現場環境で学習した知識に基づく曖昧な発話からの生活物理支援タスク

○萩原 良信　長谷川 翔一　大山 瑛　谷口 彰　エル ハフィ ロトフィ　谷口 忠大（立命館大）

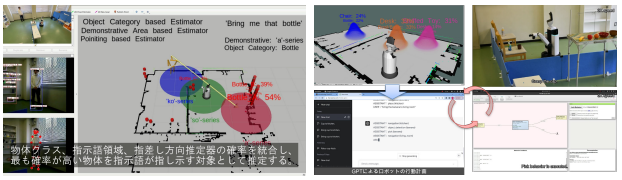
家庭環境では,「あれ取って」や「コップを持ってきて」といった曖昧な言語指示が用いられる．これらの言語指示には，持ってくる物体や取ってくる場所の情報が明示的に含まれていない．本稿では，ロボットが現場環境で学習した知識に基づいて不足している情報を補い，曖昧な言語指示から生活物理支援タスクを実現する二つの手法について述べる．一つは，現場のマルチモーダル情報を用いた指示語を含む言語指示の外部照応解析の手法である．もう一つは，場所概念モデルにより獲得された現場知識と大規模言語モデルを活用したプランニングの手法である．

1. はじめに

家庭環境において，ユーザからの言語指示に基づいてモノを持ってくるなどの生活物理支援を実現するロボットの研究開発が推進されている [1, 2]．家庭用の生活支援ロボットの研究開発において，ロボットが「あれ取って」や「コップを持ってきて」などの曖昧な言語指示を理解し，具体的な生活物理支援タスクを実現する事は重要な課題である [3]．これらの言語指示には，取ってくる物体や取ってくる場所の情報が明示的に含まれておらず，タスクを実行するために情報を補う必要がある．このような課題に対して，著者らのグループは，ロボットが現場環境で学習した知識を用いて，言語指示において不足している情報を補い，具体的な生活物理支援タスクを実現する二つの手法を開発した [4, 5]．本稿では，これらの手法の概要について述べ，生活物理支援タスクの実行例を紹介する．

一つ目の手法は,「あれ」や「それ」などの指示語が指し示す対象を現場のマルチモーダル情報に基づいて推定する外部照応解析の手法である [4]．外部照応解析とは，現場の環境から照応詞の指示対象を推定する事である．これに対して，照応詞の前後の文章から指示対象となる語を推定するのが内部照応解析であり，自然言語処理の分野において研究されてきた [6]．しかし,「あれ取って」などの指示語を含んだ言語指示をロボットが理解するには，外部照応解析が重要になる [7]．そこで，著者らのグループは，物体，指示語，指差しの情報を用いた外部照応解析により，指示語の指示対象における曖昧性を解消する手法を開発した（図 1a）．

二つ目の手法は,「コップを持ってきて」などの言語指示において，現場環境の知識に基づいて探索する場所を推論し，行動を生成するプランニングの手法である [5]．このような言語指示から探索する場所を推論するには，物体がどの場所で観測されるかといった知識が重要になる．従来，物体ラベルを占有格子地図のグリッドに付与する Semantic mapping [8] や，マルチモーダル情報（言語，位置，画像，物体）から場所のカテゴリと領域を推定する場所概念モデル [9] の研究が報告されている．これらの研究は現場の知識が獲得できる一方，現場で観測されなかった物体の予測は困難である．そこで，大規模言語モデルをロボットの行動計画に応用した研究 [1, 2] が報告されている．大規模言語モデルは，分布意味論を介した単語の意味理解の汎化を実現しており，未知語に強い性質を持っている．現場環境で得られた知識を大規模言語モデルに対してプロンプトとして与える事で現場に応じたプランニングが可能になるが，現場知識を獲得し言語化する作業コストが課題となる．そこで，著者らのグループは，場所概念モデルによりロボットが獲得した現場における場所と物体の知識を大規模言語モデル [10] にプロンプトとして与え，現場知識と一般常識に基づく行動のプランニングを実現する手法を開発した（図 1b）．



