F-045

自律的な行動学習を利用した教示の意味学習
Learning the meaning of instructions using autonomous action learning

野川 博司 † 岡 夏樹 ‡
Hiroshi Nogawa Natsuki Oka

1. はじめに

人間の意図を理解できれば，それをさまざまなことに利用できる．しかし，どのような情報から意図を推定するのか，情報から意図をどのような方法で推定するのか，について一般的なものがないため困難である．また，さまざまな状況に応じて意図を理解するための情報が異なる．そのため，状況を限定して意図を理解せざるを得ない．このように人間の意図を理解するのは困難であるので，本研究では，人間の意図学習の1つの例として以下のように状況を限定して意図の学習を行っている．

ゴールまでの最適経路を求める迷路探索タスクにおいて，迷路の各地点での最適行動を自律的に学習中のエージェントに現在の状態における最適な行動を示す行動教示（「↑」「→」「↓」「←」）と直前の行動に対する評価を示す評価教示（「○」「×」）を人間が与えていく．はじめエージェントは，与えられた教示の意味が理解できず，選択可能である行動の中から学習中の評価値により行動を選択するが，教示を繰り返し与えていくうちに，エージェントが行っている行動学習と教示を照らし合わせて，教示の意味を学習していく．そして，教示の意味を理解したエージェントは，その教示を利用して行動学習を進めていく．以上の枠組みにおいて，行動教示と評価教示を区別し，できる限り高速かつ高精度で人間の意図と教示の意味の学習を行うモデルを提案するのが本研究の目的である．ここでの意図の理解とは，エージェントが行動した時に人間が教示を与えた場合，その教示が直前のエージェントの行動に対する評価教示，現在のエージェントのとるべき行動に対する行動教示のどちらの種類の教示であるかエージェントが理解することである．また，評価教示の意味を正しく学習するには，行動や状況に対する評価が人間とエージェントとの間である程度一致する必要がある．異なる種類の教示を区別して学習するのは困難である．

異なる種類の教示を区別して学習する研究において，鈴木[1]は，迷路探索タスクにおいて強化学習のQ-Learningを用いて自律的に行動学習した評価値に基づいて，自分がとった行動が良いか悪いか，本来とるべき行動は何だったのかを判断し，状況認識を行うことにより，「Good」「Bad」の評価教示と「上」「下」「右」「左」の行動教示が混在している場合での教示の意味学習に成功している。しかし，評価値を用いて状況認識を行っているので行動学習が進んでいないと教示の意味が学習できず，行動学習のはじめの方では教示の意味学習が行われない．そのため，現在学習している迷路探索タスクでの行動学習に対して，学習した教示の意味を利用して高速化を行うことができていない．また，ここでの行動教示とは直前の行動における最適な行動を教示しており，人間ではなくコンピュータが毎回誤りのない教示を行っている．

そこで，今回提案する教示の意味学習モデルでは，行動学習の進んでない状態でも教示の意味学習を行えるように評価値を用いず，1回の学習ごとの行動学習に要したステップ数の比較を行うことで，教示の意味を判断する方法を用いた．また，ここでの行動教示とは現在の状態における最適な行動を教示しており，教示者は人間であり毎回教示を与えるわけではなく，誤りのある教示も行う可能性がある．

2. モデルとアルゴリズム

2.1 環境モデル

本研究では，単語の意味や外部の環境に対して全く知識を持たないエージェントと，外部の環境に関する知識を十分にもち単語で教示する教示者を想定して，次のようにモデル化する．

ある環境でエージェントが様々な行動をとる場合，教示者は，エージェントの行動と周囲の環境を照らし合わせ，エージェントに教示を与える．一方，周囲の環境に関する知識がない間は，エージェント自身は今自分が行った行動に対する評価ができないので，教示が何をさすのか，つまり教示の正しい意味を学習することができない．しかし，その環境内で行動をしているうちに，学習に要したステップ数の比較によって，教示者からの教示があったときエージェントの行った行動が正しいものであったか，正しくないものであったかという情報をエージェントは得ることができ，エージェントは教示の意味を学習することができる．

2.2 学習アルゴリズム

2.2.1 行動学習

行動学習には強化学習のQ-Learningを用いる．強化学習とは，試行錯誤を通じて未知の環境に適応する学習の枠組みである．[2] 一般的な教師付き学習では学習機構の外部から教師となる理想的な出力が与えられる．しかし，現実的には理想的な出力の例を得ることが困難である問題が多く存在する．これに対し，強化学習では状態入力に対する正しい行動出力を明示的に示す教師が存在せず，かわりに報酬を手がかりに学習する．環境との相互作用の繰り返しを通じて，最適または合理的な方策を学習することが強化学習の目的とされる．Q-Learningとは，状態と行動の組に対する評価値を行動を行うたびに更新して最適行動を学習する手法である．

2.2.2 教示意味学習

今回提案する教示意味学習は、エージェントの行動学習にQ-Learningを用いているが、Q-Learningの評価値

†京都工芸繊維大学大学院 工芸科学研究科, Graduate School of Engineering, Kyoto Institute of Technology
‡京都工芸繊維大学, Kyoto Institute of Technology

