強化学習と期待効用最大化と階層ディリクレ過程に基づくロボットによる最適支援行動選択と場所の分節化

The Optimal Support Action Selection and the Segmentation of Space by Robots Based on Reinforcement Learning, Expected Utility Maximixation and Hierarchical Dirichlet Process

牧野 知也[1] 岩橋 直人[1] 国島 丈生[1] 中村 友昭[2] 長井 隆行[2]

Tomoya Makino[1] Naoto Iwahashi[1] Takeo Kunishima[1] Tomoaki Nakamura[2]

Takayuki Nagai[2]

[1] 岡山県立大学

[1]Okayama Prefectural University

[2] 電気通信大学

[2] The University of Electro-Communications

本論文では，強化学習を用いて人間の移動先を予測し，期待効用最大化に基づき最適行動を選択し，階層ディリクレ過程を用いて場所の分節化を行う手法を提案する．本手法によれば，ロボットは人間に指示されることなく自らの判断で扉を開けるなどの最適な支援行動を行うことができるようになる．さらに，ロボットに部屋という概念を獲得させることもできる．シミュレーション実験により，本手法の有効性を示した．

1 はじめに

近年，少子高齢化により家庭や施設にロボットを導入し，高齢者や居住者を支援することが期待されている．そのようなロボットは，人間に指示されることなく自らの判断で最適な支援行動を行うことが望まれる．関連研究としては，ある人物の過去の行動データを元にデータを取った人間の行動を予測してロボットに支援行動を行わせる研究[1，2]や，場所ごとにどのような目的を持った人が多いのかというデータを収集し，場所ごとに行動の予測を変えて支援行動を行う研究が挙げられる[3]．しかし，これらの手法では事前に特定の人間の行動データや特定の場所を訪れた人間の行動データを多く集めておく必要がある．さらに，これらのデータはデータを収集した人間の生活や場所の状態が今後も大きく変わらないことが条件である．そのため，例えばデータを収集した人間が学生から社会人になるなどの大きな変化が起きた場合やデータを収集した場所周辺の建物が 1 つ無くなるなどの変化が起きた場合，再度データを集め直さなければならない可能性がある．

本論文では強化学習[4]を用いて人間の移動目的地を予測し，期待効用最大化[5]に基づき最適行動を選択を選択し，階層ディリクレ過程(Hierarchical Dirichlet Process，HDP)[6]を用いて場所の分節化を行う手法を提案する．まず，強化学習を用いてロボットに行動範囲内の行動価値関数を学習させ，これを元に人間の移動目的地を予測させる．そして予測した結果から期待効用最大化に基づいて最適な支援行動を選択する．これにより，ロボットは人間に指示されることなく予測した目的地までの扉を開けるなどの行動支援が行えるようになる[7]．さらに，全状態の状態価値関数を HDP によりモデル化し、状態空間を分節化する。これにより，ロボットに部屋という概念を獲得させることができる．部屋という概念を獲得することでロボットはそれぞれの部屋を廊下，キッチンなど名前で認識することが可能になり，部屋ごとに異なる行動を行わせることができる．

本論文の構成は以下の通りである．まず 2 章では行動予測の手法について述べ，3 章では行動予測から最適行動を選択する手法を述べる．4 章では HDP について述べ，5 章ではシミュレーションによる実験条件と実験結果を示し，6 章では考察を述べる．7 章はまとめである．

2　強化学習に基づく人間の行動の予測

まず，人間の移動先を予測する手法について説明する．ロボットは初期状態では何の情報も与えられていないものとする．ロボットは強化学習を用いて建物内に存在する各目的地に対しての最適方策を学習する．提案手法では強化学習の一つであるQ学習を用いて学習を行う．Q学習は以下の式(1)により表され，ロボットはε-greedy 選択を用いて建物内を自律的に行動して各目的地への最適方策を学習する．

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma max_{a'} Q(s',a') - Q(s,a)] + a)] \quad (1)$$

ここで$Q(s,a)$とは行動価値関数のことであり，Q値と呼ばれる値である．状態sで行動aを行った際，その行動がどれだけ良いかを表している．sは遷移前の状態，aは遷移時の行動，rは遷移によって得られた報酬，$max_{a'} Q(s',a')$は遷移先の状態s'で最もQ値が高くなる行動a'を行った際のQ値，αは学習率，γは割引率である．ロボットはε-greedy 選択を用いて自律的に行動し，式(1)を更新することで全ての状態，全ての行動のQ値が求められ，ロボットは各目的地への最適方策を学習することができる．

