LJ-009

“No news is good news”規準を利用した行動教示の学習
Action Command Learning Based on “No News Is Good News” Criterion

左 祥 † ZUO Xiang　田中 一晶 † TANAKA Kazuaki　嵯峨野 泰明 † SAGANO Yasuaki　岡 夏樹 † OKA Natsuki

1. はじめに

将来，日常生活の場にロボットが普及してくると予想されている．我々は人とロボットとのインタラクションにおいて，ロボットは適応能力を持っているべきであると考え，また，望ましい行動を人が一々指定するのでなく，報酬に基づく学習もできることが望ましいと考える．しかし，人がロボットの一挙手一投足の全てを明示的に評価してやることは現実的でないため，人の自然な振る舞いの中に含まれる，必ずしも意識的に発したのではない情報を有効に利用して学習することが重要となる．

言葉の意味をロボットやエージェントに獲得させる研究の中で，近年の代表的なものとしては，言葉と状況の共起性を基本とし，それにいくつかの情報を加味して，言葉と状況の対応付けを試みるもの [2, 4, 5, 6, 7, 8] が挙げられる．これらのうち，[2, 4, 6] では，報酬も利用した学習が行われるが，[2, 6] では明示的な報酬だけが利用される．[4] では，人の自然な振る舞いの中に含まれる情報の 1 つとして，音声のピッチの急激な上昇に注目し，これを警告 (負の報酬の一種) と捕えて利用し，より効率的な学習の実現を図った．これに対して，本研究では意識することなく発せられる情報の 1 つとして，発話のタイミング (さらに特定して言うと，発話の欠如) に注目し，この発話のタイミングの特性を利用した効率的な意味学習アルゴリズムを提案する．

2. 人間－ロボット間インタラクションデータの収集と分析

ここでは，予備実験として行った，人とロボットとのインタラクションにおける人の発話データの収集・分析について説明する．

2.1 実験設定

実験タスクとして，一定空間において教示者 (実験協力者) が AIBO ERS-7(SONY の 4 足歩行ロボット) をゴールまで誘導するタスクを設定した．直進だけではたどり着けないように簡単な障害物を置いた．実験では Wizard of Oz 法 [1] を用い，実験者は実験協力者の見えない場所から無線で AIBO を操作した．AIBO の操作はランダムに行動を行う場合と，ある程度学習が進んでいるように見せた状態との 2 回に分けて行った．

次の説明文を教示者に提示してから実験を行った.「アイボを骨のところまで案内してあげて下さい．アイボはまだ言葉がよくわからないので，間違った行動をしてしまいますが，気長に案内してあげてください.」実験協力者は 13 名であった．AIBO の行動の種類は，＜前進＞＜後退＞＜ 90 度右折＞＜ 90 度左折＞＜停止＞の 5 種類であった．ここで＜前進＞＜後退＞＜右折＞＜左折＞の行動は＜停止＞により停止する．

2.2 実験結果

各教示者の実験の様子を撮影したビデオ映像を手作業で分析し，次のことが分かった．

- 人の発話は主に，AIBO がとるべき行動を指示する教示 (行動教示)(全発話中の 68%) と，実行した行動の適否を評価する教示 (評価教示)(全発話中の 26%) の 2 種類であった．残りの 6%は,「うー」のような意味を特定しにくいものであった．すべての行動に対して明示的に評価教示を発するわけではないことが確認された．
- 行動教示の意味は主に，前進，後退，右折，左折，停止の 5 種類，評価教示の意味は肯定・否定の 2 種類があった．また，一つの意味に対して複数の言い方が観察された (たとえば，前進の意味で「進め」「まっすぐ」「前」など)．
- AIBO が行動教示を受けて行動し始めてから 5 秒間次の発話がないとき，99%の確率で正しい行動を行っていた．

