Graph2vecによる世界モデルの分散表現獲得と他者世界モデルの推定

○境辰也（大阪大）　堀井隆斗（大阪大）　長井隆行（大阪大／電気通信大）

1. はじめに

自律ロボットの応用が進んでいる．現状の自律ロボットは，与えられた命令を忠実に実行することで人間のタスク遂行を補助するツールである．一方で，より高度な意思決定をする自律ロボットの場合，命令を忠実に実行することが必ずしも最善の方策であるとは限らない．このような自律ロボットがユーザーからの信頼を獲得し社会で活躍するには，行動決定の理由を説明する能力が必須となる．我々は，このような説明能力を持つロボットを XAR (eXplainable Autonomous Robots) と定義し，その要件として以下の 4 点を挙げた [1].

(1) 解釈可能な世界モデルの保持
(2) ユーザーの持つプランニングアルゴリズム (または行動方策)，世界モデルの推定
(3) 方策の伝達に有用な情報の抽出
(4) ユーザーへの説明提示

この説明の過程は，実際にはロボットとユーザー相互になされるものであり，図 1 に示すようなコミュニケーションの過程そのものである．なお世界モデルとは，行動と状態変化の対応関係，すなわち環境のダイナミクスを表す内部モデル [2] を指し，今後世界モデルと環境は特に区別せず用いる．また，図 1 の Policy Sharing は自身の方策に対する自発的な情報提示であり，その情報提示に対して他者からクエリを与えられたとき，説明を生成する．

上述の 4 つの要件の中でも特に，他者世界モデルの推定は，ユーザーに応じた説明をするために重要である．人 – ロボットインタラクションの文脈において，ユーザーの内部状態を推定することの重要性は既に認識されており，Gao ら [3] や Clair ら [4] はユーザーの取った行動やインタラクションの履歴から尤もらしい行動方策を推定する枠組みを提案した．また，Huang ら [5] は，説明を方策に復元する過程に注目し，説明を受け取るユーザーが持つ復元アルゴリズムを適切に推定することの重要性を主張した．これらの研究は，内部状態の中でも特に行動方策とプランニングアルゴリズムに注目したものである．しかし一般に，実世界で動くロボットは人間が求める挙動を示すようにアルゴリズムや行動方策を設計され，最終的な目標もユーザーと共有している．そのような状況下での行動決定結果に対する疑問は，環境認識の齟齬に端を発するものが多い．

そこで本稿では，ロボット自身が持つ世界モデルとユーザーが与えた質問 (クエリ) から，ユーザーの持つ世界モデルを推定する手法を提案する．提案手法により世界モデルの差異が判明することで，ユーザーとの環境認識の齟齬を解消する説明が生成可能になる [6]. 本稿では，図 1 の仕組みを図 2 のように簡単化し，説明者が被説明者の世界モデルを他者世界モデルとして推定する問題に焦点を当てる．

