3E1-04

対話学習エージェントの戦略獲得におけるプランニングの適用

Dialogue Strategy Acquisition using Dyna-Q architecture

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 工学研究科
Graduate School of Engineering, Toyohashi University of Technology

This paper describes efficient acquisition of dialog strategies using Dyna-Q. Experiments are carried out through the interaction between two Infant Agents (IAs) to exchange the concepts held by each IA, and dialog strategies for efficient concept sharing are acquired through that. Dyna-Q is one of the Reinforcement Learning algorithms and can acquire strategies with a combination of on-line Learning and off-line Planning, however, because the Planning needs an environment model learned by agent, we propose three types of model-learning methods and compare them. In addition, to test the effectiveness of Dyna-Q, we compare Dyna-Q with Q-Learning which is used in our previous paper. The experimental results showed that (1) the method to learn a stochastic model with a forgetting factor is the most effective, (2) Dyna-Q with the previous model can acquire more efficient strategies than Q-Learning in first stage of learning.

1. はじめに

近年，人間－エージェント対話を通して，概念をエージェントに自動獲得させる研究が行われ始めた[中川 97, 赤穂 97, 金 00, 新田 02, 小玉 04]. それらの研究では，エージェントが得たセンサ情報と，人間の教示音声から，両者の対応関係を概念として学習していくため，実際の環境および直接の対話相手と接地した知識が獲得できる. ところで，対話を通した概念獲得を考えると，人間やエージェントが利用する対話戦略が概念獲得の効率に大きな影響を与えることは容易に想像できる. 例えば，教示者がランダムに概念を教えるよりも，学習者の理解状況に合わせて教示する方が効率的である. また，学習者が自身の理解状況を教示者に如何に伝えるかによっても効率が変化する. しかし，従来の概念獲得研究では，有効な対話戦略を如何にしてエージェントに与えるかという議論は行われてこなかった. 一般に対話戦略は人間が設計して与えているが，これらは対話相手と対話環境によって大きく異なるため，対話戦略もまた概念と同様に，エージェントが自ら獲得し運用するのが望ましい.

こうした背景から我々は，概念獲得を効率的に進めるための対話戦略を強化学習によってエージェントに自動獲得させる研究を行ってきた[田口 04]. また，将来的に人間との対話を通して戦略を獲得させることを考えると，可能な対話時間には限界があるため，時間をかけて最適な戦略が獲得されるような手法よりも，少ない時間でそれなりの戦略を獲得できる手法が望ましい. 前報[田口 04]で提案した手法では，小さな状態空間から学習をはじめて段階的に戦略を獲得することができるため，学習初期でも効率的な対話戦略が獲得される. 本報ではより効率的な戦略獲得を目指し，強化学習と平行して環境モデルを利用したプランニングを行う Dyna-Q[Sutton 98]を本タスクに適用する. Dyna-Q はオンラインで戦略と環境モデルの双方を学習し，学習した環境モデルを利用しオフラインでプランニングを行うことができる. そのため，オンライン学習のみの手法よりも実世界でのインタラクションが少なくてすむという利点がある. 本実験では対話戦略の獲得にこの手法を適用し，対話時間が削減できることを示す.

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

2. 強化学習とプランニング

強化学習とは環境を予めモデル化することなく，エージェントが行動した際に得られる報酬を元に取るべき行動を学習していくアルゴリズムである. エージェントが環境を認識した結果を状態と呼ぶ. 一般に学習後は，各状態における最適な行動が学習されるため，条件反射的な素早い意思決定を可能にする. 一方，プランニングとは，与えられた環境のモデルから戦略を導出，または改善するための手法であり，熟考型の意思決定を実現する. プランニングは大別すると，記号論理をベースとしたプラン空間プランニングと，動的計画法などに代表される状態空間プランニングにわけられる. 後者は強化学習との親和性が高く，強化学習とプランニングを統一的に扱うアーキテクチャが提案されている. 本報ではその一つである Dyna-Q[Sutton 98]を利用して実験を行う.

2.1 Dyna-Q

Dyna-Q のアーキテクチャを図1に示す. Dyna-Q エージェントは，環境とのインタラクションで得た経験（図中：実際の経験）から強化学習と環境モデルの学習を行う. またオフラインでは，学習した環境モデルから得られるシミュレーション上の経験を利用しプランニングを行う. 強化学習とプランニングには Q 学習[Watkins 92]が用いられる. すなわち，Q 学習に入力する経験を切り替えることで，強化学習とプランニングを併用した戦略の獲得ができる.