3. 意味学習アルゴリズムの提案

本研究では，前進，後退，右折，左折の 4 種類の行動教示の意味を報酬に基づいて学習することを目指す．強化学習の手法の 1 つである Q-learning[9] を学習アルゴリズムとして採用するが，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，人がある言葉を発したことを一つの状態と考え (したがって相異なる言葉の数だけ状態があることになる)，各状態における行動の価値 (ある言葉が発せられたときにある行動をするとどの位の報酬が期待できるか) が言葉の意味であるとする．

本研究では，報酬としては，意識的にせよ無意識的にせよ人から与えられる遅れのない報酬だけを考えることにする (簡単のため，ゴールに到達したことによる報酬等の遅れのある報酬は考えないことにする)．この場合，Q 値の更新式は以下のように簡単になる:

$$Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha r \qquad (1)$$

ここに α は学習率である．

本研究の中心となるアイディアは，上記の予備実験での知見に基づき,「AIBO が教示を受けて行動し始めてから 5 秒以上発話がない場合，その行動が教示通りに行われたと考え，正の報酬をシステムに与える.」ことである．つまり,「よしよし」,「違う」などの明示的な評価教示を受けるだけでなく，評価教示が無いということも肯定的な評価であると判断するわけである．このように発話がな

†京都工芸繊維大学　大学院工芸科学研究科
†Kyoto Institute of Technology

い (No News) 場合に，それがよい知らせ (good news) であると考え (この判断規準を NNC(No News Criterion) と呼ぶ)，これを正の報酬として強化学習を行うアルゴリズムを NNC 学習 (No News Criterion に基づく学習) と呼ぶ．

4. 意味学習アルゴリズムの評価実験

予備実験と同様に，AIBO ERS-7 を言葉でゴールに誘導するタスクを用いた．実験の様子を図 1 に示す．

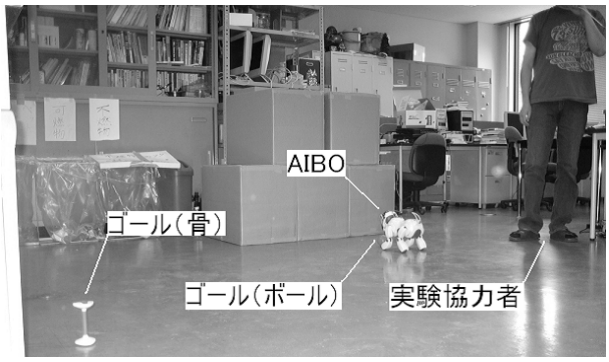

図 1: 実験の様子

4.1 比較評価したアルゴリズム

本実験で比較評価したアルゴリズムは以下の 3 つである．

アルゴリズム 1: ロボットは明示的な評価教示の意味を既に分かっているとし，明示的な評価教示だけを報酬として学習する．NNC は使わない．

アルゴリズム 2: ロボットは明示的な評価教示の意味を既に分かっているとし，明示的な評価教示及び NNC を報酬として学習する．

アルゴリズム 3: ロボットは明示的な評価教示が分からないとし，NNC だけを報酬として学習を行う．

4.2 学習対象の言葉及び音声認識の対象とする言葉

実験では Julius[3] を用いて音声認識を行うが，予備実験で得られた発話データに基づき，実験場面で発せられると予想される言葉を予め登録しておくこととする．

まず，行動教示については，前進，後退，左折，右折の 4 種類の意味の教示を学習させることにするが，予備実験で，一つの意味に対して複数の言い方が用いられることが観察された．そこで，本実験の目的である，アルゴリズムによる学習性能の違いの比較を，比較的短い時間の実験で明らかにするために，今回は，実験協力者にできるだけ同じ言葉を使ってもらえるように仕向け，学習の進行を容易にすることとした．具体的には，実験協力者に,「動かしたいときに使う言葉」として，次の 4 つの言葉が書かれた命令表を提示した．