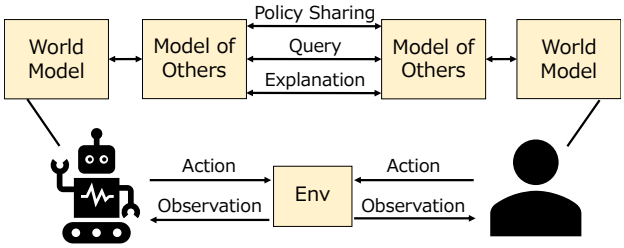

図 1 コミュニケーションとしての説明の概略図．説明の要素を明確にするために，環境の観測と他者世界モデル同士の相互作用を分割している．

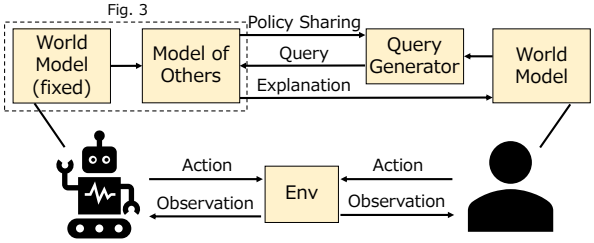

図 2 本稿で考える簡略化された説明プロセス．ロボットからユーザーへの説明プロセスのみを考える．

2. 提案手法

提案手法では，以下の手順でユーザーの持つ世界モデルを推定する．

(1) **世界モデルの分散表現獲得：** グラフ化された世界モデルを用いて，各世界モデルの分散表現を獲得する．
(2) **クエリベクトルの獲得：**ユーザーが与えたクエリを基に，世界モデルの表現空間上でクエリの意味を表現する方向ベクトルを獲得する．
(3) **他者世界モデルの推定：**世界モデルの分散表現とクエリベクトルを用いて，コサイン類似度によりユーザーの持つ世界モデルを推定する．

なお，ロボットとユーザーは状態空間，方策を共有していると仮定し，状態間の接続関係のみ非共有とする．また，本稿では説明の生成は対象としない．

2.1 世界モデルの分散表現獲得

ロボット自身の経験に基づき，環境の類似度を表す表現空間を学習する．初めに，ロボットは強化学習による方策獲得と同時に各環境の状態遷移を表す無向グラフを獲得する．具体的には，方策学習時の探索で遷

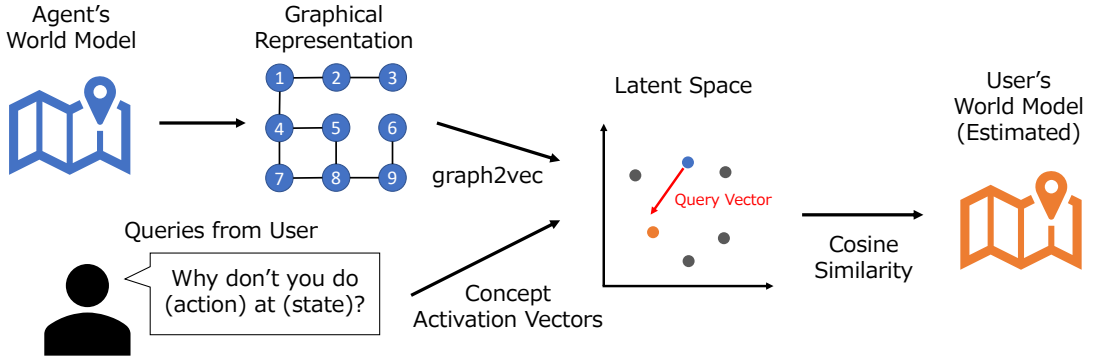

図 3　提案手法の概要図

移が観測された状態間は隣接しているとし，エッジを追加する[1]．そして全ての環境の無向グラフを獲得したのち，graph2vec [8] を適用することで各環境グラフの分散表現を得る．Graph2vec は doc2vec [9] をグラフの分散表現獲得に応用した手法であり，単語の生起に代わり，各部分グラフを表すラベルの生起を予測対象とする．なお，部分グラフを表すラベルは Weisfeiler-Lehman relabeling process [10] を用いて得る．このプロセスは隣接するノードのラベルを考慮して次のレイヤーのラベルを決定するもので，高いレイヤーのラベルほどグラフの大域的な情報を表す．Graph2vec を用いることで，同じ部分グラフが生起するグラフ，すなわち似た状態遷移構造を持った環境のグラフほど表現空間上で近くに埋め込まれる．予め世界モデルの表現空間を獲得しておくことで，ユーザーからのクエリを基にした効率的な探索が可能になる．

なお，本稿では世界モデル獲得時，どの環境の経験であるかを表す環境ラベルを明示的に与える．環境ラベルが与えられない場合には，時間的に連続した経験を用いて世界モデルを得る．さらに，近い分散表現を持つモデルは同一環境を表すとみなし，複数のモデルを併合することでより精度の高い世界モデルを獲得することも有効であると考えられる．

