Tsuneo Nitta

豊橋技術科学大学 大学院工学研究科
Graduate School of Engineering, Toyohashi University of Technology

Abstract: When applying reinforcement learning to dialogue strategy acquisition, we must design its state space based on various information about dialogue such as dialogue history, facial expression of opponent, and so on. In general, this design work is very hard for us because it requires a lot of trial and errors. Moreover, even if we can design a complex state space, long term learning will be required to acquire an efficient strategy from such complex state space. There might be some simple semi-efficient strategies that could be acquired from smaller state space. In order to find such semi-efficient strategy quickly and to find more efficient strategy according to progress of learning, we propose a parallel learning method in which multiple Q-learning modules with different state space are executed. We introduce an evaluated value onto each module for estimating its performance. By determining the agent action according to the module with highest evaluation value, the agent can select efficient act in any stage of learning. The experimental results show that the agent can act efficiently according to learning stage.

1. はじめに

我々は，計算機が人間との対話を円滑に進めるために必要な対話戦略を，強化学習によって計算機に自動獲得させる研究を行っている [田口 03]. 強化学習を対話戦略の獲得に応用する場合，対話相手の要求や意図を直接参照することはできないため，相手の言動や表情，しぐさ，対話履歴等の多様な情報を組み合わせて，相手の要求や意図を間接的に表現しうる状態空間を設計しなければならない. その際，状態空間に用いる情報が少ないと，効率的な戦略を獲得させることはできず，逆に情報が多いと複雑で効率的な戦略を獲得させることはできるが，戦略の獲得に多くの学習時間が必要となる. 従って設計者は，獲得される戦略の性能と学習時間も考慮して最良の状態空間を選定しなければならない. また，実世界での学習では環境への即応性が求められるため，時間をかけて一つの複雑な戦略を獲得するよりも，学習時間に応じて単純な戦略から複雑な戦略へと段階的に獲得されることが望ましい. しかし，Q 学習をはじめとした一般的な強化学習の手法では，状態空間に対する最適な戦略は獲得されるが，その過程で獲得される戦略の有効性については考慮されていない.

そこで本報ではこの状態空間の選定と，段階的な戦略の学習を実現する方法として，状態空間の異なる複数の学習器を用いた並列学習を提案する. 提案手法は一回の行動とその結果得られる報酬を利用して，複数の学習器を同時に学習させる. そして，一回の試行で得られた報酬を基に試行に用いた学習器の評価値を計算し，行動選択に用いる学習器を段階的に切り替えていく. 一回の試行で複数の学習器を同時に学習させることができるため，状態空間の選定に必要な実験コストが低減できると考えられる. また，段階的に戦略を切り替えていくことで環境への即応性が高まる. 後述の実験では本手法をエージェント同士の対話学習に適用し，その有効性を示す.

連絡先：Ryo Taguchi [taguchi@vox.tutkie.tut.ac.jp]

2. 並列学習

先述したように対話戦略の獲得では，相手の要求や意図を直接参照することはできない. そのため，学習環境は不完全知覚状態を有する部分観測マルコフ決定過程（POMDP）と定義することができる. 本研究ではエージェントが取得するセンサ情報や，対話の履歴情報を組み合わせることでマルコフ性の回復を試みる. この場合，どの組み合わせを用いて状態空間を設計するかが問題となる. そこで，状態空間の異なる複数の学習器を並列に利用し，環境に適した学習器を自動的に選択することを考える.

