1D2-7

Scaffolding(足場づくり)を利用した学習系の構築

Building a learning system that utilizes scaffolds

田中 一晶　　岡 夏樹
TANAKA Kazuaki　OKA Natsuki

京都工芸繊維大学 大学院工芸科学研究科
Graduate School of Science and Technology, Kyoto Institute of Technology

In the future, robots are expected to support human's work. However, it is not realistic to program all the action into them. It therefore is imperative that they have learning ability to learn new actions through interaction with humans and environments.

It is known that Scaffolding is effective in studying through interpersonal interactions. We can hence predict that it is also effective in studying through human-robot interactions. However, it has not become clear: 1) whether the scaffolding actually occurs in everyday human-robot interactions, 2) under what situations it occurs if it could occur, and 3) how a robot can utilize the scaffolds that are given by the ordinary people.

In this work, we set up a situation that a human teaches a robot new actions, and try to clarify the three points by experiment.

1. はじめに

近年、人と触れ合うロボットが日常生活の場に現れ始めた。しかし、それらのロボットは予め実装された行動を実行するものや、人が操作して動くものがほとんどであるため、社会性が低く、すぐに飽きられるという問題がある [中田 00]。また、将来、人の日常生活の場で、ロボットが人の仕事をサポートすることが予想されるが、そのようなロボットに予め必要な行動を全てプログラムしておくことは現実的ではない。よって、人と触れ合うロボットは人や環境とのインタラクションを通して新たな行動を獲得する能力が必要不可欠である。

人が人とのインタラクションを通じて学習する場合には、Scaffolding が有効であることが知られている [Wood 76]。Scaffolding はその時々の学習者の能力に応じて簡単な学習課題から徐々に難しい学習課題を与えて行く方法であり、ロボットも人との触れ合いを通して学習する場合には、Scaffolding が有効である可能性がある。ロボットに段階的に課題を与えることよって学習の効率化を図った研究としては、[Asada 96, Thomaz 06] があり、簡単な課題から徐々に難しい課題を与えることによって、効率的に学習できることが明らかとなったが、これらは、研究者が工夫して段階的な課題を設定し、評価実験を実施したものであった。したがって、日常的な場面で一般の人との触れ合いを通してロボットが学習する場合、以下の点は未だ明らかになっていない。

- 実際に、Scaffolding は生じるのか。
- Scaffolding はどのような条件下で生じるのか。
- 日常的な場面で、一般の人によって与えられた足場をロボットは利用できるのか。

本研究では、人がロボットに新たな行動を教える場面を具体的に設定し、これらの問題を実験的に明らかにする。

連絡先: 田中一晶, 岡夏樹, 京都工芸繊維大学 大学院工芸科学研究科, e-mail: d8821007 @ edu.kit.ac.jp, oka @ dj.kit.ac.jp

2. 人間－ロボットインタラクションの予備実験

ここでは、ロボットが獲得の行動獲得実験のタスクとして何が適切であるかを検討する。そして、検討した学習タスクを通して、人とロボットのインタラクションの観察を予備実験として行い、その結果を示す。

2.1 学習タスクの検討

本研究ではロボットが獲得する新たな行動として、ゲームの獲得を選択する。我々は、一般の人に馴染み深く、ペットロボット AIBO ERS-7 でも実行可能なゲームを検討した結果、「あっち向いてホイ」が良いと考えた。

あっち向いてホイは一般的には、「あっち向いてホイ」という掛け声を用いて行うが、容易に使用できる通常の音声認識システムによる音声認識にはタイムラグがあり、リアルタイム性が失われてしまうため、音声認識の代わりに距離センサを用い、物体が動いたタイミングを利用する。また、AIBO の画像認識では、人の手の認識よりもピンク色の認識精度が高いため、ピンク色のボールを指の代わりに使用する。

本研究では、ゲーム獲得には 2 つの過程が必要であると考え、1 つ目はゲームの手順を学習する「手順獲得過程」、2 つ目はどのような状態が目標であるかを学習する「目標状態獲得過程」と設定した。各過程における学習内容は以下とする。

手順獲得過程: 人がボールを AIBO の顔の前に持ってくるとゲームの開始であり、ボールが動いた瞬間、タイミング良く顔を 4 方向 (上下左右) どこかに向けるというゲームの進行手順を学習する。