2.2　クエリベクトルの獲得

ユーザーが与えたクエリを基に，世界モデルの表現空間上でクエリの意味を表現する方向ベクトルを獲得する．Kim ら [11] はニューラルネットワークの中間層に注目し，ある概念を満たす特徴量とそうでない特徴量が入力されたときの潜在表現の差を計算することで概念ベクトル (CAV：Concept Activation Vectors) を生成した．本研究ではこの手法を応用し，クエリを満たす環境とそうでない環境の分散表現の差を取ることで，式 (1) としてクエリベクトルを定義する．なお，クエリは「状態 s_{query} では行動 a_{query} を選択すべきだ」という形式を仮定する．

$$\boldsymbol{v}_{pos} = \frac{\sum_i \boldsymbol{v}_i \cdot P(a_{query}|\boldsymbol{v}_i, s_{query})}{\sum_i P(a_{query}|\boldsymbol{v}_i, s_{query})}$$

$$\boldsymbol{v}_{neg} = \frac{\sum_i \boldsymbol{v}_i \cdot (1 - P(a_{query}|\boldsymbol{v}_i, s_{query}))}{\sum_i (1 - P(a_{query}|\boldsymbol{v}_i, s_{query}))}$$

$$\boldsymbol{v}_{cav} = \boldsymbol{v}_{pos} - \boldsymbol{v}_{neg} \tag{1}$$

ここで，$\boldsymbol{v}_i$ は各環境の分散表現である．また，確率的に行動が選択される方策に対応するため，各分散表現 $\boldsymbol{v}_i$ の係数には行動の選択確率値を用いる．必要に応じて係数を $\{0, 1\}$ の二値で表現することも可能である．

本稿では世界モデル推定のためにクエリベクトルを導入したが，ユーザーの属性や過去のインタラクション履歴に基づいたユーザーベクトルを考えることもできる．その場合，graph2vec で獲得した表現空間が CAV の生成に適しているとは限らず，新たな表現空間への写像といった工夫が必要である．

2.3　他者世界モデルの推定

世界モデルの分散表現とクエリベクトルを用いて，コサイン類似度により他者世界モデルを推定する．環境 i の推定環境としての尤もらしさは式 (2) で表される．

$$S(\boldsymbol{v}_i, \boldsymbol{v}_{obs}) = \sum_j similarity(\boldsymbol{v}^j_{cav}, \boldsymbol{v}_i - \boldsymbol{v}_{obs}) \tag{2}$$

ここで，$\boldsymbol{v}_{obs}$ はエージェントが現在観測している環境の分散表現であり，$\boldsymbol{v}^j_{cav}$ は任意の数のクエリベクトルである．また，$similarity(\boldsymbol{a}, \boldsymbol{b})$ はベクトル $\boldsymbol{a}, \boldsymbol{b}$ のコサイン類似度 $[-1, 1]$ を出力する関数である．

実環境で推論する際は，ロボットとユーザーはほぼ同一の環境を観測している．したがって，両者の持つ世界モデルは類似していると考えられるため，$\boldsymbol{v}_{obs}$ からみた各環境の方向とクエリベクトルの方向の類似度を推定基準に用いる．以上の定義を用いると，求める他者世界モデルは式 (3) で表される．

$$Env_{est} = \arg\max_i S(\boldsymbol{v}_i, \boldsymbol{v}_{obs}) \tag{3}$$

本稿ではすべてのクエリの重要度が等価であると仮定し，各クエリベクトル $\boldsymbol{v}^j_{cav}$ に対する類似度の総和で評価関数 S を設計した．しかし，実世界ではそれぞれのクエリの重要度が異なることも考えられ，その場合には類似度に係数 λ^j を乗じる必要がある．また，式 (2) に表現空間上の距離に関する評価項や分散表現そのもののコサイン類似度の項を加えることで，ロボットの観測環境との類似度を明示的に考慮することもできる．

[1]本手法は離散状態空間での利用を仮定しており，実ロボットに適用する場合には状態空間を離散化する必要があるが，その方法については本研究の対象外である．離散化には，各状態が方策上持つ意味を考慮した手法 [7] の利用が考えられる．