このようにロボットが自律的に行動してQ値を求めることにより，ロボットが人間の行動を観測した時に建物内に存在する各目的地に対しての行動確率を求めることができる．目的地がN個ある場合，目的地$T_i (i = 1, \dots, N)$への行動確率は以下の式(2)によって表されるものとする．

$$P(T_i) = \frac{1}{k} \sum_{t=1}^{k} \frac{Q^{T_i}(S_t, a_x)}{\sum_{j=1}^{N} Q^{T_j}(S_t, a_y)} \quad (2)$$

ここで$P(T_i)$は目的地T_iへの行動である確率，tは何度目の行動であるかを示し，$Q^{T_1}(S_t, a_x)$は$t-1$回目の遷移先の状態で行動a_xを行った際の目的地T_iへのQ値，kは行動回数である．この確率を建物内の各目的地に対して求めることで，ロボットはどの目的地への行動であるかを予測することができる．

3　期待効用最大化に基づくロボットによる最適支援行動の選択

次に，ロボットの最適支援行動選択について説明する．ロボットは人間の行動を予測する際，正確に移動先を予測しなければならない．しかし，人間の行動を観測するだけで正確に行動先を予測することは不可能である．そのため異動先を予測するだけでなく，予測の不確からしさによってはロボットが人間に移動先を質問する手法を提案する．提案手法ではこの判断は期待効用を用いて行う[8].

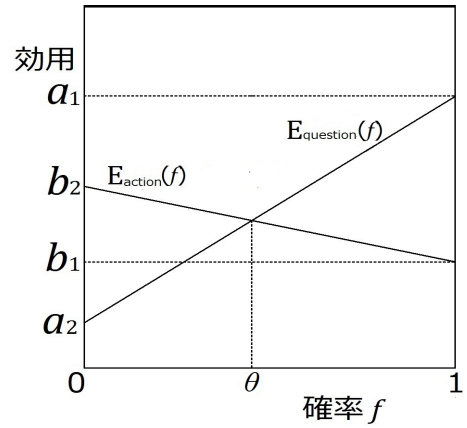

図 1：効用と確率の関係

まず，ロボットの動作として，人間の現在位置から予測した目的地までの間にある扉を開ける行動$action$と人間へ行動先を質問する行動$question$のどちらかの動作を行うものとする．この時，ロボットの行動$action$と行動$question$の期待効用$E_{action}(f), E_{question}(f)$は式(4)，式(5)のように示すことができ，図1は2つの式を図に表したものになる．

$$E_{action}(f) = a_1 f + a_2(1-f) \quad (0 \le f \le 1) \quad (4)$$

$$E_{question}(f) = b_1 f + b_2(1-f) \quad (0 \le f \le 1) \quad (5)$$

図 1 において，式(2)より求めた確率の中で最も高い確率をfとしてa_1, a_2を行動$action$の効用，b_1, b_2を行動$question$の効用とする．

この時，$E_{action}(f) = E_{question}(f)$は 0< θ <1 なる解θを持つ．このθを境界として$f < \theta$ならば$E_{action}(f) < E_{question}(f)$であり，$\theta < f$ならば$E_{question}(f) < E_{action}(f)$であることが図1より分かる．つまり，$f < \theta$ならば行動$question$，$\theta < f$ならば行動$action$を行う方が良い．よって，提案手法ではこの$\theta$の値を元にどちらの動作を行うかを決定する．

4　HDP による場所の分節化

HDP とは，ディレクレ過程を階層化することで文書のようなデータ(単語)の集合の分類を可能とした生成モデルであり，文書集合全体のトピック数およ

び文書ごとのトピック数を推定することができることが特徴である．

本論文では各状態を文書とし，各目的地を単語とし，目的地への状態価値関数の大きさを単語の出現頻度とした．トピック数を推定しながら，各状態をトピックに分類することができ，部屋という概念が生まれる．なお，状態価値関数は以下の式(3)より求められる．

$$V(s) = \sum_a \pi(s,a) Q(s,a) \quad (3)$$

$\pi(s,a)$とは状態 s で行動 a を行う確率である．式(3)を用いることにより，各状態の状態価値関数を求めることができる．ロボットは部屋という概念を獲得し，行動予測を行う際に現在人間のいる部屋への行動はありえないなどと判断することができる．また，各部屋に名前をつけることで部屋ごとに異なる行動をロボットに行わせることができる．