目標状態獲得過程: 4 方向どこかを向いた結果、目の前にボールが無ければ勝ち (人がボールを動かした方向以外を向いた)、目の前にボールがあれば負け (人がボールを動かした方向を向いてしまった) であることを学習する。

2.2 実験方法

AIBO には予め、前述の手順獲得過程 (2.1 節) で述べたゲームの手順を学習させておき、目標状態 (どのような状態が勝ち (または負け) か) を人とのインタラクションを通じて獲得する実験を行った。

実験協力者 7 名には、AIBO の表情 (ボールの有無、嬉しい、悲しい、混乱、通常の 6 種類) と、AIBO が勝ったときには「嬉しい」表出を、負けたときには「悲しい」表出を行うことが目標状態であると伝えておいた。使用できる教示は「撫でる」と「叩く」の 2 種類とした。

2.3 実験結果

予備実験の結果、7 名の実験協力者全員が AIBO に勝利条件を教えることができた。実験協力者の教示は、はじめ、AIBO が勝ったときには「撫でる」、負けたときには「叩く」という単純なものだったが、実験が進むと以下の 1. や 2. のような変化が観察できた。また、3. のような問題も起こった。

1. AIBO が勝利状態・敗北状態を学習し、正しい表出を行うようになると、実験協力者は評価教示を与えなくなっていく (7 名)。
2. AIBO がゲームに敗北し、初めて「悲しい」表出を行ったとき、「その表出は正しい」という意味で「撫でる」教示を与えた実験協力者がいた (2 名)。また、同様の状況において、思わず撫でようとして手を引っ込めたり (1 名)、撫でようと思ったがやめた (3 名) 実験協力者もいた。
3. ボールを動かした方向を AIBO が向いた際、AIBO がボールを認識できなかったとき、「ボールが無い」ときの表情を表出していても、実験協力者はそれに気付くことなく「叩く」の教示を与えていた (5 名)。

従来の学習システムは、単一の学習課題おいて最適解を学習するものであるため、同一状態には一貫して同じ教示 (報酬) を与えることが望ましい。しかし、人とのインタラクションでは、ロボットにとって同一の状態でも、上記の 1. や 2. の変化が起こるため、その変化に追随してしまい、学習がうまくいかない。本研究では、このような教示の変化を足場の一種と捉え、対応する手段を次節で提案する。

3. Scaffolding を利用した学習系

予備実験の結果 (2.3 節) から、学習が進むにつれて、人がロボットに与える教示は変化することがわかった。本節では、人が与える教示の変化 (これを足場と捉えて) に対応し、新たな行動を獲得する学習システムを提案する。

本研究で提案する学習システムは手順獲得部 (3.1 節)、目標状態獲得部 (3.2 節)、感情表出部 (3.3 節) の 3 つの要素で構成する。感情表出部では、人から適切な足場が与えられるように、現在の学習状態をフィードバックする。システムの構成を図 1 に示す。

3.1 手順獲得部

手順獲得の学習アルゴリズムは Q-Learning[Watkins 92] を採用する。Q-learning では、状態 s における行動 a の価値 (行動価値と呼ぶ) $Q(s,a)$ を行動を行うたびに報酬 r に基づいて更新して最適行動を学習する。本研究では入力は以下の 7 つとし、入力に従って状態が定義される。各状態 $s0$〜$s8$ は図 2-(a) を参照。

i_0： 目前の物体の有無 (距離センサによる)。目の前に物体があれば 1、無ければ 0 とする。

i_1： i_0 の前状態からの変化。変化があれば 1、無ければ 0 とする。同一状態に遷移する場合は値を変更しない。

i_2〜i_5： 人がボールを動かした方向 (AIBO Remote Framework の画像認識による)。それぞれ、上 (0001)、下 (0010)、左 (0100)、右 (1000) とコード化する。状態遷移時に初期化 (0000) する。

