Q学習を用いた協調行動のための戦略獲得

Strategy Acquisition for Cooperative Actions by Q-Learning

溝口勇太 Yuta Mizoguchi　木村優志 Masashi Kimura　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

We propose a method that an agent learns cooperative action strategies concerning a task after interacting with another agent. Cooperative actions can be acquired by Q-Learning in which an agent gets a reward not only from it's action but also from the other agent's action. We evaluated our method by the experiment in which an agent puts some scattered objects into a certain boxes with another agent cooperatively. The experimental results show that the acquired action strategies are effective for completing the task quickly. In addition, the results also show it is important to share "role of an agent" and "usage of utterance" with a partner agent in order to perform cooperative action.

1. はじめに

近年，ロボット研究，脳研究，および音声言語を含むマルチモーダル対話研究の進展に伴い，ロボットは人間に身近なパートナーとして実社会で用いられるようになってきた．将来は，生活の中に家事ロボットなど，多くのパーソナルエージェントが参加する時代が到来すると予想される．一方，こうした知的エージェントが，人間生活を補佐するのに必要な知識は膨大なため，設計者が予め与えることは将来も不可能であろうと考えられる．このためエージェントが，他者との対話を通して必要な知識や言葉をいかに自律的に学習・獲得する研究が行われてきた [中川 95, 赤穂 97]．しかし，例えば，相手から“エンピツ”と言われた時に，「鉛筆を持って行くのか」，あるいは「鉛筆を使って何かを書くのか」といった判断をすることは語意学習のみからでは不可能である．従って，発話の意味は状況や相手によって変わるため，変化に合わせて行動戦略を修正していく必要がある．

本稿では片付けタスクを例に，他者との対話を通して，相手に合わせた行動，すなわち協調行動を行うための行動戦略を獲得する機構を提案する．二人で片付けタスクを実行する際には，発話の意味が相手の意図によって変わる様な状況が起こる．例えば，同じオブジェクトを指差すという行為でも，それが片付けの依頼を意味していたり，片付ける場所の質問である場合などがある．互いの行動によって得た報酬をエージェント同士が共有することで，Q学習によって上記のような協調行動が獲得できることを示す．

2. 協調行動を獲得するエージェント

エージェントは，他の主体と対話を行いながら語意と行動戦略を学習する．ここで，語意とは，ラベルが示す属性と，その特徴の確率分布のことである．語意学習には，Taguchi らの手法 [Taguchi 06] を用いる*1．また，行動戦略の獲得には，Q学習を用いて質問・応答などの戦略を獲得する手法 [Taguchi 04] を応用している．

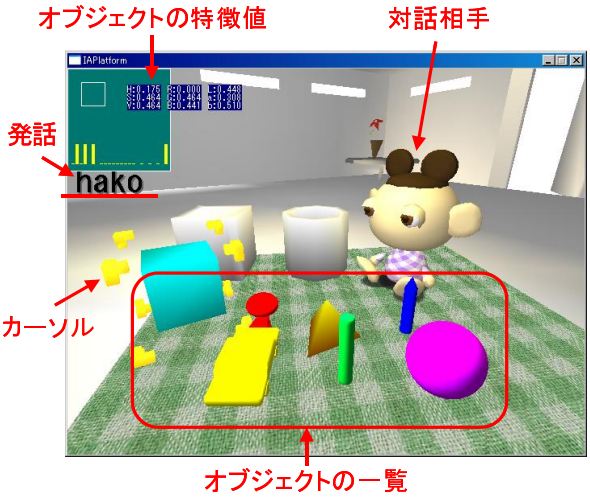

図 1: エージェント実験プラットフォーム

2.1 エージェント実験プラットフォーム

エージェントを用いた協調作業などのシミュレーション実験では，様々な人やエージェントが混在した環境が必要である．我々は，3D 空間内仮想空間にオブジェクトが配置された環境でネットワークを介して対話実験を行うことができるプラットフォームの開発している [溝口 08]．仮想空間内では，オブジェクトを介してエージェント同士がインタラクションを行うことができる．プラットフォームの画面例を図 1 に示す．

2.2 片付け戦略獲得

エージェントは，Q学習を用いて他のエージェントと行動を繰り返しながら片付け行動戦略を学習する．Q学習では，ある

連絡先: 木村優志，豊橋技術科学大学，〒 441-8580 愛知県豊橋市天伯町雲雀ヶ丘 1-1，0532-44-6884, kimura@vox.uttkie.tut.ac.jp