(a) 外部照応解析手法　　(b) プランニング手法

図 1: 現場知識に基づく曖昧な発話からの生活物理支援

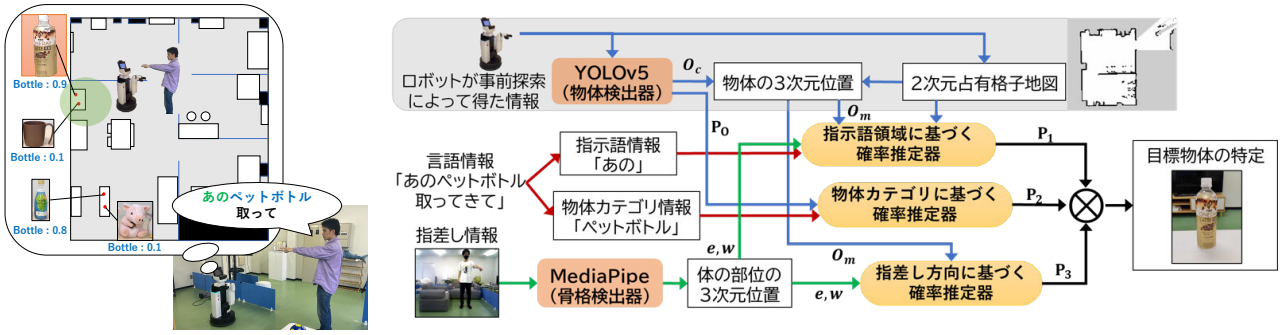
2 章は指示語を含む言語指示の外部照応解析，3 章は場所概念モデルと大規模言語モデルを活用したプランニング，最後に，まとめと今後の展望について述べる．本稿は，HSR コミュニティにおける立命館大学サイトの活動を代表する成果について報告するものである．

2. 指示語を含む言語指示の外部照応解析

2.1 概要

図 2a に指示語を含む言語指示の外部照応解析の概要を示す．例えば,「あのペットボトル取って」という発話がされた場合,「ペットボトル」を物体カテゴリ情報,「あの」を指示語情報とする．ロボットは，事前に環境を探索し，地図上における物体の 3 次元位置の情報を持っているとする．ここで，ロボットは,「あの」が意味する指示語領域（緑色の円）を推定し,「ペットボトル」に対応する物体カテゴリの確率（青色の数値）を全ての候補物体について計算する．さらに，ユーザの指差し方向から計算される確率分布を用いて確率的に指示対象の物体を推定する．

図 2b に開発した手法の処理の流れを示す．青色矢印



(a) 概要図　　(b) 処理の流れ

図 2: 指示語を含む言語指示の外部照応解析手法 [4]

は，ロボットが家庭環境の事前探索により収集した情報，赤色矢印は，ユーザの言語指示から得た言語情報，緑色矢印は，ロボットのカメラから得た視覚情報の流れを示す．$\boldsymbol{O_c}$ と $\boldsymbol{O_m}$ は，それぞれカメラ座標系とマップ座標系における物体の 3 次元位置情報，$\boldsymbol{P_O}$ は，各物体の信頼度スコア，$\boldsymbol{e}$ と $\boldsymbol{w}$ は，それぞれユーザの目と手首の 3 次元位置，$\boldsymbol{P_1}, \boldsymbol{P_2}, \boldsymbol{P_3}$ は，各推定器から出力される確率分布である．この手法は，指示語，物体カテゴリ，指差しの 3 種類の外部照応情報を解析し，ユーザが指示した対象物体を推定する．各推定器は，環境内で観測された全ての候補物体についてその物体が指示対象である確率（以後，対象確率）を計算する．ロボットは，現場環境で事前に得た地図上における候補物体の 3 次元位置情報，ユーザの指示において得られた指示語情報，物体カテゴリ情報，指差し情報を用いた推定器から対象確率を計算し，この乗算により指示対象の物体を推定する．