2.1 並列学習の構成

並列学習を用いた学習エージェントの構成を図 1 に示す. 本報では学習器の学習アルゴリズムに Q 学習[Watkins 92]を用いた. エージェントは状態空間の異なる複数の学習器と一つの行動選択器を持つ. 各学習器は感覚入力からの情報を基に，それぞれ状態を認識する. 行動選択器は，各学習器の学習結果に基づいて行動を決定する. 行動の結果としての報酬と新たな感覚入力が各学習器に与えられ，それぞれが以下に示す Q 学習の更新式によって Q 値を更新する.

$$Q(x,a) \leftarrow (1-\alpha)\,Q(x,a) + \alpha\,(r + \gamma \max_{b} Q(y,b)) \qquad \cdots(1)$$

ここで，x は遷移前の状態，a は遷移時の行動，r は遷移によって得られた報酬，max Q(y,b)は遷移先の状態での最大 Q 値，α は学習率，γ は時間遅れ報酬に対する割引率となっている. Q 学習は，マルコフ決定過程(MDP)において全ての状態行動対が更新され続ければ，行動の選択方法に依存せずに最適行動を獲得することが保証されている[Sutton 98]. つまり，エージェントが常にランダム行動をとっていても，最適行動は獲得される. そのため，行動選択器がどのような行動を選択したとしても，全ての学習器は Q 値を更新することができる.