*1 Taguchi らの手法では，ある属性のそのラベルにとって不要であると判断するための距離尺度として，確率分布どうしの相関を用いている．しかし，この方法は分布形の変化という本質的でない部分にも影響を受けてしまう．そこで，本報告では，対称化した KL ダイバージェンス ($D = I_{KL}(p(x_a|l), p(x_a)) + I_{KL}(p(x_a), p(x_a|l))$) を用いている．ここで，$a$ は属性，x_a は属性ごとの特徴ベクトル，$I_{KL}(p(x), q(x))$ は $p(x)$ からみた $q(x)$ の KL ダイバージェンス ($I_{KL}(p(x), q(x)) = \int p(x) \log p(x)/q(x)\, dx$) である．

状態 s で行動 a をとり，次の状態に遷移し報酬 r を得た時，次式のようにQ値を更新する．

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

ここで，α は学習率，γ は割引率である．そして，エージェントは，現在の状態に対する各行動のQ値を取得し，最大Q値を持つ行動を選択する．エージェントの行動を表 1 に，行動の実行規則を図 2 に，Q学習の状態空間を表 2 に，報酬系を表 3 に示す．

エージェントは話題となるオブジェクトに関連する三つの行動と，箱に関連する三つの行動，二つの発話行動，および何もしないという行動の計九つの行動を取る．片付けタスクでは監督エージェントと作業エージェントが交互に行動を実行する．行動を選択し実行する権利を「行動権」と呼ぶ．行動権を得たエージェントは，表 1，および，図 2 に示した行動を自身の戦略に基づいて選択し実行する．そして，「(a2) 指差し」，「(a3) 把持」，「(a5) 運搬」，「(a6) 箱指差し」，「(a8) 発話」，「(a9) 何もしない」のいずれかを実行すると，相手に行動権が移る．

片付けに成功した場合，エージェントに正の報酬が与えられる．また，協調的な片付けを実現するには，片付けに成功したとき一緒に片付けを行った相手にも報酬が与えられるように，両者の間で報酬を共有する必要がある．そこで，エージェントに快，不快，平常の三つの表情をもたせ，これらを直前の相手の行動に対する評価の伝達に使用するようにした．具体的には，対話相手の表情が快であったときにも，表 3 に示すように正の報酬が与えられるようにした．エージェントはオブジェクトを箱に入れる際，相手が自分の知識や要望と一致する箱を選んだなら快の表情を返す．逆に，相手が自分の知識や要望と矛盾する箱を選んだなら不快の表情を返す．また，相手が正しく片付けられているオブジェクトを箱から持ち上げた場合にも不快の表情を返す．それ以外の場合には，平常の表情を返す．表情を伝達しそれに従って報酬を与えることで，対話相手の取った行動の意図を学習することができるようになる．

3. 片付けタスク実験

エージェント同士で対話を行いながら，仮想空間内に散らばるオブジェクトを決められた位置へ片付けるタスクを行う．片付け対象のオブジェクトは，鉛筆，本，ペン の 3 種類，計 45 個である．事前に行った語意学習から，各オブジェクトは，それぞれ，“PENCIL”, “BOOK”, “PEN” というラベルと対応づけられている．また，オブジェクトを片付けるための箱を 3 個用意した．各オブジェクトは種類に応じて正しい片付ける箱がそれぞれ決まっている．実験に使用するエージェントは，監督エージェントと作業エージェントの二体である．監督エージェントは，「どのオブジェクトをどこに片付けるか」を知っている．しかし，自らは片付け行動を行わず，作業エージェントに指示を与える．作業エージェントは指示に従って片付けを実行する．両エージェントは，片付けが効率的に行われるように，指示や質問などの行動を学習する．本実験では，監督エージェントと作業エージェントがそれぞれ 1 回ずつ行動することを 1 ターンと呼ぶ．まず 15,000 ターンの対話を行い，片付けタスクのための対話戦略を両エージェントに学習させる．そして両エージェントが獲得された対話戦略を使い，新たに 1,000 ターンの対話を行うことで対話戦略を評価する．

評価時の各ターンでの片付け済みオブジェクトの割合を図 3 に示す．破線のグラフは初期の Q 値をランダムに与えた場合（ランダム戦略），実線のグラフは獲得した戦略を用いた場合

表 1: エージェントの行動

(a1)	オブジェクト選択	ランダムにオブジェクトを選択する
(a2)	指差し	選択したオブジェクトを指差す
(a3)	把持	選択したオブジェクトを持つ
(a4)	箱選択	ランダムに箱を選択する
(a5)	箱指差し	選択した箱を指差す
(a6)	運搬	持っているオブジェクトを選択した箱の中へ入れる
(a7)	単語追加	注目す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
(a8)	発話	発話レジスタの内容を発話する
(a9)	何もしない	何もせずに相手に行動権を移す