- 「まえ」「うしろ」「ひだり」「みぎ」

しかし，命令表を渡しても実験協力者は必ずしもその言葉だけを使用するとは限らないので，予備実験から想定されるその他の言葉も登録して音声認識可能にしておく．学習アルゴリズムとしては，音声認識の登録語でありさえすれば，命令表に記載した言葉と記載されていない言葉を区別することなく，意味の学習をすることが可能であるが，命令表に記載していない言葉は使用頻度が少なくなるため，今回の実験中には，実験を実施した時間内に学習が進むだけのデータが得られることはなかった．したがって，以降では，命令表に記載した言葉の意味の学習結果についてだけ論じる．

続いて，評価教示について説明する．予備実験において，実際によく使用された言葉の一つである以下を命令表に「評価したいときに使う言葉」として記載した．

- 「そうそう」「違う」

また，行動教示と同じように，他の想定される言葉も Julius に登録して認識可能にしておいた．

4.3 状態-行動と言葉の意味

3. 節で述べたように，本研究では，人がある言葉を発したことを一つの状態と考え，各状態における行動の価値 (ある言葉が発せられたときにある行動をするとどの位の報酬が期待できるか) をその言葉の意味であるとする．学習精度を分析する 4 つの言葉に対応する状態を，今後は以下のラベルで参照する．

s1: 行動教示「まえ」を音声認識した状態
s2: 行動教示「うしろ」を音声認識した状態
s3: 行動教示「ひだり」を音声認識した状態
s4: 行動教示「みぎ」を音声認識した状態

4.4 実験設定

実験で使用するアルゴリズムは前述の通り 3 つであり，6 名の実験協力者に，3 つのアルゴリズムを実装した AIBO を，実験協力者ごとに異なる順序で誘導してもらった．1 人の実験協力者に対して 1 つのアルゴリズムで 20 分，合計 1 時間の実験を行った．

また，実験の説明表と実験で使用する命令表 (4.2 節を参照) を A4 の紙 2 枚で実験協力者に提示してから実験を行った．実験の説明表には,「言葉で AIBO にゴールまでの道を案内してください．初めは AIBO は言葉があまりわかりませんが AIBO を褒めたり叱ったりしながら道を教えているとだんだんわかるようになってきます．命令表に書いてある言葉を使ってください.」
という言葉と，AIBO に実装されている＜前進＞＜後退＞＜左折＞＜右折＞ 4 種類の行動の説明を記載した．なお本実験においては，＜左折＞＜右折＞は 45 度回転してから前進する実装とした．

本実験では，学習率 $\alpha = 0.1$ とした．また，人の評価による報酬の値は正または負の 2 種類で，その値は $+0.2$ 及び -0.2 とした．行動の選択方法はボルツマン選択を使い，ボルツマン温度は 0.06 に設定した．

5. 実験結果

各状態での行動 (すなわち，各状態に対応する言葉の意味) の学習結果を図 2，図 3，図 4，図 5 に示す．横軸は実験協力者が各行動教示を発話した回数である．縦軸は正しい行動が選択される確率 (Q 値により決まる) である．図上にプロットした点は 6 名の実験協力者の結果を平均したものである．