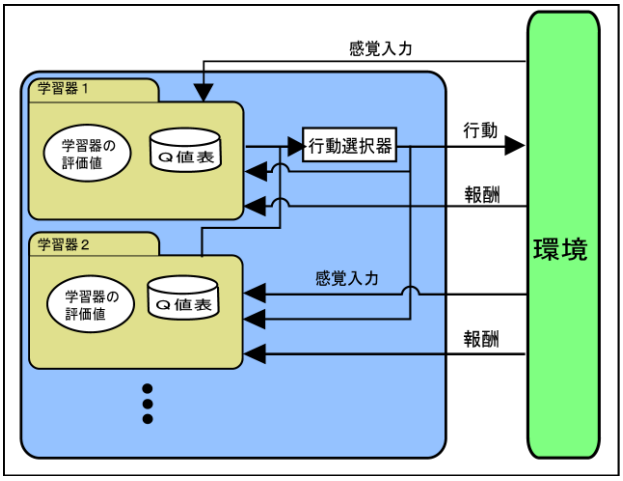

図 1:並列学習の構成

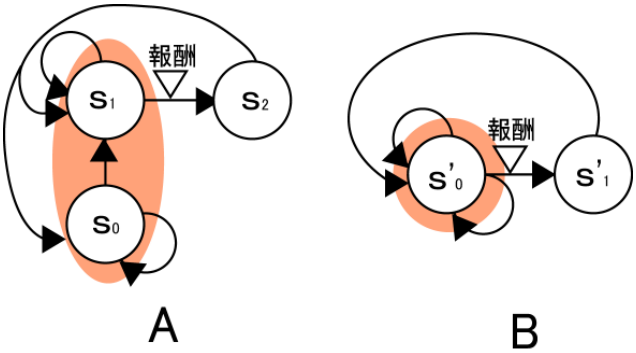

図 2:不完全知覚となる環境

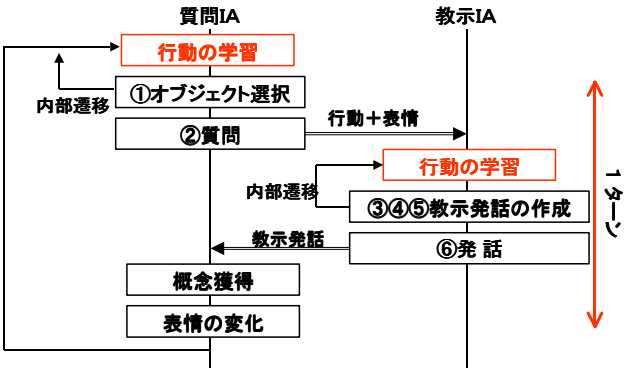

図 3:対話学習の流れ

2.2 行動選択の方法

Q 学習では学習後，ある状態で最も Q 値の高い行動をとることが最適行動となる．この行動を学習器の推薦行動と呼ぶことにする．各学習器から得られる推薦行動が競合した場合，どれを利用するのが最も有効であるかを判断しなければならない．競合の例を図 2 に示す．図 2 の A が学習対象となる環境である．環境を完全に識別できる学習器は環境と同様のモデルを持つ．図 2 の B では不完全知覚によって実際には異なる S_0 と S_1 という状態が一つの状態 S'_0 として認識されている．エージェントが状態 S_0 にいる場合，上に行く行動を取るのが最適な戦略となるが，不完全知覚の学習器は状態 S_0 と S_1 が同一の状態 S'_0 として認識されるため右に行く行動が獲得される．また，状態 S'_0 として学習された Q 値は状態 S_0 と S_1 を混同した結果であるため，S'_0 の Q 値と S_0 や S_1 の Q 値を直接比較し優劣を判定することは出来ない．そのため Q 値を用いない方法で行動の価値を判断しなければならない．そこで本報では，行動選択を行う学習器を試行毎に切り替え，得られた報酬を評価値として利用する．試行に使用した学習器の評価値 V_i の更新式を以下に示す．

$$V_i \leftarrow V_i + \alpha_i \{ \Sigma r / n \} - V_i \} \quad \cdots(2)$$

ここで i は学習器，α_i は学習率，Σr は一試行で得られた報酬の合計，n は一試行の行動数を表す．そしてエージェントは各学習器の評価値を基に確率的に学習器を切り替える．

3. 並列学習を用いた対話戦略の獲得実験

提案手法をエージェント同士の対話学習に適用した実験を行う．実験では，我々が開発している Infant Agent（IA）[新田 02，田口 03]同士が，対話を通して概念を共有化していく過程を対象に，対話学習を効率的に進めるための対話戦略をそれぞれの IA に自動獲得させる．

3.1 Infant Agent による概念獲得

IA は人間や他の IA との対話を通して，オブジェクト特徴と音声特徴の対応関係を概念として獲得する．本実験の対話学習は，人間の教示によって全ての概念を獲得した IA（教示 IA）が，学習途中の IA に概念を教示する 形で進められる．対話学習の流れを図 3 に示す．質問 IA は，仮想空間内にあるオブジェクトから話題とするオブジェクトを選択し，それを指差すか，移動することで教示 IA に質問する．学習に用いるオブジェクトは形（丸，三角，四角），色（赤，青，白），位置（上，下，左，右）および動作（移動）の計11個の特徴を持ち，教示 IA は質問されたオブジェクトが持つ特徴の呼称を1～3語で発話する．質問 IA は教示された音声特徴に未知語が含まれている場合は，その音声特徴を概念辞書に新規登録する．また，既に登録済みの音声特徴が教示された場合は，対応するオブジェクト特徴の頻度を概念辞書に保存する．そして学習を進める中で，ある音声特徴に対応するオブジェクト特徴の一つが0．9以上の頻度で現れた場合，その対応関係を概念として獲得する．先行研究の [新田 02] では，聴覚情報に音声特徴ベクトルを用いたが，今回の実験では対話戦略を獲得する手法の確立に的を絞るため，モーラ単位のシンボル列を入力とし，かつシンボル列に誤りはないものとした．

IA 同士が対話を行う場合，効率良く対話学習を進めるための戦略が重要になる．例えば，正しい単位で未知語を登録させるためには，少ない単語数での教示が有効であり，効率よく概念を確定させるためには，一度に多くの単語を教示した方が有効である．また，質問 IA は教示された内容が理解できたら話題を変える等の理解度に合わせた質問戦略が有効である．但し，人間との対話も考慮に入れると，相手の理解度を直接参照することはできない．そこで質問 IA には理解度を間接的に表示するための表情を与える．通常，IA は平常の状態にあるが，概念を獲得した場合と，教示発話に易しい未知語（4モーラ未満）が含まれている場合に，その数に応じて3段階の快の表情になり，難しい未知語（4モーラ以上）が含まれている場合は不快になる．また，快・不快は持続せず，その都度平常に戻るとする．