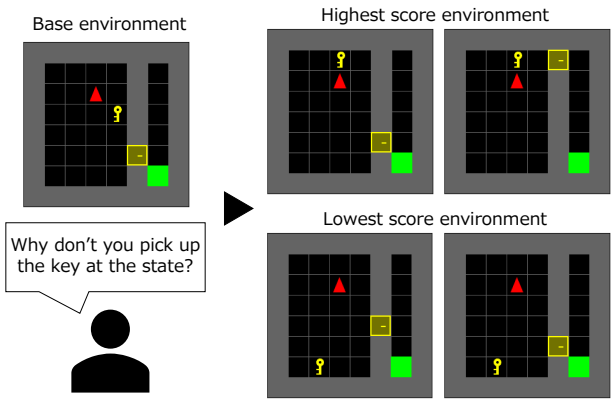

図 4　他者世界モデルの推定結果

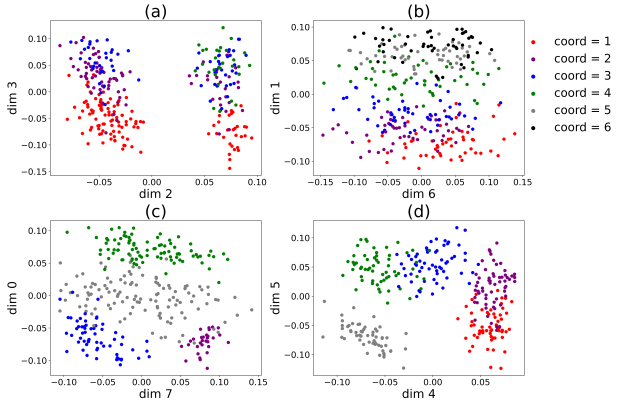

図 5 表現空間の可視化結果．(a) 鍵の X 座標 (b) 鍵の Y 座標 (c) ドアの X 座標 (d) ドアの Y 座標によりデータを色分けしている．

3.　実験

シミュレーション環境において，行動方策を PPO (Proximal Policy Optimization) で獲得するエージェントに対して提案手法を適用し，有用性を評価する．実験には，複数のオブジェクトが配置されたグリッド環境 [12] を一部改変し用いた (図 4)．この環境では，エージェント (赤色) は鍵をとり，黄色のドアを開け，緑色のゴールに到達することで報酬を得る．報酬の最大値は 1 であり，獲得に要した行動回数によって減衰する．また．ゴールの位置は不変であるが，鍵とドアの位置は毎試行変化する．エージェントの行動は直進，左を向く，右を向く，前方のグリッドにある鍵をとる，ドアを開けるの 5 種類である．またエージェントの観測は，鍵の絶対位置 $(x, y$ 座標$)$，ドアの絶対位置 $(x, y$ 座標$)$，自身の絶対位置 $(x, y$ 座標$)$ と向き，鍵の保持/不保持，ドアの開/閉の計 9 次元である．

3.1　実験 1：世界モデルの分散表現獲得結果

PPO による行動方策の学習と同時に各環境の状態遷移を表すグラフを獲得し，graph2vec で 16 次元の分散表現を得た．表現空間を ICA (Independent component analysis) を用いて 8 次元に圧縮し，可視化した結果を図 5 に示す．なお，本実験では鍵とドアの絶対位置を環境ラベルとして用い，ノードの特徴量にはそれらを除く 5 次元の観測を用いた．実験の結果，鍵，ドアの絶対位置によるクラスタが表現空間上で形成されていることがわかる．特にドアの位置に関するクラスタが明確に形成されているが，これは鍵と比較して周囲の状態遷移関係を大幅に変化させるためである．本実験では，鍵やドアの絶対位置は更新対象となる環境グラフを識別する目的でのみ利用しており，グラフそのものには埋め込まれていない．したがって graph2vec により，状態遷移構造に表出した鍵やドアの位置情報による差異を適切に反映した表現空間が形成されたといえる．