2.2　指示語領域に基づく推定器

指示語系列には異なる性質があり，コ系列は話し手から近い物体や人を参照する場合，ソ系列は聞き手から近い物体や人を参照する場合，ア系列は話し手からも聞き手からも遠い物体や人を参照する場合にそれぞれ用いられる．この推定器では，各系列の性質に基づいた 3 次元ガウス分布によって指示語領域を表す．まず，ユーザやロボットの位置，指差しの方向を用いてガウス分布のパラメータを決定する．次に，このガウス分布に各候補物体の 3 次元位置を入力する．最後に，得られた確率を対象確率 (P_1) として出力する．各系列におけるガウス分布のパラメータの設計指針は以下である．

- コ系列の平均：ユーザに近い手首の座標
- ソ系列の平均：ロボットに近いユーザとロボットの内分点の座標
- ア系列の平均：ユーザから遠い指差しベクトルの延長線上の座標
- 各系列の分散：ユーザまたはロボットの位置と指示語領域の平均の座標との距離に比例する値

2.3　物体カテゴリに基づく推定器

物体カテゴリに基づく確率推定器は，物体カテゴリ情報とロボットが事前に探索して得た候補物体のカテゴリ確率を入力とし，対象確率を出力する．対象確率を予測するために Objects365 [11] で事前学習済みの You Only Look Once version 5 (YOLOv5) [12] を物体検出器として用いる．YOLOv5 を用いて事前に環境中の候補物体を検出し，各物体が属するカテゴリの信頼度スコアを計算する．言語指示から抽出した物体カテゴリ情報に対応する各物体の信頼度スコアを正規化したものを対象確率 (P_2) として出力する．

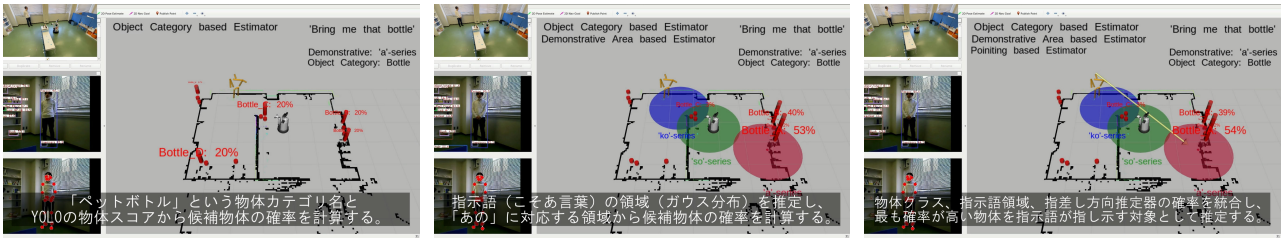
2.4　指差し方向に基づく推定器

指差し方向に基づく確率推定器は，骨格検出により得たユーザの身体部位の 3 次元位置とロボットが事前に環境内を探索して得た候補物体の 3 次元位置を入力とし，対象確率を出力する．指差し動作の骨格検出には MediaPipe [13] を用いる．ユーザの目を始点，手首を終点とする指差しベクトルとユーザの目から各候補物体へのベクトルの内積から 2 つのベクトルのなす角 θ を得る．この θ を用いた 2 次元フォンミーゼス分布から得られる確率を対象確率 (P_3) として出力する．