5　実験

提案手法を用いてシミュレーション実験を行った．図 2 にシミュレーションに用いた建物の内部を示す．T_1，T_2，T_3，T_4，T_5，T_6をそれぞれ目的地とし，◯をロボットとする．また，太い線は壁を表しており，壁で囲まれた各部屋と部屋の間には扉があるものとした．ロボットは建物内を自律的に行動してQ学習を行い，人間が動いた時にその行動から行動先を予測して行動$action$もしくは行動$question$を選択して行動を行う実験を行った．なお，行動$action$を行った場合，ロボットは人間の現在位置から予測した目的地までの間にある扉を全て開けるものとする．また，行動$question$を行った場合は回答された目的地までの間にある扉を全て開けるものとする．以下，人間の行動予測，最適行動選択，および場所の分節化それぞれに関する実験について，順番に記述する．

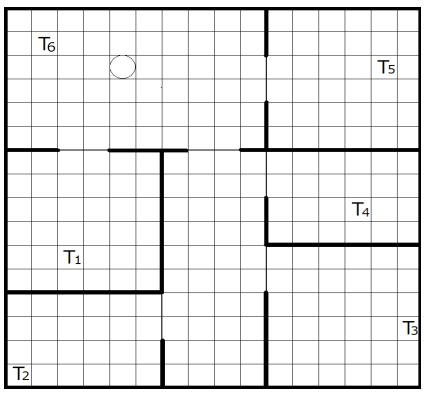

図 2：実験に用いた建物の内部

5.1 人間の行動予測に関する実験

5.1.1　条件

実験ではロボットの初期位置はランダムとし，上下左右にのみ行動できることとした．また，Q学習の学習率αを 0.1，割引率γを 0.99 とした．Q値の初期値は壁に対しての行動のみ−10 を与え，その他は全て 0 とする．報酬は学習中の目的地へ到着する行動を行った時のみ与えられ，報酬の値は 1 とした．つまり，目的地T_1への最適方策を学習している場合，ロボットは初めランダム方策を行い，目的地T_1へ到着する行動を行った場合のみ報酬 1 が与えられる．また，ロボットの行動選択方法として ε -greedy 選択を用い，ε の値は 0.1 とした．ε -greedy 選択とは，多くの場合は最適な行動を行うが，ある確率 ε においてランダムな行動を行う行動選択手法である．つまり，提案手法では毎回 10%の確率でランダム方策を行い，90%の確率では最適方策を行うことになる．また，ロボットは人間の現在位置を知っているものとする．

5.1.2　結果

ロボットは建物内に存在する全ての目的地への最適方策を獲得することができ，最適方策の獲得には 1 つの目的地に対して約 25 万回Q値の更新が必要であった．また，Q値の最高値はいずれも 1 となり，最低値は様々であったが，いずれも約 0.75 が最低値となった．

人間の行動予測に関する実験結果において，例として実験で行った行動経路を図 3 に示す．なお，図３に示す行動 1 は目的地T_6から目的地T_1への最適行動を示し，行動 2 は目的地T_6から目的地T_2への最適行動を示し，行動 3 は目的地T_6から目的地T_5への最適行動を示す． ただし，それぞれの行動経路は人間の行動を 10 回観測するか，いずれかの目的地への行動確率が θ を超えるまでを示している．これは本論文の実験においては人間が 10 回行動するまでにロボットが支援行動を行うようにしたためである．

図 3 に示した行動を行った際のそれぞれの目的地への行動確率の変化を図 4，図 5，図 6 に示す．左上の 1 回目から順番にそれぞれ目的地T_6から 1 回人間の行動を観測した際の各目的地への行動確率，2 回人間の行動を観測した際の各目的地への行動確率，3 回人間の行動を観測した際の各目的地への行動確率を示す．これらを人間が 10 回行動するか，いずれかの目的地への行動確率が θ を超えるまでを表している．また，各目的地への行動確率は式(2)を用いて計