3.2　実験 2：他者世界モデルの推定結果

得られた表現空間にクエリベクトルを適用し，他者世界モデルを推定する．本検証では鍵を取得する場面，ドアを開ける場面での質問を想定し，「状態 s_{query} では { 鍵をとる/ドアを開ける } べきだ」というクエリを考える．基準となる世界モデルとクエリを与えた時の，他者世界モデルの推定結果を図 4 に示す．クエリを満たす環境の評価値が最も高く，グリッド環境上で鍵の位置が対極にある環境の評価値が最も低くなった．この結果から，得られたクエリベクトルが世界モデルの推定に適当であることが示唆される．

表 1 最適環境出現順の累積度数と平均値．試行回数は各 100 回である．

Method \ Order	1	2	3	Order Average
Our method	35	49	60	7.1
Random	6	8	11	23.1

また，エージェントの世界モデルにクエリを満たすよう最小限の変更を加えた世界モデルを，最適環境と定義する．例えば，「状態 s_{query} では鍵をとるべきだ」というクエリが与えられたときには，ドアの位置はそのままに，鍵の位置のみクエリを満たすよう変更されたものが最適環境となる．無作為に選んだエージェントの世界モデルとクエリのペアに対し，式 (2) を用いて得られた各環境の評価値を降順にソートし，最適環境の出現順を得た (表 1)．なお，比較対象として，クエリを満たす環境をランダムにソートした場合の出現順も併記している．実験の結果，提案手法により最適環境が上位にソートされることが確認された．変更すべき状態遷移関係を明示的に与えられていないにもかかわらず，提案手法は比較的早い段階で最適環境を推定していることが分かる．

3.3　実験 3：複数クエリによる検証

説明エージェント A と被説明エージェント B を用意し，A が B の世界モデルを正しく推定するために必要なクエリの数を評価する．実験の概要は以下の通りである．

(1) A，B は初期状態 (エージェントの絶対位置，向き，鍵とドアの状態) と，どの環境でどの方策を使うかという環境・方策の対を共有する．また，鍵とドアの位置が異なる任意の世界モデルをそれぞれ

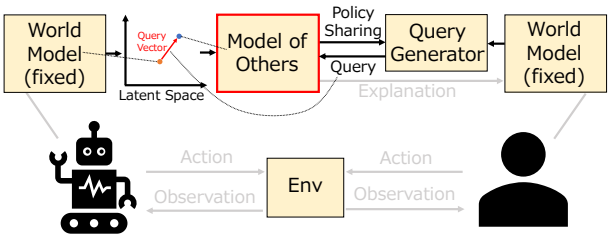

図 6 実験 3 の概略図．エージェントは自身の行動方策を共有し，ユーザーから与えられたクエリを基に他者世界モデルを更新する．

表 2　他者世界モデルの推定に要した更新回数．

Method	Number of updates	Standard deviation
Our method	**5.53**	4.83
AND search 1	20.72	17.43
AND search 2	8.69	4.71

持つ．

(2) A は，自身の行動系列を順に提示する（Policy Sharing）．

(3) B は，自身の方策における最適行動と与えられた行動が異なるとき，「状態 s_{query} では行動 a_{query} を選択すべきだ」というクエリを追加する．既存のクエリは削除しない．

(4) A はクエリに基づき他者世界モデルを更新し，s_{query} を初期状態として再び更新した他者世界モデルにおける行動系列を順に提示する．

(5) (3)，(4) を繰り返し，A が B の持つ世界モデルを他者世界モデルとして得るのに要した環境更新回数を評価する．

なお，一度選択した環境は選択せず，2 番目以降の候補を採用する．また，行動系列を全て提示した段階で同一の環境が得られていない場合は，クエリを増やさず環境の更新を続ける．本検証では他者世界モデルの更新方法として，提案手法とクエリの AND 検索を比較する．