2.5　生活物理支援タスク

開発した手法をロボット（トヨタ自動車社製 HSR）に実装し，「あのペットボトルを持ってきて」という曖昧な言語指示からの生活物理支援タスクを行った [1]．図 3 に同タスクにおけるロボットによる対象物体の推定過程と結果を示す．(a) は，物体カテゴリに基づく確率推定の結果を示している．環境で観測された全ての候補物体について，「ペットボトル」という物体カテゴリ名に基づいて計算された対象物体の確率が表示されている．この推定器のみでは，対象物体を 1 つに絞り込む事は難しい．(b) は，次に実行された指示語領域に基づく確率推定の結果を示している．こ，そ，あ系列に対応する指示語領域がそれぞれ青，緑，赤の円で可視化されている．あ系列の領域のガウス分布に基づいて計算された対象物体の確率が示されている．ここまでの推定により，高い確率を持つ二つの候補物体が得られた．(c) は，その後に実行された指差し方向に基づく確率推定の結果を示している．指差し方向は，黄色いベクトルで可視化されている．このベクトルから指差し方向の確率分布を仮定し，これに基づいて計算された候補物体の確率が示されている．最終的に得られた候

[1] https://youtu.be/mUZ5EdQgdSw



(a) 物体カテゴリ推定器の実行　(b) 指示語領域推定器の実行　(c) 指差し方向推定器の実行

図 3: 対象物体の推定過程と結果

補物体の確率において，最も高い確率を持つ Bottle_A が対象物体として推定された．この結果に基づいてロボットは対象物体の近くまで移動し，物体を検出する事によりタスクを達成する事が出来る．

3. 場所概念モデルと大規模言語モデルを活用したプランニング

3.1 概要

ロボットが「コップを持ってきて」という言語指示を受けた場合，対象の物体がどこにあるかを予測する必要がある．物体のカテゴリ名などが分かっている場合，ChatGPT [2]などの大規模言語モデルを活用して物体がありそうな場所を推論させ，ロボットの行動列を生成させる方法がある [2]．このとき，現場における物体の配置情報を場所概念モデル [9] により獲得する事により，一般常識のみならず現場の知識を活用した行動列の生成が期待される．図 4 に構築した手法の概要を示す．この手法は，場所概念モデルに基づく現場知識の獲得と記述，ChatGPT による行動計画，FlexBE [14] による逐次的な行動実行から構成される．まず，場所概念モデルにより獲得された現場知識，ロボットのスキルセット，ユーザの言語指示をプロンプトとして ChatGPT に入力する．次に，ChatGPT は入力されたプロンプトにおいて，言語指示を達成するロボットの行動を生成する．最後に，プランニングエンジンの FlexBE が生成された行動を逐次的に実行し，ロボットによる動作の成功と失敗を受けて次の行動を ChatGPT から受け取る．この繰り返しにより，現場知識と一般常識に基づくプランニングによる生活物理支援が実現できる．

3.2 場所概念モデルによる現場知識獲得

場所概念モデルは，位置，言語，画像，物体のマルチモーダル情報を観測とし，潜在変数である場所カテゴリと領域のインデックスを推定する確率的生成モデルである．このモデルにおける確率分布のパラメータ推定により，物体情報が与えられた元での各場所における観測確率 (P(場所 | 物体情報)) を計算する事が出来る．この確率を現場における物体の配置情報に関する知識として ChatGPT のプロンプトに記述する．場所概念モデルの生成過程やパラメータの詳細は，文献 [5] を参照されたい．

[2]https://openai.com/blog/chatgpt/

3.3 ChatGPT による行動計画

場所概念モデルにより現場環境で獲得した，各場所名における物体名（ラベル）の観測確率，観測された場所の名前，観測された物体名（ラベル），そして，ユーザからの言語指示（例：Bring a cup to the kitchen），ロボットのスキルセットをプロンプトに記述し，ChatGPT に言語指示を達成する行動を生成させる．

3.4 FlexBE による行動実行

ChatGPT が生成した行動を FlexBE を用いて逐次的に実行する．実世界では，ロボットが実行した動作が必ずしも成功するとは限らない．このため，動作の成功と失敗を FlexBE を通して ChatGPT に入力し，結果に応じた次の行動を逐次的に生成させる．これにより，不確実な実世界における生活物理支援タスクを達成する．