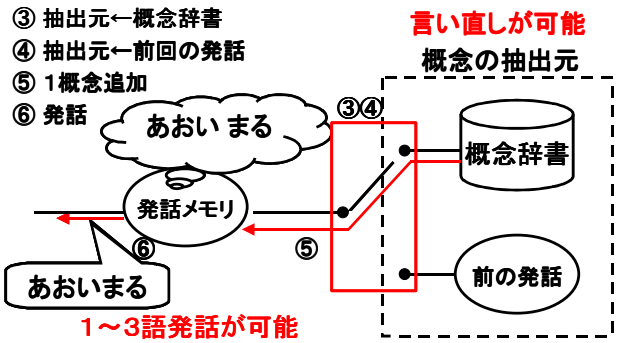

図 4:教示の流れ

3.2 対話戦略の獲得

概念獲得を効率的に進めるための対話戦略を，対話学習を通して両 IA 同時に獲得させる実験を行う．実験では予めいくつかの概念を初期概念として質問 IA に与え，初期概念数の違いに依存しない戦略を獲得させる．学習アルゴリズムには上記の Q 学習を用いた．以下では Q 学習に用いる行動，状態，報酬について説明する．

(1) 行動

質問 IA の行動は①話題とするオブジェクトをランダムに選択する，②指差し／移動によって質問する，の二つとした．次に図 4 を用いて教示 IA の行動を説明する．教示 IA は概念の抽出元から，いくつかの概念を発話メモリに格納(図中⑤)した後，その発話メモリの内容を発話(図中⑥)する．この際，質問された内容に含まれない概念は発話メモリに格納できないとした．本稿では，概念の抽出元として，概念辞書と前回の発話内容の二つを用意した．教示 IA は教示のたびに，二つの抽出元から一方を選択(図中③④)する．概念の抽出元に概念辞書を選択した場合は，質問内容に含まれる全ての概念が発話の候補となる．抽出元に前回の発話内容を選択した場合は，質問内容に含まれる概念のうち前回発話した概念だけが，発話の候補となる．つまり抽出元を選択することで，教示 IA は通常の教示と言い直しによる教示とを適宜切り替えることができる．

(2) 状態

質問 IA の状態空間は，表 1 に示す六つの情報(以下，それぞれを状態属性と呼ぶ)を用いて作られる．現在の表情と前回の表情を持つことで，表情の変化(つまり理解度の変化)を利用した戦略が獲得されることを期待している．教示 IA には，表 2 に示す七つの状態属性を用いた．但し，未教示概念とは，未だ快または平常の表情が得られていない概念，つまり理解されているということが確認できていない概念のことを指す．

表 1:質問 IA の状態属性

分類	状態属性
質問戦略	時間 (5 ターン毎に 1 増加させ，最大値は 5 とする)
	前ターンの話題変更の有無
	現在の表情
	聞き取れた概念数
	話題となるオブジェクトの既知概念数
制御	現在の話題変更の有無

表 2:教示 IA の状態属性

分類	状態属性
時間戦略	時間 (5 ターン毎に 1 増加させ，最大値は 5 とする)
表情戦略	話題の変更があったか(変更あり/ 変更なし)
	現在の質問 IA の表情(平常/ 快/ 不快)
	前回発話した単語数(0 ～ 3)
履歴戦略	発話メモリ内の未教示概念数(0 ～ 3)
制御	概念の抽出元メモリ (未定義/ 概念辞書/ 前の発話)
	発話メモリ内の概念数(0 ～ 3)

(3) 報酬

両 IA に与える報酬には質問 IA の表情を用いる．具体的には快の場合にその段階に応じて 10,20,30 の報酬を与え，平常または不快の場合には時間コストとして-4 の負の報酬を与える．また，①話題の変更や③～⑤の教示発話の作成に関する行動による内部遷移については報酬を 0 とした．

3.3 並列学習を用いた対話戦略の獲得実験

(1) 実験条件

並列学習を対話戦略獲得に適用した場合の有効性を検証するため実験を行う．実験では表 3 で示すように教示 IA の状態空間の一部を用いて作られた 7 つの学習器を並列に用いた場合と，全ての状態属性を用いた学習器(表 3 の学習器 7 と同じ)を単独で用いた場合を比較する．