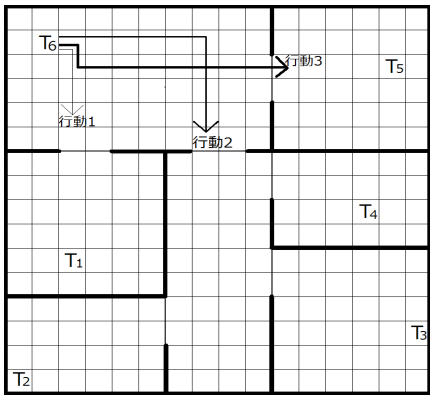

図 3：実験に用いた行動経路

算した．ただし，ロボットは現在人間が目的地T_6の部屋にいることを知っているため目的地T_6への支援行動は必要ないものとし，目的地T_6への行動確率は考えないものとする．そのため目的地は 5 つとなり N=5 として計算した．

図 3 に示した行動を行った際のそれぞれの目的地への行動確率の変化を図 4，図 5，図 6 に示す．左上の 1 回目から順番にそれぞれ目的地T_6から 1 回人間の行動を観測した際の各目的地への行動確率，2 回人間の行動を観測した際の各目的地への行動確率, 3 回人間の行動を観測した際の各目的地への行動確率を示す．これらを人間が 10 回行動するか，いずれかの目的地への行動確率が θ を超えるまでを表している．また，各目的地への行動確率は式(2)を用いて計算した．ただし，ロボットは現在人間が目的地T_6の部屋にいることを知っているため目的地T_6への支援行動は必要ないものとし，目的地T_6への行動確率は考えないものとする．そのため目的地は 5 つとなり N=5 として計算した．