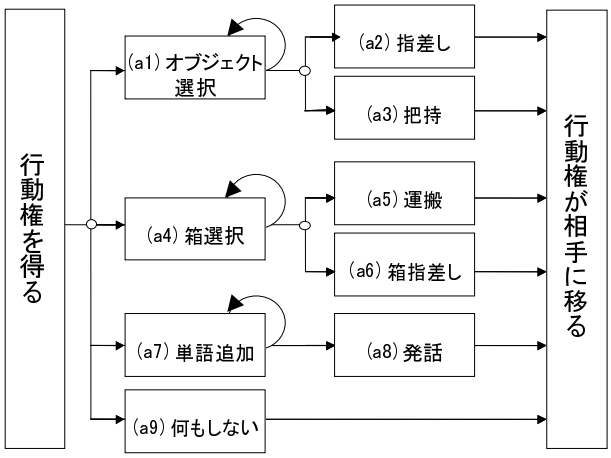

図 2: エージェント行動の実行規則

表 2: Q学習の状態空間

(s1)	相手の表情	(快，平常，不快)
(s2)	相手の行動	
(s3)	相手の把持しているオブジェクトの数	
(s4)	相手の把持しているオブジェクトと発話の概念の一致数	
(s5)	相手の表情	(快，平常，不快)
(s6)	自分の前の行動	
(s7)	自分の把持しているオブジェクトの数	
(s8)	自分の把持しているオブジェクトと発話の概念の一致数	
(s9)	発話と話題オブジェクトの概念の一致数	
(s10)	発話と話題箱オブジェクトの概念の一致数	
(s11)	話題オブジェクトの変化	(0: 変化なし，1: 増, 2: 減)
(s12)	話題オブジェクトが箱内にあるか	

表 3: Q学習の報酬系

(r1)	話題オブジェクトが指示通り，かつ，正しく片付けられた	+25
(r2)	話題オブジェクトが指示通り	0
(r3)	話題オブジェクトが間違った場所に片付けられた	-5
(r4)	(r1), (r2), (r3) 以外の場合	-1
(r5)	相手の表情が快	+25
(r6)	相手の表情が不快	-5
(r7)	相手の表情が平常	-1

の結果である．グラフは 20 回の試行の平均したものである．図 3 から，獲得した戦略を用いた場合は，約 400 ターンでほぼ全てのオブジェクトが片付けられることが分かる．戦略を獲得した方が，ランダムに片付けを行ったグループよりも早く片付けを行えた．これは，協調して片付けを行う戦略を獲得することが，片付けタスクにおいて有効であることを示している．さらに，獲得した戦略に着目し考察する．戦略を獲得させたエージェントは，片付け手順によって 2 種類に分類できる．そ

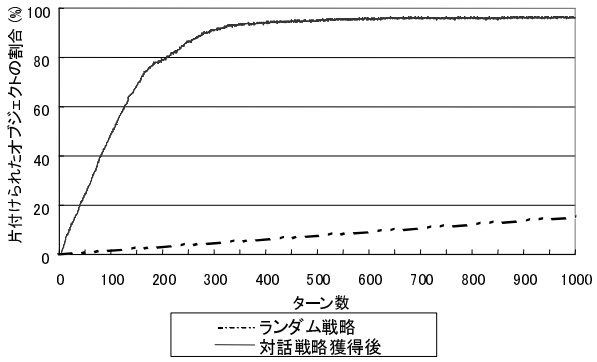

図 3: 片付けられたオブジェクトの割合

のときに獲得されたエージェントの戦略をそれぞれ「質問-応答戦略」と「依頼-遂行戦略」と呼び分け，詳細を以下に示す．

□ 質問-応答 戦略

質問-応答戦略では以下の流れで片付けが実行された．

1. 作業エージェント：「把持」（なにかを持ち上げる）
2. 作業エージェント：「発話」（持ち上げたオブジェクトについて質問する）
3. 監督エージェント：「指差し」（正しい場所を指示）
4. 作業エージェント：「運搬」（指示された場所へ運ぶ）

まず作業エージェントは，片付けが終わってないオブジェクトをランダムに選び持ち上げる．しかし，片付け先については知らないため，知っている監督エージェントにそのオブジェクトについて発話を行う．それを受け，次に監督エージェントは，「指差し」によって作業エージェントに片付け先を指示する．作業エージェントの発話が片付ける場所を相手に尋ねるという，質問の意味を担っていることがわかる．

□ 依頼-遂行 戦略

依頼-遂行戦略では，以下の流れで片付けが実行される．

1. 監督エージェント：「発話」（片付けるオブジェクトについて発話する）
2. 作業エージェント：「把持」（発話されたオブジェクトを持ち上げる）
3. 監督エージェント：「指差し」（正しい場所を指示）
4. 作業エージェント：「運搬」（指示された場所へ運ぶ）