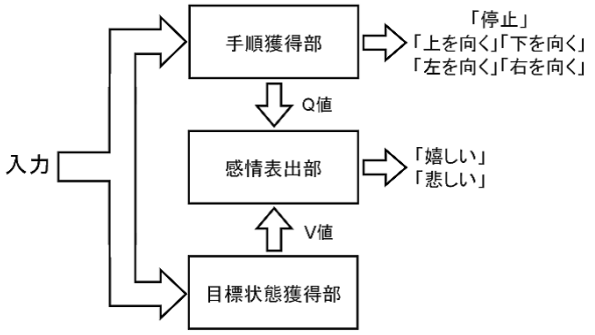

図 1: 学習システムの構成

学習システムは手順獲得部、目標状態獲得部、感情表出部の 3 つの要素で構成する。手順獲得部は Q-Learning によって行動価値を学習し、目標状態獲得部は即時報酬に基づいて状態価値を学習する。また、感情表出部はその時々の行動価値と状態価値に従って感情を決定し、表出を行う。システムへの入力は 3.1 節を参照。

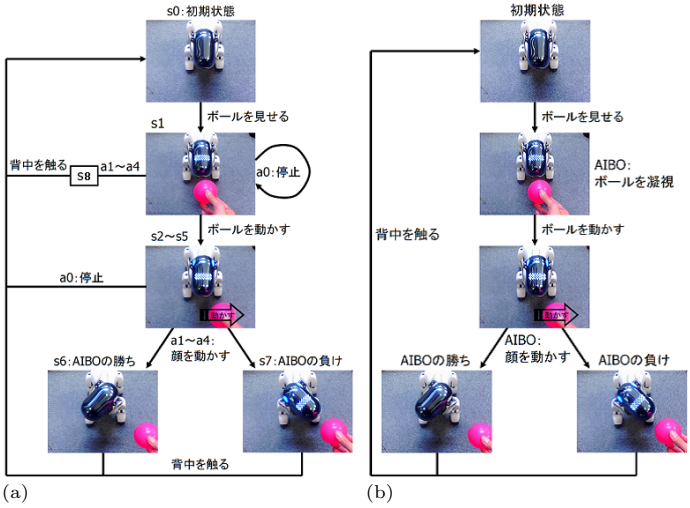

図 2: 学習における各状態とその遷移条件

(a)：各状態とその遷移条件
$s8$ は、状態 $s1$ でボールを凝視せず、誤ったタイミングで顔を動かした状態である。
(b)：評価実験 (4. 節) で実験協力者に提示した手順
図中の矢印に付けた遷移条件のうち、AIBO によるものは「AIBO：(AIBO の行動)」とし、それ以外は人によるものである。

i_6： AIBO の首の位置。初期位置であれば 0、それ以外なら 1 とする。

また、AIBO が実行する行動 a_n は以下の 5 種類とする。

a_0： 停止し、現在の姿勢を維持する。

a_1〜a_4： 顔をそれぞれ上下左右に向ける。

状態遷移が起こると、AIBO は遷移後の状態 s' から行動 a_n を選択し実行する。また、AIBO の背中のタッチセンサに触れると初期状態 s_0 へと戻る。タスクの流れを図 2 に示す。

人から報酬 r(評価教示) が与えられると、以下の更新式に従って実行した行動 a_n の Q 値を更新する。後述の感情表出部 (3.3 節) では、Q 値と勝利条件獲得部 (3.2 節) の状態価値とを比較して感情表出を行う。本研究では、目標状態獲得部は遅れのある報酬は考慮しないため、手順獲得部も同様に遅れのある報酬は考慮せず、Q 値の更新式は以下を用いる。学習率 $\alpha = 0.15$ とした。

$$Q(s,a) \leftarrow Q(s,a_n) + \alpha\{r - Q(s,a_n)\} \quad (1)$$

AIBO の行動選択の方法は Boltzmann 選択を用い、Boltzmann 温度 $t = 0.02$ とした。

また、学習が進んでいない状態では、全ての行動の選択確率は同一であるため、ランダムに選択される。そこで、「ボールの動きにつられて、ボールが動いた方向を向く」を表現するため、状態 $s2$〜$s5$ において、ボールが動いた方向を向く各行動 $a1$〜$a4$ の Q 値にバイアス +0.05 を足して行動の選択確率を計算する。例えば、状態 $s2$(ボールが上に動いたことを認識した状態) においては、行動 $a1$(顔を上に向ける) の Q 値に +0.05 して行動の選択確率を計算する (実際に Q 値を更新するわけではない)。