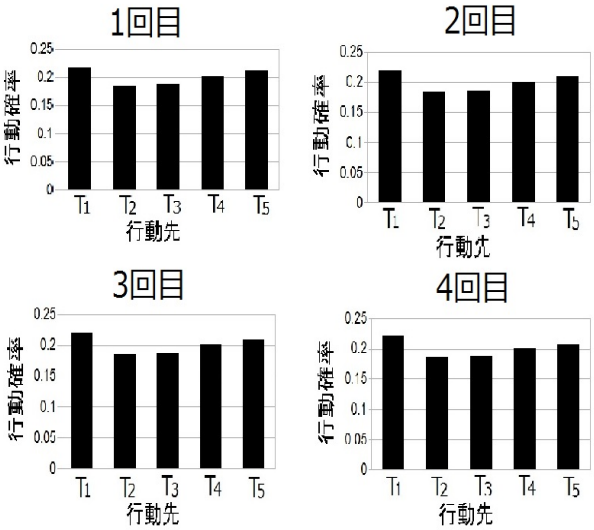

図 4：行動 1 を行った時の行動確率の変化

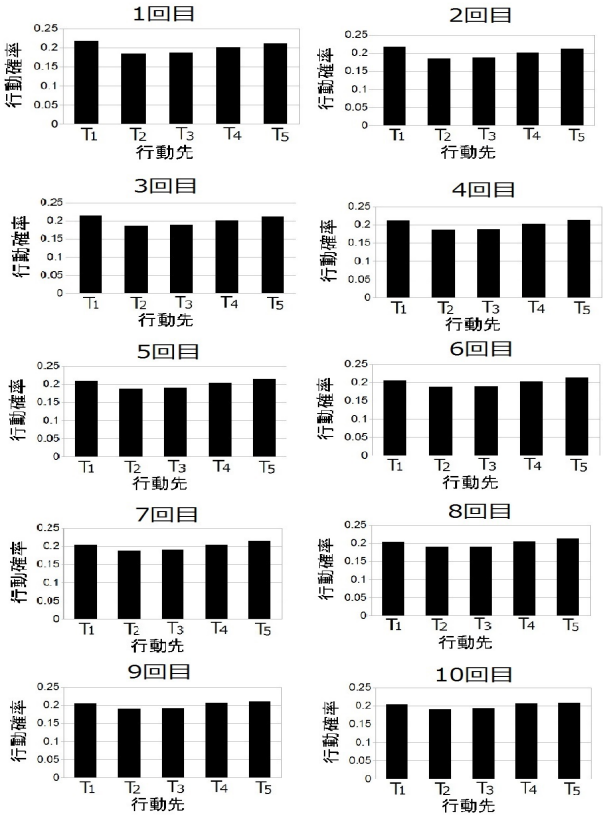

図 5：行動 2 を行った時の行動確率の変化

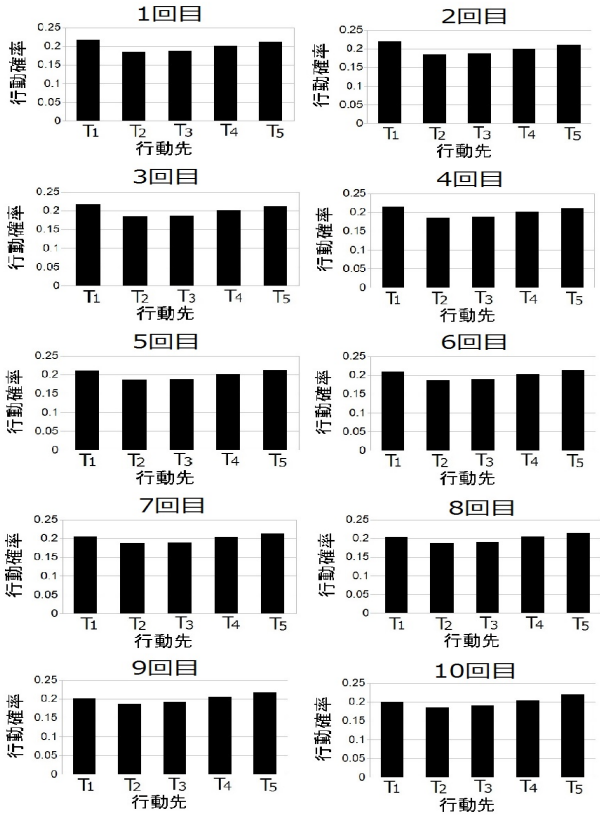

図 6：行動 3 を行った時の行動確率の変化

表 1：効用の設定値

動作 \ f	0	1
$action$	5.0	15.1
$question$	8.0	7.0

5.2 最適行動選択に関する実験

5.2.1 条件

提案手法で用いた効用の値を表 1 に示す. この時, θ の値は 0.220 である. 人間が 10 回行動するまでにいずれかの目的地への行動確率が θ を超えた場合には動作$action$を行い, 人間が 10 回行動しても θ の値を超えなければ動作$question$を行うこととした.

5.2.2 結果

図 4 より, 行動 1 では 4 回目の行動で目的地T_1への行動である確率が最も高く θ より値が大きくなるため, ロボットは目的地を正しく予想して行動$action$を行っていることが分かる. また, 図 5 より行動 2 では 10 回行動を行ってもどの行動確率も θ の値より大きくならないため, 行動$question$を行い人間へ行動先を質問するであろうことが分かる. 最後に図 6 より, 行動 3 でも 10 回行動を行ってもどの行動確率も θ の値より大きくならないため, 行動$question$を行い人間へ行動先を質問するであろうことが分かる. これらのことから, ロボットは人間が目的地T_6から目的地T_1, 目的地T_2, 目的地T_5へ移動した際に正しく目的地を予測して動作を行っていることが分かる.

5.3 場所の分節化に関する実験

5.3.1 条件

Q学習を用いて得られた行動価値関数から式(3)を用いて状態価値関数を求め, 状態価値関数を用いて HDP による場所の分節化を行った.

5.3.2 結果

次に場所の分節化に関する実験結果を記す. 図 7 はトピックごとに場所の色分けを行った結果である.

図 7 より, 大雑把ではあるがそれぞれの部屋ごとに分節化することができたことが分かる. また, 目的地がない真ん中の部屋においても周りの部屋とは違うトピックを持つことができたことが分かる.