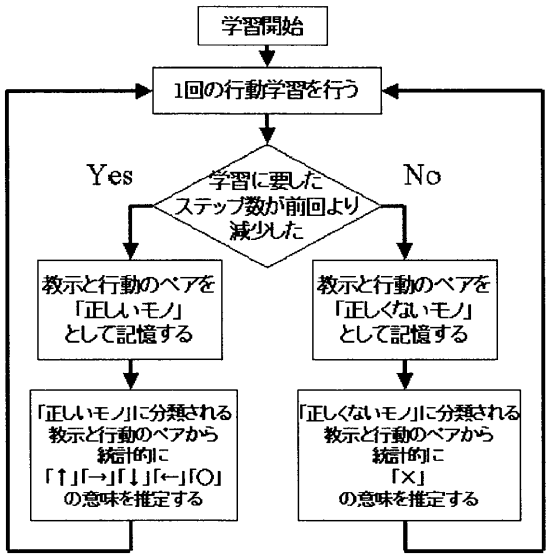

図 1: 教示意味学習の流れ

を用いて教示意味を学習するものではなく，行動学習の進んでない状態でも教示の意味学習を行えるように，学習に要した行動ステップ数の比較により教示意味を学習するものである．以下に，エージェントが人間に与えられた教示の意味を学習する手順について示す．

人間から与えられた教示とそのときエージェントが行った行動をペアにして，覚えておき学習終了時に今回の学習に要したステップ数と前回の学習に要したステップ数を比較する．そのとき，学習に要したステップ数が少なくなっていれば，今回の学習中に与えられた教示に対して，エージェントが行った行動が正しかったとエージェントが判断し教示と行動のペアを覚えておき，学習に要したステップ数が多くなっていれば，正しくなかったとして教示と行動のペアを覚えておくようにする．このような学習を何度も繰り返すことにより，多くの教示と行動のペアが取得できるので，そこから統計的に教示の意味を決定する．(図 1)

教示の意味を決定する方法は，行動教示と評価教示で異なる．

行動教示の意味学習では，学習に要したステップ数が少なくなっていることより正しかったと判断できる教示と行動のペアから，教示ごとの各行動の割合を求めて，他の行動全てに対して，ある一定以上割合が大きい行動があれば，それを教示の意味として推定する．

評価教示の意味学習では，学習に要したステップ数が少なくなっている場合，教示と行動のペアから教示ごとの各行動の割合を求めて，全ての行動である一定以上のばらつきがあり，教示と行動のペアの合計数が最大のとき，それを「良い行動であった」として推定する．学習に要したステップ数が多くなっている場合，教示と行動のペアから教示ごとの各行動の割合を求めて，全ての行動

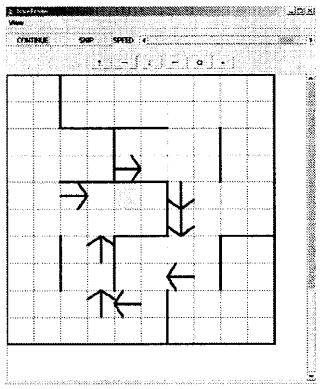

図 2: 実験画面

である一定以上のばらつきがあり，教示と行動のペアの合計数が最大のとき，それを「悪い行動であった」として推定する．これは，前回よりステップ数が少なくなったということは，今回の学習は結果として見れば良いものであり「○」の教示が「×」の教示より多く与えられたはずである。逆に，前回よりステップ数が多くなったということは，今回の学習は結果として見れば悪いものであり「×」の教示が「○」の教示より多く与えられたはずである．という考えをもとにしている．

3. 実験

3.1 実験環境

提案したモデルで，実際に教示学習が可能か検証するために，エージェントによる迷路探索タスクをシミュレーションするソフトウェアを作成して，実験を行う．教示者に提示する画面は，図 2 のようになっている．

3.2 教示を与える場面

エージェントが行動する環境は迷路探索タスクとし，エージェントはゴールまでの最適経路を学習する．このとき行動学習すべき内容は，迷路の各地点での最適行動である．行動学習に用いる強化学習のパラメータは，学習率 $\alpha = \frac{1}{1+(\text{状態での行動回数})}$，割引率 $\gamma = 0.80$，選択法は ϵ-greedy（$\epsilon = 0.3$），評価法は Q-Learning で実験を行った．教示者はエージェントの行動に対して教示を行い，エージェントはこの教示の意味も学習するものとする．迷路探索タスクを学習タスクに選んだのは，タスクの学習が比較的容易で，学習に対する適切な教示を示すのが容易であるためである．

3.3 教示者

人間を教示者として想定している．教示は「↑」「→」「↓」「←」「○」「×」があり，エージェントが試行錯誤で行動している現在の状態に対して，できるだけ早くゴールに着くように教示するものとする．