Q 学習の学習率 α，および．並列学習に用いる α_i は学習回数に応じて 1～0 へと減少させる．割引率 γ は 0.9 と設定した．学習時の行動選択には，ε - greedy 手法($\varepsilon = 0.1$)を用いた．質問 IA が 11 個全ての概念を獲得するか，対話が 100 ターンを超えるまでを 1 試行とし，試行が終了するごとに質問 IA の概念を 0～10 個からランダムに選択し初期化する．このとき戦略の学習結果は保持する．獲得実験後，獲得戦略の効率を評価する実験を行った．

表 3:並列学習に用いた教示 IA の学習器

学習器 1	時間戦略＋制御
学習器 2	表情戦略＋制御
学習器 3	履歴戦略＋制御
学習器 4	時間戦略＋表情戦略＋制御
学習器 5	時間戦略＋履歴戦略＋制御
学習器 6	表情戦略＋履歴戦略＋制御
学習器 7	時間戦略＋表情戦略＋履歴戦略＋制御

(2) 実験結果

学習器を単独で用いた場合と並列で用いた場合，それぞれについて 10 万回の試行を 5 回行った．並列学習を用いた戦略学習時における各学習器の評価値の推移を図 5 に示す．図から試行回数に応じて学習器 3，学習器 5，学習器 7 と順に評価値が高くなっていくことが解る．これは学習器が状態空間の小さなものから大きなものへと段階的に切り替わっていったことを示す．次に，各学習段階において獲得された戦略の評価実験を行った．評価実験では，初期概念 0～10 個それぞれについて 100 回の対話学習を行い，1概念を獲得させるために必要な平均ターン数を調べた．尚ここでは，行動選択および学習器の選

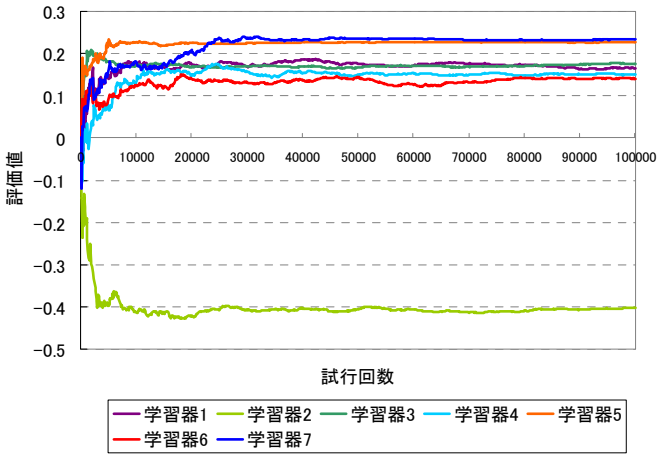

図 5:学習器の評価値

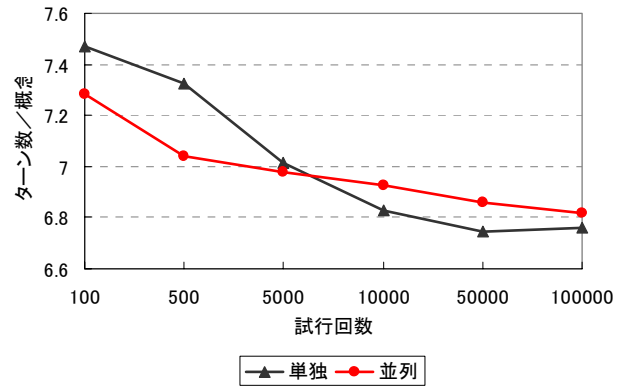

図 6:各学習段階における性能比較

択はもっとも評価値の高いものを選ぶとした. 評価実験の結果を図 6 に示す. 図の横軸が試行回数, 縦軸が平均ターン数である, 図から学習初期においては学習器を単独で用いるよりも, 並列で用いる方が効率よく行動できることが解る. これは状態空間の小さな学習器の学習結果を利用し, 行動選択を行ったためである. しかし, 学習の後半では学習器を単独で用いた方が高い性能が得られるという結果になった. これは他の学習器を利用して行動選択を行うことで学習サンプルに偏りが生じ, 学習器 7 の最適行動の獲得が遅れたためであると考えられる. この問題を解決するためには, 学習器の評価値だけを基準にして切り替えを行うのではなく, 学習進度のような他の指標も同時に利用し, より効率的に学習器を切り替えていく方法が必要である.