まず，監督エージェントが，片付けて欲しいオブジェクトについて発話する．次に作業エージェントは監督エージェントの発話を受けて，仮想空間内に存在するオブジェクトの中から発話と一致するオブジェクトを探して持ち上げる．そして，監督エージェントはその持たれたオブジェクトを確認し，正しい片付け先を指差す．これを受け手，作業エージェントは持ち上げたオブジェクトを指指された場所で運び，片付けを達成していた．監督エージェントの発話や，指差しが作業エージェントに対する作業の依頼という意味を担っていることがわかる．「質問-応答戦略」では作業エージェントが発話を行うことをきっかけとして監督エージェントと協調行動を始めている．一方，「依頼-遂行戦略」では，監督エージェントが作業エージェントに逐次行動を促すように行動することで協調作業を行っている．両戦略における行動の違いは，監督エージェントと作業エージェントとの間で互いの役割が共有されたためだと考える．つまり，対話戦略を学習した 15,000 ターンの対話の中で，片付けタスクを最も効果的に実行するため，エージェントは対話相手に合わせて自分の行動を選択できるようになった．このことから，協調行動の獲得には相手との対話の中で試行錯誤を繰り返し，自分と相手の役割を共有していくことが重要だと考える．

また両戦略におけるエージェントの役割の違いは発話行動にも影響している．「質問-応答戦略」では「発話」は質問の形式で使われ，一方「依頼-遂行戦略」では「依頼」の形式で使われていることがわかる．しかしどちらの場合も，対話相手となるエージェントがこの発話行動の違いに合わせて適応し，片付けに繋がる行動を返している．このことから，対話相手と協調的に作業を行う場合，相手との間で発話行動を共有することが重要だと考える．

「質問-応答戦略」では作業エージェントが積極的に質問を行い，監督エージェントがそれに協調的な行動，すなわち質問に対する応答を返すことで対話が成り立っている．一方「依頼-遂行戦略」では，監督エージェントが積極的に作業エージェントに依頼を行い，作業エージェントが協調的な行動，すなわち指示されたオブジェクトを把持したり運搬したりすることで片付けが成立する．

4. まとめ

本論文では，エージェントによる協調行動の獲得を目指し，確率分布に基づく語意学習方法と，強化学習を用いた対話戦略の学習方法を組み合わせた手法を提案した．実験の結果から，片付けタスクにおいて協調行動を獲得することで早く片付け行動が行えることをしめした．獲得した行動戦略は，「質問-応答戦略」，「依頼-遂行戦略」の 2 種類であった．前者は，知識のないエージェントが必要とすることを積極的に質問し，それに対する答えを持っているエージェントが答えるという協調行動がみられれた．後者では，知識のあるエージェントが積極的に知らないエージェントに指示を出し，作業エージェントはそれに合わせて行動した．この結果より，協調行動では，相手や状況に応じた自分の役割や行動を決めることが重要だと考える．

しかし，タスクに応じて報酬系・状態空間を事前に設計することは困難である．さらに，今回提案した手法では，獲得した戦略を状況に応じて使い分けるようなことができない．今後は，強化学習の報酬系・状態空間をタスクに応じて自律的に獲得する手法や，状況に応じて戦略を使い分けるような機構を実装していきたい．

参考文献

[中川 95] 中川聖一, 升方幹雄, “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”. 人工知能学会誌, Vol. 10, No. 4, pp. 619-627, 1995.

[赤穂 97] 赤穂, 速水, 長谷部, 吉村, “EM 法を用いた複数情報源からの概念獲得”, 信学論 Vol.J80-A No.9 pp.1546-1553, 1997.

[Taguchi 04] Ryo Taguchi, Kouichi Katsurada and Tsuneo Nitta, “Automatic Acquisition of Dialog Strategies for Concept Learning through Interaction among Agents”, 2004 IEEE Conference on Cybernetics and Intelligent Systems, pp1048-1053

[溝口 08] 溝口 勇太，田口 亮，木村 優志，土井岡　伴哉，桂田 浩一，新田 恒雄, “知的エージェント学習実験プラットフォームの構築”，人工知能学会全国大会，3E3-8 (2008-6)

[Taguchi 06] Ryo Taguchi, Masashi Kimura, Shuji Shinohara, Kouichi Katsurada, and Tsuneo Nitta, “Implementation of Biases Observed in Children’s Language Development into Agents”, Symbol Grounding and Beyond (LNAI 4211): In Proc. of the Third International Workshop on the Emergence and Evolution of Linguistic Communication, pp. 143-167. Springer. (Roma, 2006.9)