3.5 生活物理支援タスク

構築した手法をロボットに実装し，「コップをキッチンに持っていって」という曖昧な言語指示からの生活物理支援タスクを行った．図 5 に同タスクにおけるロボットによる現場知識の獲得とプランニングの例を示す．(a) は，場所概念モデルにより獲得された現場の各場所における物体の観測確率を示している．(b) は，ChatGPT に与えたプロンプトの一部を示している．各場所における物体の観測確率などの現場知識とユーザからの言語指示，ロボットのスキルセットを入力して，行動を生成させる．(c) は，生成された行動に対応するノードを FlexBE が実行する様子である．ノードの実行により HSR が動作し，移動，物体検出，把持などの動作を逐次的に遂行する．このとき，動作の結果（True or False）は，FlexBE を介して ChatGPT に入力され，結果に応じて次の行動が生成される．

4. おわりに

ロボットが現場環境で学習した知識を用いて，不足している情報を補い，生活物理支援タスクを実現する二つの手法について述べ，生活物理支援タスクの実行例を紹介した．これらの手法の有効性や限界，他手法との定量的な比較評価については，文献 [4, 5] を参照されたい．今後の展望として，これらのモデルを確率的に統合したモデルの開発により，「それ片付けて」などの指示語を含み現場知識を必要とする言語指示に対して柔軟な理解を可能とする手法の実現を目指したい．

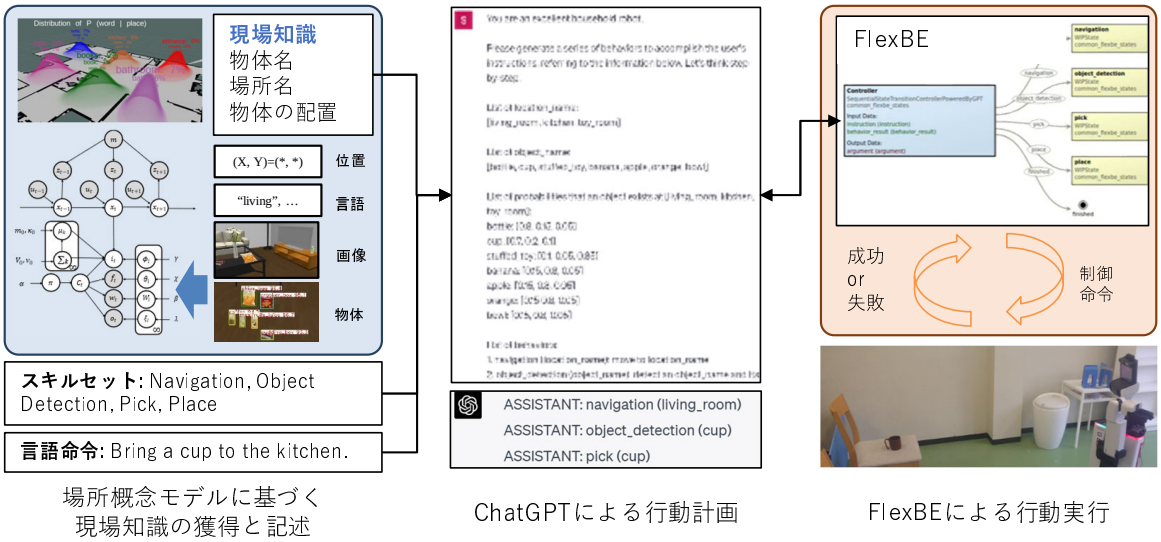
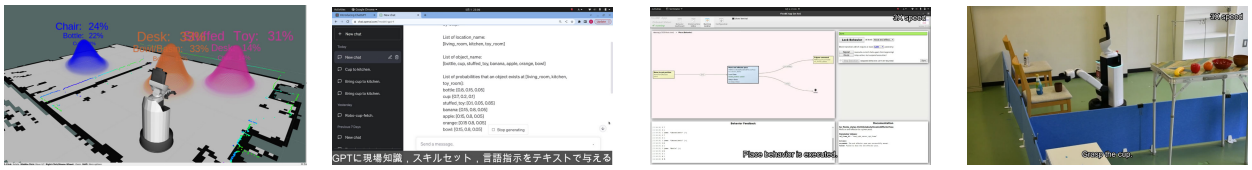