2.2 環境モデルの学習方法

Dyna-Q では，どのように環境のモデルを学習するかが問題となる. 本報では以下の三つのモデル学習手法を提案し，その比較実験を行う.（以下の(1)～(3)のモデル学習方法を利用した Dyna-Q をそれぞれ DQ-決定，DQ-確率，DQ-忘却と呼ぶ）

(1) 決定論的モデルの学習

各状態行動対において，最近の遷移一つのみを保持する. その遷移で得られた報酬も，直接報酬一つのみを保持する.

(2) 確率論的モデルの学習

各状態行動対における次状態とその頻度を全て保持し，状態遷移確率を計算する. また，各遷移で得られた報酬は平均する.

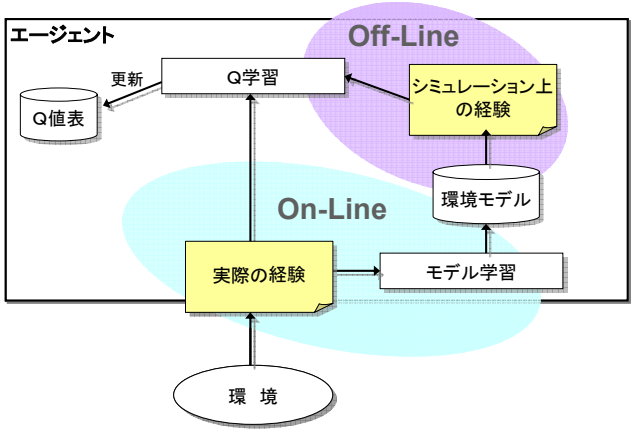

図1：Dyna-Q アーキテクチャ

(3) 忘却因子つき確率論的モデルの学習

(2)と同様の方法で学習するが，各遷移情報および報酬を時間と共に忘却するようにする．

3. Infant Agent 同士の対話学習

本実験に用いるエージェントである Infant Agent（以下 IA）について説明する．IA は人間の幼児の概念獲得過程をモデルとして，人間または他の IA との対話を通して概念と対話戦略を獲得していくエージェントである．本報では，人間の教示によって全ての概念を獲得した IA と，概念を持たない IA の2体が，対話を通して概念を共有化する過程を対象に，戦略の獲得実験を行う．

3.1 概念の獲得

対話実験はコンピュータの仮想空間で行われる．この仮想空間には 9 個のオブジェクトがあり，それぞれのオブジェクトは色や形，位置といった複数の視覚特徴を持つ．オブジェクトの視覚特徴は 10 種類（丸，三角，四角，赤，青，白，上，下，左，右）にカテゴリ化され，それぞれの有無を表す 0/1 のベクトルとして IA に渡される．このカテゴリ化されたそれぞれの特徴を以下ではオブジェクト特徴と呼ぶ．本報では，このオブジェクト特徴とその特徴を指す音声特徴との対応関係を概念と呼ぶ．IA は指差し等で指示されたオブジェクトに関連した獲得済みの概念を1～4 語で発話（教示発話）することができる．教示発話を受け取った IA は，オブジェクト特徴と音声特徴との対応関係から概念を獲得する．具体的な概念の獲得アルゴリズムには[中川 95]の方法を用いた．

このタスクには以下の二つの問題がある．

① 発話された単語がそれぞれどのオブジェクト特徴のことを指しているかについての情報は相手に与えられない．

② 発話は単語毎に区切られておらず，連続した音声として与えられるため，連続して 2 語以上の未知語を教示すると，1 つの未知語として受け取られてしまう．

①の問題から，例えば「あか」という概念を獲得させるために教示者は，赤い丸や赤い四角などの複数の赤いオブジェクトに対して「あか」と教示し，それが色の概念であることを確率的に学習させる必要がある．複数単語による教示はそれを効率的に行うために有効である．しかし，②の問題があるため単純に「できるだけ多く発話する」だけでは効率的な対話は実現しない．本実験ではこうした教示のための戦略に加え，「いつ聞き返すべきか」や「どのオブジェクトを話題にするか」といった学習者側の戦略も同時に獲得させる．尚，本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，音声特徴には音素単位のシンボル列を用いた．また，未知語を正確に聞き取ることは，既知の単語を聞き取るよりも困難であると仮定し，未知語を受け取った IA は 1 音素当たり 0.1 の確率で認識誤り（ランダムに音素を変化）を起こすとした．