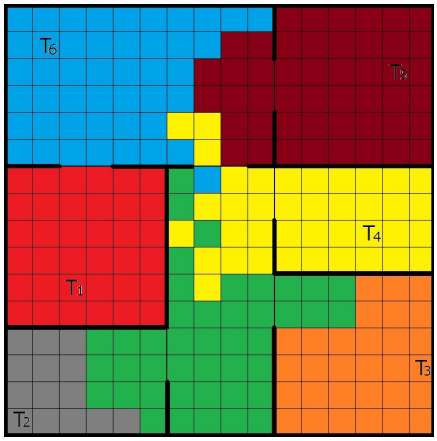

図 7：場所の分節化を行った結果

6 考察

人間の行動予測に関する実験結果である図 4, 図 5, 図 6 より, ロボットは目的地T_6から目的地T_1, 目的地T_2, 目的地T_5へ行動した際に正しく行動先を予測していることが分かる. また, 最適行動選択に関する実験において行動 1 では人間に指示されることなく人間の現在位置から目的地T_1までにある扉を開けており, 行動 2 と行動 3 においても人間に質問することで目的地T_2, 目的地T_5までの扉を開けていることが分かる.

しかし, 実際には行動 2 と行動 3 でも移動先を目的地T_2, 目的地T_5と予測し, 動作$action$を行うことが望ましい. よって今後の課題としては以下の 2 つが挙げられる.

まず, 効用値の設定である. 実験では効用値は予備実験を行い, 経験的に決定した. しかし, それでは常に最適な効用値を設定することはできない. また, 他の場所からの行動でも正しく動作する保証はない. そのため効用値を自動で更新し, 最適化する手法などを考える必要がある.

次に, ロボットが行動を行うタイミングを変えることが挙げられる. 実験では 10 回行動した時に判断を行うように設定したが, それでは早すぎる場合がある. 例えば実験で行った行動 2 の場合, 20 回行動を行えばロボットは目的地T_2への行動と予測することができた. また, 今回用いた内部状態では目的地T_2, 目的地T_3, 目的地T_4の判断は難しいため, θ を超えても一定条件を満たすなら動作$action$を行わないなどの手法も考えられる. その他にも多層マルチモーダル LDA を用いることで概念学習や知識獲得を行い, 廊下を歩く, キッチンで飲む, などの行動を理解することで現在の人間の行動から次に人間の行う行動をより正確に予測できる可能性がある[9].

以上より, 効用の自動更新の手法やロボットが動

作$action$もしくは動作$question$を行う条件を考えることが今後の大きな課題となる.

7 まとめ

本論文では強化学習と期待効用最大化に基づいて人間の行動先を予測し，ロボットが自らの判断で最適な支援行動を行う手法，および行動支援を行うための情報を用いて HDP により場所を分節化する手法を提案した. シミュレーション実験の結果により，提案手法を用いることでロボットが人間の行動先を比較的良好に予測し，支援行動を行え，ある程度期待通りに場所を分節化できることを示した. 今後は，各部屋ごとに異なる支援行動戦略を設定するなどして，ロボットがより自律的に自らの判断で行動できるようにし，実用性を高めてゆく予定である.

参考文献

[1] 真部靖弘，服部元史，田所諭，高森 年：ペトリネットによる人間の行動パターンモデルと行動予測, 日本機械学会論文集 C 編, Vol. 63, No.609，pp.1693-1700，1997.

[2] 福田司，中内靖，野口勝則，松原隆：自律移動ロボットとタッチパネルを利用した調理作業支援システム，日本機械学会論文集 C 編，Vol.72，No.716，pp.1215-1222，2006.

[3] 佐竹聡, 神田崇行, Dylan F. Glas, 塩見昌裕，石黒浩，萩田紀博：環境情報を理解してサービス提供を行うロボットの実現，情報処理学会 インタラクション，2009.

[4] 三上貞芳，皆川雅章：強化学習，森北出版，2000.

[5] Stuart Russell，Peter Norvig : エージェントアプローチ人工知能第 2 版，吉川康一監訳，共立出版，2008.

[6] 岩田具治：トピックモデル，購談社，2015.

[7] 牧野知也, 岩橋直人, 国島丈生 : 強化学習と期待効用最大化に基づくロボットによる人間の行動予測と最適支援行動選択, 第 17 回 IEEE Hiroshima Student Symposium，2015.

[8] 杉浦孔明，岩橋直人，柏岡秀紀，中村哲：言語獲得ロボットによる発話理解確率の推定に基づく物体操作対話，日本ロボット学会誌，Vol.28，No.8，pp.978-988，2012.

[9] 長井隆行，中村友昭，アッタミミ・ムハンマド，持橋大地，小林一郎，麻生英樹：多層マルチモーダル LDA と強化学習による意味理解に基づく行動決定，人工知能学会全国大会論文集，29，1-4，2015.