図 4: 場所概念モデルと大規模言語モデルを活用したプランニング手法 [5]



(a) 現場知識の獲得　(b) ChatGPT による行動計画　(c) FlexBE による行動実行　(d) HSR によるタスク実施

図 5: ロボットによる現場知識の獲得とプランニングの例

謝　辞　本研究は，JST ムーンショット型研究開発事業 JPMJMS2011，JSPS 科研費 JP22K12212 の助成を受けたものである．

参　考　文　献

[1] B. Chen, F. Xia, B. Ichter, K. Rao, K. Gopalakrishnan, M. S. Ryoo, A. Stone, and D. Kappler, “Open-Vocabulary Queryable Scene Representations for Real World Planning,” *arXiv preprint arXiv:2209.09874*, 2022.

[2] S. Vemprala, R. Bonatti, A. Bucker, and A. Kapoor, “ChatGPT for Robotics: Design Principles and Model Abilities,” Microsoft, Tech. Rep., 2023.

[3] T. Taniguchi *et al.*, “Survey on frontiers of language and robotics,” *Advanced Robotics*, vol. 33, no. 15-16, pp. 700–730, 2019.

[4] 大山瑛, 長谷川翔一, 中川光, 谷口彰, 萩原良信, and 谷口忠大, “実世界のマルチモーダル情報に基づく指示語を含んだ言語指示の外部照応解析,” in **言語処理学会年次大会**, 2023, pp. 2296–3001.

[5] 長谷川翔一, 伊藤昌樹, 山木良輔, 坂口太一, 萩原良信, 谷口彰, エルハフィロトフィ, and 谷口忠大, “生活支援ロボットの行動計画のための大規模言語モデルと場所概念モデルの活用,” in **日本ロボット学会学術講演会**, 2023.

[6] 飯田龍, 乾健太郎, and 松本裕治, “文脈的手がかりを考慮した機械学習による日本語ゼロ代名詞の先行詞同定,” **情報処理学会論文誌**, vol. 45, no. 3, pp. 906–918, 2004.

[7] 杉山治, 神田崇行, 今井倫太, 石黒浩, 萩田紀博, and 安西祐一郎, “コミュニケーションロボットのための指さしと指示語を用いた 3 段階注意誘導モデル,” **日本ロボット学会誌**, vol. 24, no. 8, pp. 964–975, 2006.

[8] S. Garg, N. Sünderhauf, F. Dayoub, D. Morrison, A. Cosgun, G. Carneiro, Q. Wu, T.-J. Chin, I. Reid, S. Gould, P. Corke, and M. Milford, “Semantics for Robotic Mapping, Perception and Interaction: A Survey,” *Foundations and Trends® in Robotics*, vol. 8, no. 1-2, pp. 1–224, 2020.

[9] A. Taniguchi, Y. Hagiwara, T. Taniguchi, and T. Inamura, “Improved and scalable online learning of spatial concepts and language models with mapping,” *Autonomous Robots*, vol. 44, pp. 927–946, 2020.

[10] OpenAI, “GPT-4 Technical Report,” *arXiv preprint arXiv:2303.08774*, 2023.

[11] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, “Objects365: A Large-scale, High-quality Dataset for Object Detection,” in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8430–8439.

[12] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “You Only Look Once: Unified, Real-Time Object Detection,” in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[13] C. Lugaresi, J. Tang, H. Nash, C. McClanahan, E. Uboweja, M. Hays, F. Zhang, C.-L. Chang, M. G. Yong, J. Lee *et al.*, “Mediapipe: A Framework for Building Perception Pipelines,” *arXiv preprint arXiv:1906.08172*, 2019.

[14] P. Schillinger, S. Kohlbrecher, and O. von Stryk, “Human-Robot Collaborative High-Level Control with Application to Rescue Robotics,” in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2796–2802.