3.4 考察

実験の結果から, 並列学習を用いることで, 環境への即応性の高いものから段階的に学習器を切り替えることができることを示した. 人間との対話で戦略を獲得させる場合など, 1 回の試行に多くのコストがかかる学習問題では, 状態属性の組み合わせごと実験を行い, 組み合わせの最適化を図るのは現実的ではない. 本手法は 1 回の試行で複数の学習器を同時に学習させることができるため, 実験時間の短縮が期待できる.

状態空間が異なる学習器の並列利用に関しては, いくつかの先行研究がある. 伊藤らは完全知覚と不完全知覚の学習器を同時に使い, 学習初期は不完全知覚の学習器で行動を選択し, 成功例を集め, 途中で完全知覚の学習器に切り替えることで, 完全知覚での学習を高速化する手法を提案している[伊藤 01]. 各学習器が行う行動の選択は, 行動価値に基づいて確率的に行われる. 学習初期では全ての行動が等確率で選ばれるが, 学習が進むにつれ確定的に行動が選択されるようになる. すなわち学習は行動の不確定性を減らしていく作業と捕らえることができる. この行動の不確定性が一定の閾値以下になった場合, 十分な学習を行ったとみなして学習器の切り替えを行う. 但し, この手法は予め学習器を切り替える順番を定義しておく必要がある. また, 学習を行っても行動の不確定性が減らないような学習器があった場合には, 学習器を切り替えることができない. また, 内部らは重点サンプリングの原理を利用し, 学習アルゴリズムや状態空間の異なる複数の学習器を並列利用する手法を提案している[内部 03]. この手法は各学習器のアルゴリズムに重点サンプリングを利用することで, 他の学習器が行動を決定する場合でも効率的に学習を進めることができる. しかし[Precup 99]で指摘されているように重点サンプリングは値の変動が大きく, 収束し辛いという問題がある. また, POMDP において重点サンプリングが有効に機能するという保証はまだ得られていない.

4. まとめ

状態空間の選定と, 段階的な戦略の学習を実現する方法として, 状態空間の異なる複数の学習器を用いた並列学習を提案した. 実験の結果から, 複雑な状態空間を持つ学習器を単独で用いるよりも即応性が高いことが示された. 実験では教示 IA だけが並列学習を行ったが, 質問 IA と教示 IA の両者が並列学習をした場合の実験については今後の課題となっている. また同時に, 先行研究の知見を利用し効率的に学習器を切り替える手法を検討していきたい.

参考文献

[Precup 99] D. Precup etal : Eligibility Traces for Off-Policy Policy Evaluation, ICML, pp759-766, 1999

[Sutton 98] R.S. Sutton etal : Reinforcement Learning, MIT Press, 1998 (三上ほか 訳:強化学習, 森北出版, 2000)

[Watkins 92] C.J.C.H. Watkins etal : Q-learning, Machine Learning 8 , pp.279-292 , 1992 .

[伊藤 01], 伊藤ほか:知覚情報の粗視化によるマルチエージェント強化学習の高速化 : ハンターゲームを例に, 信学会論文誌, D-1 情報・システム I-情報処理, Vol. J84-D-1 Num. 3, pp.285-293, 2001

[内部 03] 内部ほか :重点サンプリングを用いた複数強化学習器の同時学習, 信学技報 NC2002-233, pp179-184, 2003.

[田口 03] 田口ほか :Infant Agents 間相互対話による対話戦略の自動獲得, 人工知能学会研究会資料 SIG-SLUD-A302-4, 2003.

[新田 02] 新田ほか : Infant Agents 間での対話による概念知識獲得, 人工知能学会全国大会, 2002 1A1-07.