行動教示として，エージェントの現在の状態に対する最適な行動（「↑」「→」「↓」「←」）を教示する．

評価教示として，エージェントが直前に行った行動がゴールに近づく上で最適であれば「○」教示を与え，そ

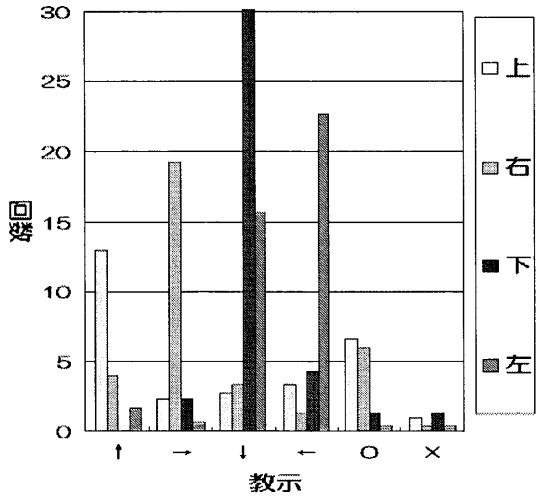

図 3: ステップ数が少なくなったときの教示・行動ペア数の平均値（回）

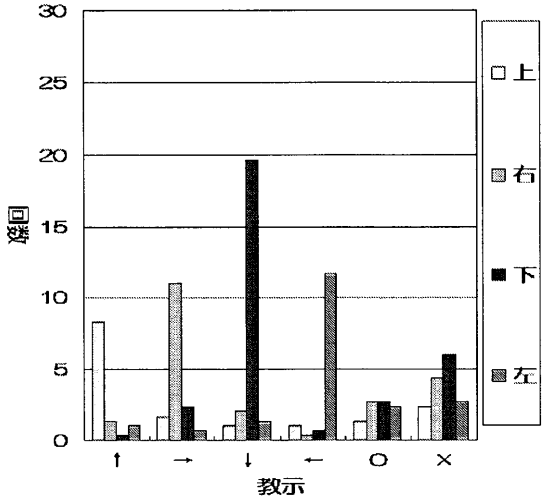

図 4: ステップ数が多くなったときの教示・行動ペア数の平均値（回）

うでなければ「×」教示を与える．

3.4 実験内容

シミュレータ上で行動学習するエージェントに「↑」「→」「↓」「←」「○」「×」の教示を人間が判断して与える．それにより，今回考案した教示意味学習モデルが行動教示と評価教示を区別して，どの程度学習可能であるか調べる実験を行った．

3.5 実験結果

表 1，表 2 に実験で求まった教示と行動のペア数の平均値を示す．

今回の実験結果では，考案したモデルは，迷路探索タスクにおいて人間に与えられる行動教示と評価教示をエージェントが区別でき，行動教示，評価教示の意味学習ともに正しい理解ができた．しかし表 1，表 2 の「○」「×」教示での教示・行動ペア数にあまり差が出ていないため，評価教示の意味学習に対しては「○」と「×」の意味が反対に学習される可能性があり，評価教示の意味学習において今回のモデルは高い精度を持つとはいえないことが示された．これは，考案したモデルがステップ数が少なくなることで，その回の学習が結果として良いものであり「○」の教示が「×」の教示より多く与えられたはずである，という前提をもとにして決定していることによると考えられる．しかし，エージェントの迷路探索タスクに対する行動学習がほとんど行われない状態であっても，行動教示が高い精度で学習でき評価教示においても，ある程度の精度で学習可能であるといえる。

4. おわりに

今後の課題として，エージェントが自律的に行動学習した評価値を用いる教示意味学習アルゴリズムを，今回提案した学習モデルに追加することが考えられる．今回提案した教示意味学習と評価値を用いる教示意味学習を並列で行うことによって，行動学習がほとんど行われていない状態において，今回提案した意味学習により，行動教示学習と評価教示学習を行い，それにより推定した教示を利用して行動学習を進め，行動学習が進んでくると評価値を用いる教示意味学習により教示の最終的な意味を確定するようにする．こうすることにより，行動学習がほとんど行われていない状態から行動学習を完了するまでの全ての状態において，教示の意味学習を行え，片方のみの教示の意味学習アルゴリズムを用いるモデルよりも高速で高精度な意味学習が行えるはずである．

また今回の評価教示の判定法である，教示と行動のペアから教示ごとの各行動の割合を求めて，全ての行動である一定以上のばらつきがあるものが評価教示であるという方法以外に評価教示特有の判定方法を見つける必要がある．なぜなら，今回の方法は行動教示と評価教示を区別して意味学習する場合おいてのみ有効であり，ここにさらに別の種類の教示を加えると今のままの評価教示学習では，新たな種類の教示の意味学習を妨げる可能性があるからである．

今後これらを含めた新しいモデルを考案し評価を行う．

参考文献

[1] 鈴木 健太郎，植田 一博，開 一夫 (2002)．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(2)，200-212.

[2] Richard S. Sutton, Andrew G. Barto. 著，三上貞芳，皆川雅章 訳（2000）.『強化学習』．森本出版.