3.2 目標状態獲得部

目標状態獲得部は即時報酬だけを考慮して (2) 式により、状態価値 $V(s')$ を更新する。s' は報酬が得られたときの状態であり、学習率 α は 0.2 とする。

$$V(s') \leftarrow V(s') + \alpha\{r - V(s')\} \tag{2}$$

予備実験 (2. 節) では、AIBO が勝ったときに喜び、負けたときに悲しむことを学習させたため、教示の対象は「AIBO の感情表出」であった。これに対し、目標状態獲得部では、教示の対象は教示者が定めた「目標状態」とし、報酬に従って状態価値を学習する。

3.3 感情表出部

感情表出部では、教示者に AIBO の学習状態をフィードバックするため、Q 値に従って行動を選択した結果、その行動の Q 値よりも高い状態価値を持つ状態に遷移すれば「嬉しい」、低い状態価値を持つ状態に遷移すれば「悲しい」表出を行う。また、AIBO が正しい行動を選択するようになると、報酬は与えられなくなる (2.3 節の 1. より)。これに応じて、AIBO も感情表出を行わなくなることで、AIBO の飽き・慣れを教示者にフィードバックする。

まず、行動価値 $Q(s,a)$ と次状態 s' の状態価値 $V(s')$ の差分 d をとり、その値に応じて感情を決定し、感情に対応した表情を表出する。また、学習が進むにつれて $Q(s,a)$ は $V(s')$ に漸近するため、差分 d は 0 に近づき、AIBO は感情表出を行わなくなる。

感情 e とそれに対応する表情は以下の 3 種類とする (「感情」→表情)。また、AIBO の感情とは関係なく、ボールが目前にあるときには「ボールを見ている」ときの表情となる。

e_0：「通常」→表出なし
e_1：「嬉しい」→喜ぶ
e_2：「悲しい」→悲しむ

感情 e の決定は以下によって行う。ここで、θ は閾値であり、0.04 とする。

$$d = V(s') - Q(s,a) \tag{3}$$

$$e = \begin{cases} e_0 & (-\theta < d < \theta) \\ e_1 & (d \geq \theta) \\ e_2 & (d \leq -\theta) \end{cases} \tag{4}$$

3.4 NNC の利用

2.3 節で述べた通り、AIBO が正しい行動を選択できるようになるにつれて、人から得られる評価教示は減少する可能性がある。そこで、手順獲得部と目標状態獲得部は、教示が与えられないことを肯定的な評価と捉える暗黙的な評価規準 NNC[左 07] を利用して、人からの報酬が途絶えた後も一定のモチベーションを維持する。

目標状態獲得部では、人からの報酬が無かった場合、NNC による報酬 R はその状態の状態価値 $V(s')$ と同じ値とする。また、手順獲得部に与える報酬は、勝利条件獲得部と同様に $V(s')$ と同じ値とする。つまり、報酬が途絶えた後、状態価値は現在の値を維持し、行動価値は状態 s' で過去に得られた報酬の期待値 $V(s')$ に従って更新される。

$$R = V(s') \tag{5}$$

4. 評価実験

我々は 3. 節で紹介した学習システムが実際に人の教示の変化に対応可能か実験的に評価した。ここではその実験方法と結果について述べる。

4.1 実験方法

AIBO にあっち向いてホイを実験協力者に教えてもらい、その様子を撮影する。実験の時間は 30 分とした。

本実験では、5 名の実験協力者に A4 のプリント 2 枚で AIBO の表情とゲームの手順 (図 2-(b)) を予め提示した。使用できる教示は予備実験と同様、「撫でる」と「叩く」の 2 種類とした。

予備実験の結果 (2.3 節の 3.) より、短い時間の実験では AIBO の表情が多いと、表情と状態との対応がわかりにくく、実験協力者の負担となることが分かったため、本実験では、初期状態には何も表出せず、AIBO の表情は次の 4 つとした (嬉しい、悲しい、ボール有り、よくわからない[*1])。