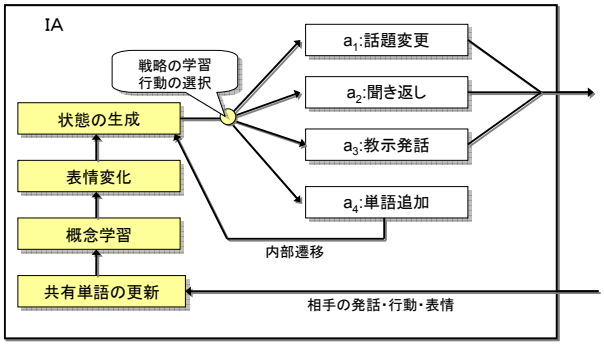

図2：IA の行動

3.2 状態・行動・報酬

前報では，教示と質問に役割を固定した 2 体の IA を用いて実験を行った．本実験では同じ IA を 2 体用いて，それぞれの初期概念数の違いだけから役割に応じた戦略を獲得させる．

(1) 行動

IA は以下の4つの行動を持つ．

- 話題変更：ランダムにオブジェクトを選択し指差す．
- 聞き返し：相手の教示発話を繰り返す．
- 単語追加：話題となるオブジェクトに関連する単語を一つ発話レジスタに追加する．
- 教示発話：発話レジスタの内容を発話（1～4 語）する．

対話はどちらかの IA が「話題変更」することによって始まる．その後は，学習中の戦略に従って上記の 4 つの行動のどれかを選択し実行していく．「話題変更」，「聞き返し」，「教示発話」を実行した場合は相手に行動の権利が移るが，「単語追加」は話題に関連する単語を全て追加するまで（最大 4 語）繰り返し実行することができる．また，IA は「聞き返し」や「教示発話」があった場合，その中で正確に発話された単語を「両者で共有された単語」（共有単語）として共有単語メモリに保持する．IA の行動の流れを図 2 に示す．

(2) 表情変化

IA は快，不快，平常の感情（生得的な内部状態）を持ち，以下の規則に従って変化する．これらの感情は表情モダリティを介して他の IA に伝えられ，強化学習および環境モデルの状態と報酬に利用される．

- 快　：新たな概念を獲得した場合
  共有単語が増加した場合
- 平常：快でも不快でもない場合
- 不快：相手の発話に未知語が含まれている場合

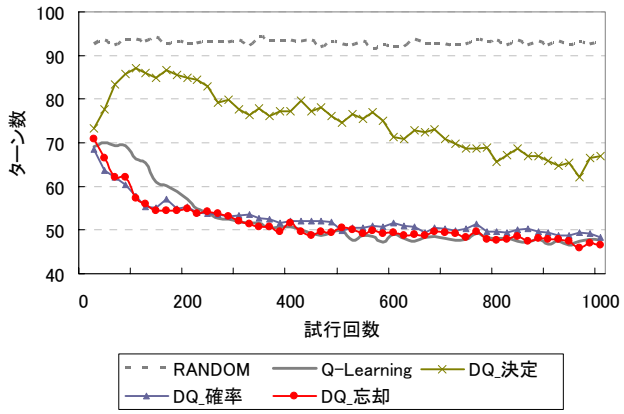

図3：実験結果（試行 1,000 回）

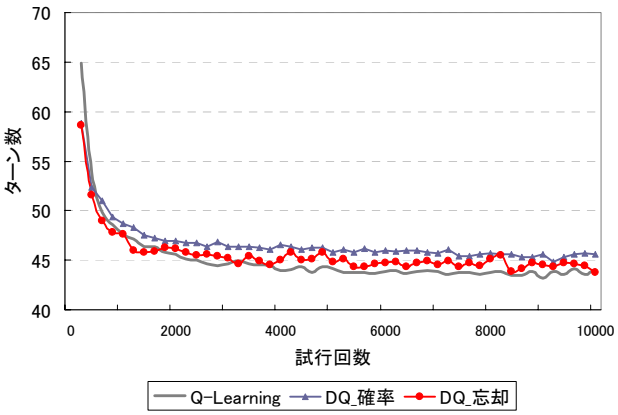

図4：実験結果（試行 10,000 回）

(3) 状態

IA は，相手の行動，それによって変化した自身の表情，今何を発話しようとしているのか，といった情報を用いて状態を生成する．具体的には，以下の 9 次元で状態を表現する．