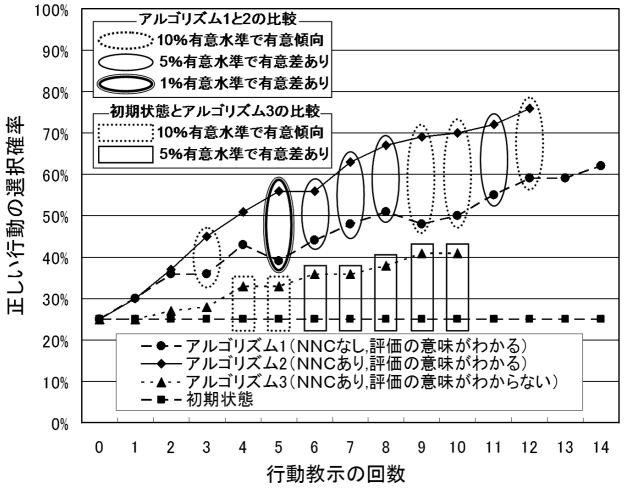

図 2: 状態 s1 での行動学習 (「前」の意味学習) の進行

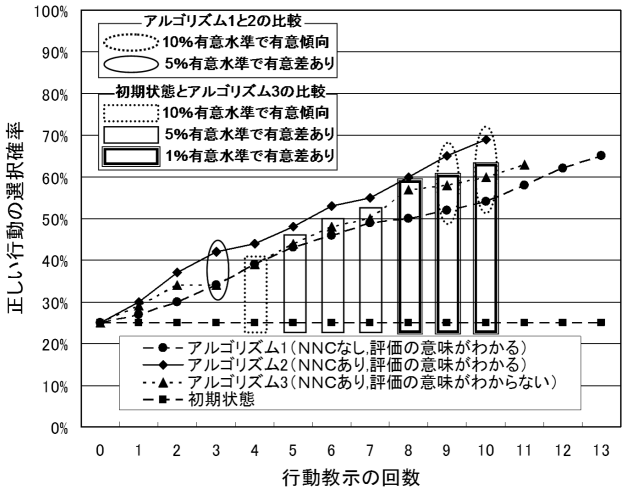

図 3: 状態 s2 での行動学習 (「後ろ」の意味学習) の進行

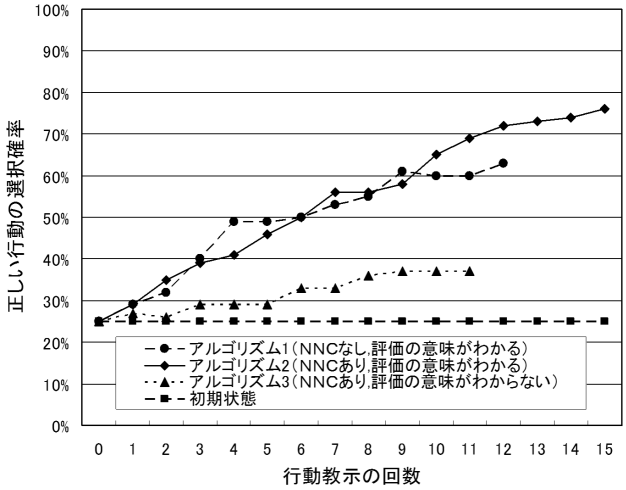

図 4: 状態 s3 での行動学習 (「左」の意味学習) の進行

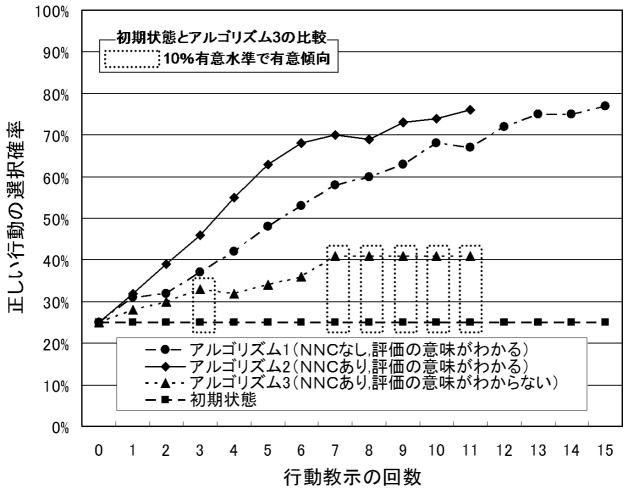

図 5: 状態 s4 での行動学習 (「右」の意味学習) の進行

6. 考察

6.1 アルゴリズム 1 とアルゴリズム 2 の性能比較

アルゴリズム 1 とアルゴリズム 2 の正しい行動の選択確率 (以下，正行動選択確率) の比較を各行動教示回数において，対応ありの t 検定で行った．結果は図 2〜5 に示されており，点線で囲まれた 2 点は 10%の有意水準で有意傾向があり，実線は 5%，二重線は 1%の有意水準で有意差があることを表している．