4.2 実験結果

5 名の実験協力者に実験 (4.1 節) を行ってもらった結果、全員が AIBO にあっち向いてホイを教えることができた。実験中に観察された教示の変化は 3 つに分類することができ、我々はそれぞれ IR(Increasing Rewards)、DR(Decreasing Rewards)、FR(Flipping Rewards) と呼ぶ。表 1 にそれぞれがどのような変化であるかと、それぞれの例を示す。

表 1: 評価実験で観察された教示の変化の例と分類

分類	教示の変化	・例
IR(Increasing Rewards): 報酬が与えられるようになる変化	なし→撫でる	・AIBO にボールを見せた際、ボールを凝視したとき。 ・AIBO がゲームに勝利したとき。
	なし→叩く	・AIBO にボールを見せたが、ボールを凝視しなかったとき。 ・AIBO がゲームに敗北したとき。
DR(Decreasing Rewards): 報酬が与えられなくなる変化	撫でる→なし	・AIBO にボールを見せた際、ボールを凝視するようになったとき。 ・AIBO がゲームに勝利し、「嬉しい」表出を行うようになったとき。
	叩く→なし	・AIBO にボールを見せたが、ボールを凝視しなかったとき。 ・AIBO がゲームに敗北したとき。
FR(Flipping Rewards): 逆の報酬が与えられる変化	叩く→撫でる	・AIBO がゲームに敗北し、「悲しい」表出を行うようになったとき。

この内、我々が提案したシステムでは IR と DR には追随・対応することができたが、FR には対応することができなかった。

5. 考察

5.1 人から与えられる教示の変化への追随・対応

前述の通り、人の教示の変化は 3 つに分類することができる。IR(Increasing Rewards) は報酬が与えられていなかった

*1 「よくわからない」の表情は随時表出システム [田中 06] による人の教示の識別でのみ使用し、「撫でられた」のか「叩かれた」のか識別できないときに表出する。

状態で、報酬が与えられるようになる変化であり、学習者に新しい足場 (課題) が与えられたときに起こる。これは追随すべき変化であり、我々が提案したシステムでは Q 値の重み付け平均により、新しく与えられた教示を重視することによって追随することができた。これに対し、DR(Decreasing Rewards) は報酬が与えられていた状態で、報酬が与えられなくなる変化であり、学習者が与えられた課題を達成したときに起こる。これは追随すべきでない変化であるが、NNC によって対応することができた。

また、逆の報酬が与えられる変化である FR(Flipping Rewards) も観察された。これは、予備実験と同様に、AIBO が負けた際、「悲しい」表出を行ったことに対して「その表出は正しい」という意図で「撫でる」教示が与えられたものであった。予備実験では勝利 (敗北) に応じた感情表出を行うことを目的としていたため、教示の対象は AIBO の感情表出であったが、評価実験ではそのようなインストラクションは行っていなかったにも関わらず、同様の変化が観察された。このように、人とロボットとのインタラクションの際には、学習者のその時々の状態に応じて、教示の対象は変化する (学習者の行動から感情表出へなど)。また、この FR が起こるタイミングは、学習者が与えられた課題を達成したときであるため、DR の代わりに起こっていることがわかる。

このように、学習者が新たな行動を獲得する際には IR と DR(または FR) が交互に起こることによって学習が進んでいく。しかし、DR と FR のどちらが起こるかは教示者によって異なり、予想することが難しいため、FR に対応するためには、学習者は教示の対象が何であるか理解する必要がある。また、DR や FR が起こるのは、学習者の表出 (学習状態のフィードバック) を肯定する意図があるからであり、行動価値の学習には NNC によって、状態価値の学習にはそのときの学習者の表出に応じた報酬を与えることで対応できると考えられる。つまり、学習が進行し、学習者が表出を行うようになった際には、教示者が与える教示の種類に関わらず、「嬉しい」表出を行ったときには正の報酬、「悲しい」表出を行ったときには負の報酬によって状態価値を更新する。例えば、AIBO がゲームに敗北し、「悲しい」表出を行うようになったときには、その表出を肯定する意図で教示が与えられるため、状態価値は負の報酬によって学習する。