・相手の表情（快，平常，不快）
・相手の行動（話題変更，聞き返し，教示発話）
・相手の発話単語数（0～4）
・相手の発話に含まれる未共有単語数（0～2）
・自分の表情（快，平常，不快）
・自分の獲得概念数（0～10）
・話題となるオブジェクトに関する既知の概念数（0～4）
・発話レジスタ内の単語数（0～4）
・発話レジスタ内の未共有単語数（0～4）

(4) 報酬

協調的な対話戦略を獲得させるために，報酬は両 IA とも共通とした．具体的には，IA の感情が快になった場合に + 10，平常／不快の場合に - 1，単語追加時の内部遷移の場合に 0 とし，その報酬と相手の表情から算出される報酬とを足したものを報酬として利用した．

4. 獲得実験

4.1 実験条件

Dyna-Q 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，(1)環境モデルの学習方法が異なる3種類の Dyna-Q の比較実験，および(2)Q 学習と Dyna-Q の比較実験を行う． 両 IA が 9 個以上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，試行が終了するたびに両 IA の概念を初期化（それぞれの初期概念数は 10 個と 0 個）する．尚，このとき戦略の学習結果は保持する．Dyna-Q によるオフラインでのプランニングは 10 試行毎に 1 万ステップ（行動）行う．また，それぞれの学習率 α 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．行動選択には，ε - greedy 手法（ε = 0.1）を用いた．

4.2 実験結果

実験結果（20 回の平均）を図 3, 4 に示す．図の横軸は試行回数，縦軸は各試行における平均終了ターン数（9 個の概念を獲得するまでの時間）となっている．比較のためにランダムで行動した場合の結果も載せる．なお，プランニングに要した時間（ステップ数）は載せていない．

図 3 から学習初期において，確率論的なモデルを用いた Dyna-Q が， Q 学習よりも効率的な対話戦略を獲得できることが解る．しかし，図 4 の結果をみると，忘却因子なしの確率論的モデルを利用すると（DQ-確率），最適な戦略が獲得できないことがわかる．これは，両 IA が同時に戦略獲得を行うことで学習初期と後半では環境のモデルが変化し，それによって生じる環境モデルと実際の環境とのずれが原因となっている．一方，忘却因子つきのモデル（DQ-忘却）は，そのずれを解消していくことができるため，学習後半においても Q 学習と同等の戦略を獲得することができる．

5. まとめ

Dyna-Q を利用して対話戦略の獲得実験を行った．実験の結果から，忘却因子つきの環境モデルを利用することで， Q 学習を用いた場合よりも早く，効率的な対話戦略が獲得できることが示された．また，100 対話程度でもそれなりの戦略が獲得できることは，人間との対話を通した戦略獲得にも応用可能であることを示唆している．今後は，より正確な環境モデルの学習方法を検討すると共に，人間との対話を対象とした実験を行っていきたい．

参考文献

[赤穂 97] 赤穂，速水，長谷川，吉村，麻生: EM 法を用いた複数情報源からの概念獲得，信学会論文誌，Vol.J80-A pp.1546-1553 , 1997.

[金 00] 金，岩橋: 知覚情報の統合に基づく言語音声単位の獲得アルゴリズム，信学技報，TL200-21, pp.9-16, 2000.

[小玉 04] 小玉，田口，桂田，岡部，新田: オンライン学習による Infant Agent のための効率的概念獲得，人工知能学会全国大会，2004，3F3-03.

[中川 95] 中川，升方: 視聴覚情報の統合化に基づく概念と文法の獲得システム，人工知能学会，Vol.10，No.4，pp.619-627，1995.

[新田 02] 新田，越坂，桂田: 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

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

[田口 04] 田口，桂田，新田: 並列学習を利用した対話戦略の獲得，人工知能学会全国大会，2004 1A2-02.

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