「前」および「後ろ」の意味学習では，明示的な評価教示のみによる学習 (アルゴリズム 1) より，明示的な評価教示と NNC 両方を用いた学習 (アルゴリズム 2) の方が学習速度が有意に早いことが示された．一方,「左」および「右」の意味学習においては，アルゴリズム 1 とアルゴリズム 2 の正行動選択確率に有意差は無いという結果となった．

我々はこの原因を次のように考えている．AIBO の行動パタン＜右折＞＜左折＞は 45 度旋回してから前進するというものであったため，たとえ AIBO が指示通りの方向に回ったとしても，回転後の進行方向が実験協力者の望む方向とは異なることが多かった．そのため，NNC の判定基準とした 5 秒経過を待たずに次の指示が発せられることが多くなり，結果として NNC を満たす回数が少なく，学習結果に有意差が出なかったのであろう．

6.2 アルゴリズム 3 の性能について

アルゴリズム 3 では，明示的な評価教示の意味を知らないという設定の下で，NNC だけを利用して学習する．図 2〜5 を見ると，アルゴリズム 1 やアルゴリズム 2 と比べると正行動選択確率は低い傾向があるが，学習はある程度進んでいることが分かる．

アルゴリズム 3 の各行動教示回数における正行動選択確率と学習が進んでいない初期状態 (行動の候補が 4 種類であるため各行動の選択確率が 25%である状態) との比較を対応ありの t 検定で行った．図に示すように，この検定結果においても,「前」および「後ろ」の意味学習では，多くの時点でアルゴリズム 3 と初期状態の正行動選択確率との間に有意差がある (すなわち，NNC だけで有意に意味学習が進んでいる) という結果となった．また,「左」および「右」の意味学習では,「前」「後ろ」の学習と比べて，アルゴリズム 3 と初期状態との間の有意差が少なかった．ここでの「前・後ろ」と「左・右」の結果の違いも，前節の考察と同じ原因により生じたと我々は考えている．

6.3 NNC の判定基準とする時間

今回の実験で，行動教示「前」と「後ろ」の意味学習では，NNC を用いることによる学習の加速や，NNC 学習のみによる学習効果が統計的に認められた．しかし，「左」や「右」の意味学習では効果が十分には見られなかった．すでに考察してきたように，この両者の差は，NNC の判定基準とした 5 秒という時間に依存して生まれたものであると考えられる．

図 6 は，NNC の判定基準とする時間と，NNC の適合率 (NNC を満たしたとき，その中の何%が実際に正しい行動だったか)，再現率 (正しい行動を行ったとき，その中の何%が NNC を満たしたか) の関係をグラフにしたものである．今回設定した 5 秒という判定時間では,「前・後ろ」と「左・右」で再現率に差があることがグラフから読み取れる．再現率により NNC の適用回数が変わるため，これが NNC の効果の差となって表れたと考えられる．同様に適合率にも差があるが，この影響は再現率の差による影響より小さいと考える．なぜならば，再現率は正解行動 (たとえば,「前」という発話に対して前進する行動) に報酬を与えた割合であり，1 種類の行動の Q 値に直接影響を与えるが，これに対して，[100 − 適合率] は正解行動以外の 3 種類の行動に誤って報酬を与えた割合であるため，適合率は 3 種類の行動の Q 値に分散して影響を与えることになるからである．よって，平均的には，適合率が与える影響は再現率が与える影響の 3 分の 1 程度になると考えられる．