5.2 ロボットの表出方法の妥当性

次に、表出方法の妥当性について検討する。AIBO が教示によって「学習していること」は、DR や FR といった教示の変化からわかるように、教示者にフィードバックできていた。また、DR が起こった後、AIBO が表出を行わなくなると再び教示を与えるようになった (IR) ことで、学習は促進されたが、実験協力者の意図は「まだ学習できていないと思ったから」であり、我々の意図 (飽き・慣れのフィードバック) とは異なっている。これは、AIBO が表出を行わなくなった後も探索行動 (最大の Q 値を持つ行動以外を実行) を行っていたことが原因の一つと考えられる。本研究では、Boltzmann 選択を行動の決定方法として採用したが、Boltzmann 温度は 0.02 に固定して実験を行ったために学習が進んだ後にも探索行動を行っていた。一般的には学習が進むにつれて温度を 0 へ近づける方法が知られているが、Scaffolding では、新しい足場 (課題) が次々と与えられるため、この方法は適切とは言えない。そこで、足場が与えられた際には適切な温度を再設定する方法が必要があると考えられる。また、学習の途中段階でも、AIBO が自信を持って行動を選択した場合 (最大の Q 値を持つ行動を選択) と、探索行動を行った場合とで、AIBO の表出を区別すべきである。つまり、どのような意図で行動を実行したのかをフィードバックできれば、より適切に足場が与えられると考えられる。

6. まとめ

我々は、ロボットが日常生活で活躍するためには人や環境とのインタラクションを通して新たな行動を獲得することが必要不可欠であると考え、Scaffolding に注目した。しかし、人とロボットとのインタラクションにおいて、実際に Scaffolding が生じるのか、それはどのような状況下で生じるのか、与えられた足場をロボットは利用できるのか、という 3 つの重要な問題は明らかになっていない。

そこで、これらの問題を明らかにするため、我々は手順獲得部、目標状態獲得部、感情表出部の 3 つから成るシステムを構築し、人がロボットにゲームを教えるインタラクション実験を行った。その結果、人はロボットとのインタラクションにおいても、足場を与えることがわかった。人はロボットの学習状態に応じて、教示の与え方を様々に変える。それらの変化は IR、DR、FR の 3 種類に分類することができ、新しい足場 (課題) が与えられた際には IR、課題を達成した際には DR または FR が起こり、これらを交互に繰り返すことで学習が進行することが明らかとなった。このことから、ロボットは「学習途中であること」、「学習が完了したこと」を人にフィードバックする必要がある。

また、我々のシステムでは、IR と DR には追随・対応して学習できることが明らかとなったが、FR には対応することができなかった。ロボットが人から与えられた足場を有効に利用して学習するためには、FR にも対応できなければならない。そこで、今後の展開として、以下に取り組む予定である。

- FR への対応: 教示の対象の識別方法の検討
- 表出方法・表情の再検討
- Boltzman 温度を自動設定する方法の検討

参考文献

[中田 00] 中田亨, ペット動物の対人心理作用能力のロボットにおける構築, 東京大学大学院工学系研究科, 学位論文, 2000.

[田中 06] 田中一晶, 岡夏樹, ペットロボットによる感情表出のタイミングがユーザとのインタラクションに与える影響, HAI シンポジウム 2006, 1B-1, 2006.

[Wood 76] Wood, D., Bruner, J. S., and Ross, G., “The role of tutoring in problem-solving”, Journal of Child Psychology and Psychiatry, Vol. 17, pp. 89-100, 1976.

[Asada 96] Asada, M., Noda, S., Tawaratsumida, S., and Hosoda, K., “Purposive Behavior Acquisition for a Real Robot by Vision-Based Reinforcement Learning.” Machine Learning, Vol. 23, pp. 279-303, 1996.

[Thomaz 06] Thomaz, A. L., and Breazeal, C., “Tutelage and Socially Guided Robot Learning”, IEEE/RSJ International Conference, Vol. 4, pp. 3475-3480, 2006.

[Watkins 92] Watkins, C. J. C. H., Dayan, P., Q-learning, Machine Learning, Vol. 8, No. 3-4, pp. 279-292, 1992.

[左 07] 左祥, 田中一晶, 嵯峨野泰明, 岡夏樹, “No news is good news” 規準を利用した行動教示の学習, 情報科学技術レターズ, pp. 319-322, 2007.