提案手法 式 (3) を用いて尤もらしい環境を選択する．

AND 検索 1 クエリを満たす環境の中から，ランダムに環境を選択する．

AND 検索 2 「初期状態から状態 s_{query} に到達するまでのすべての状態で，B の方策における最適行動が選択されること」を制約条件に，ランダムに環境を選択する．実質的には前述の 2 つの更新方法と比較して制約条件が追加され，与えられる情報が増加する．

実験の結果，提案手法が最も少ない更新回数で他者世界モデルを推定できることが分かった (表 2)．また，対応あり両側 t 検定の結果，提案手法と AND 検索 1 では $t(100) = 8.07$，$p < .01$，提案手法と AND 検索 2 では $t(100) = 4.59$，$p < .01$ となり，共に有意差が確認された．直接的に与えられる情報は「AND 検索 2」が最も多いが，提案手法ではクエリをベクトル化することで，表現空間に埋め込まれた事前知識を付加的な情報として活用し，更新回数を削減できたと考えられる．

4. おわりに

本稿では XAR の実現に向け，ロボット自身が持つ世界モデルとユーザーが与えたクエリから，ユーザーの保持する世界モデルを推定する手法を提案した．そして，提案手法により，クエリの AND 検索に比べ効率的に他者世界モデルが推定可能であることを示した．今後の課題として，ユーザーベクトルの導入や，世界モデルの差異を用いた説明生成手法の検討などが挙げられる．

謝　辞　この成果は，国立研究開発法人新エネルギー・産業技術総合開発機構 (NEDO) の委託業務の結果得られたものである．

参　考　文　献

[1] Tatsuya Sakai and Takayuki Nagai. Explainable Autonomous Robots: A Survey and Perspective. *ArXiv*, Vol. abs/2105.02658, , 2021.

[2] Ha, David and Schmidhuber, J¨urgen. Recurrent World Models Facilitate Policy Evolution. In *Advances in Neural Information Processing Systems 31*, pp. 2450–2462. Curran Associates, Inc., 2018.

[3] Xiaofeng Gao, Ran Gong, Yizhou Zhao, Shu Wang, Tianmin Shu, and Song-Chun Zhu. Joint Mind Modeling for Explanation Generation in Complex Human-Robot Collaborative Tasks. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 1119–1126. IEEE.

[4] A. S. Clair and M. Matarić. How Robot Verbal Feedback Can Improve Team Performance in Human-Robot Task Collaborations. In *2015 10th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, pp. 213–220, 2015.

[5] Sandy H. Huang, David Held, Pieter Abbeel, and Anca D. Dragan. Enabling Robots to Communicate Their Objectives. *ArXiv*, Vol. abs/1702.03465, , 2017.

[6] Tatsuya Sakai, Kazuki Miyazawa, Takato Horii, and Takayuki Nagai. A framework of explanation generation toward reliable autonomous robots. *arXiv preprint arXiv:2105.02670*, 2021.

[7] Lunjun Zhang, Gengcong Yang, and Bradly C. Stadie. World Model as a Graph: Learning Latent Landmarks for Planning. *ArXiv*, Vol. abs/2011.12491, , 2020.

[8] A. Narayanan, Mahinthan Chandramohan, R. Venkatesan, Lihui Chen, Y. Liu, and Shantanu Jaiswal. graph2vec: Learning distributed representations of graphs. *ArXiv*, Vol. abs/1707.05005, , 2017.

[9] Quoc Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, Vol. 32 of *Proceedings of Machine Learning Research*, pp. 1188–1196, Bejing, China, 22–24 Jun 2014. PMLR.

[10] Nino Shervashidze, Pascal Schweitzer, Erik Jan van Leeuwen, Kurt Mehlhorn, and Karsten M. Borgwardt. Weisfeiler-Lehman Graph Kernels. *Journal of Machine Learning Research*, Vol. 12, No. 77, pp. 2539–2561, 2011.

[11] Been Kim, M. Wattenberg, J. Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and R. Sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *ICML*, 2018.

[12] Maxime Chevalier-Boisvert, Lucas Willems, and Suman Pal. Minimalistic Gridworld Environment for OpenAI Gym, 2018.