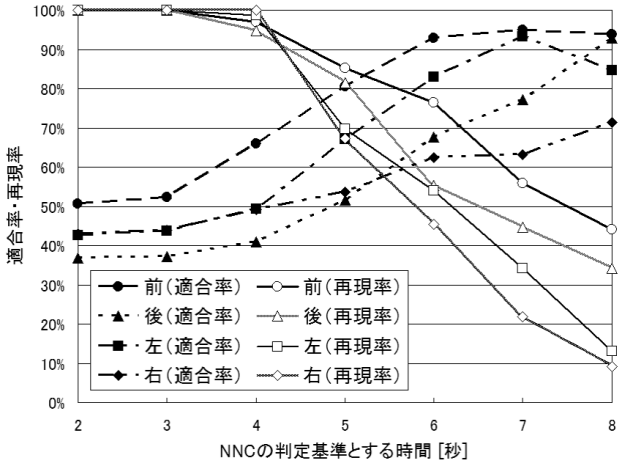

図 6: NNC の判定基準とする時間 (発話の無い時間) と NNC の適合率・再現率の関係

適合率と再現率はトレードオフの関係にあり，今回設定した 5 秒という判定時間は，一見両者のバランスが取れた点であるように見える．しかし，すでに考察したように，適合率が学習に与える影響は再現率が与える影響の 3 分の 1 程度に過ぎないことを考慮すると，判定時間を 4 秒に設定した方が NNC 採用の効果が高まると予想できる．再現率の向上と引き換えに適合率が同程度下がってしまうが，学習への影響が小さい適合率は多少犠牲にしても構わないと考えられるからである．

7. まとめと今後の展開

本論文では，人とロボットとのインタラクションを通じて，簡単な行動教示の意味を学習する際に，NNC が有効な情報を提供しうることを実験的に明らかにした．また，NNC を使用する場合，NNC の判定基準となる時間を適切に設定する必要があるという課題も浮かび上がった．

今後は，以下について取り組む予定である．

- NNC の判定基準として最適な時間の決定方法や，NNC がどのような状況下で有効かを検討する．
- 意味を学習すべき未知のフレーズの候補を音声列から切り出す方法を開発し，意味学習の際に，登録単語の音声認識を前提としないようにする．
- 「NNC 学習」で行動教示の意味の学習だけでなく，他の種類の発話の意味も学習できるアルゴリズムを提案する．

謝辞

本研究の一部は科学研究費補助金　基盤研究 (C)「3 項間インタラクションを通した認知発達メカニズムの解明」の支援を受けた．

参考文献

[1] Fraser, N. M. and Gilbert, G. N., Simulating Speech Systems, Computer Speech and Language, 5(1), 81–99, 1991.

[2] 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, 18(1), 49–58, 2003.

[3] 河原達也, 李晃伸, 連続音声認識ソフトウエア Julius, 人工知能学会誌, 20(1), 41–49, 2005.

[4] Komatsu, T., Utsunomiya, A., Suzuki, K., Ueda, K., Hiraki, K., and Oka, N., Experiments Toward a Mutual Adaptive Speech Interface That Adopts the Cognitive Features Humans Use for Communication and Induces and Exploits Users' Adaptations, International Journal of Human-Computer Interaction, 18(3), 243–268, 2005.

[5] Roy, D. K., Learning Words from Sights and Sounds: A Computational Model, PhD Thesis, Massachusetts Institute of Technology, 1999.

[6] Steels, L. and Kaplan, F., AIBO's first words: The social learning of language and meaning, Evolution of Communication, 4(1), 3–32, 2000.

[7] Sugita, Y. and Tani, J., A Holistic Approach to Compositional Semantics: a connectionist model and robot experiments, Advances in Neural Information Processing Systems 16 (NIPS2003), (Eds.) Thrun, S., Saul, L. K. and Scholkopf, B., The MIT Press, 969–976, 2004.

[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, 9(2), 200–212, 2002.

[9] Watkins, C. J. C. H. and Dayan, P., Q-learning, Machine Learning, 8(3-4), 279–292, 1992.