大規模言語モデルによるタスク実行管理器生成法と RoboCup JapanOpen @Home League GPSR タスクへの応用

○大日方慶樹 金沢直晃 河原塚健人 矢野倉伊織 金淳暁 岡田慧 稲葉雅幸 (東京大学)

1. 序論

ロボットによる生活支援は古くから期待されている．本研究では，人がロボットにタスクを自然言語で伝え，ロボットが自分の持つ行動関数 (プリミティブ関数) 列にコンパイルし順次実行していくシステムを提案する．

Attention [1] の出現やコンピュータハードウェアの性能向上によって，言語タスクでの高い性能を示す大規模言語モデルが出現した．最近では Web データを大量に学習し，人間の世界の一般的な知識を有する言語モデルが存在する．本研究ではこのモデルでロボットのプリミティブ関数列を生成する．また大規模言語モデルを画像と言語のマルチモーダルタスクに転用する取り組みがあり，これを本研究の認識器に利用する．

本論文に示す提案手法によって，RoboCup Japan Open 2022 @Home の General Purpose Service Task 競技で 1 位になった．

2. 人のコマンドからの動作計画法

2.1 タスクの計画と実行管理法

本研究では オープンボキャブラリーなロボットタスク実行システムを目指し，RoboCup@Home Command Generator [2] のコマンドの実ロボットによる実行を目指す．引数の場所に移動する move_to (ARG)，画像に対して引数の質問に回答する visual_question_answering (ARG)，人に追従する follow，引数の物体をつかむ grasp (ARG)，引数の場所または人へ物体を置くまたは渡す pass_to (ARG)，引数の文章を話す speak (ARG)，人の発話に答える answer (ARG) の 7 つのプリミティブ関数を用意する．

各関数の成功，失敗をプログラム上で定義し，その結果に応じて，成功の場合は次に進む，失敗の場合は初期位置に戻るといった動作を明確に定義するために，言語モデルから得られるロボットのプリミティブ関数は，状態機械ベースのタスク実行管理器 SMACH [3] によって動作が管理される．

本研究では，大規模言語モデルに GPT-3 [4] を用いる．ロボットが人の命令を GPT-3 によって解釈するプロンプトを図 1 に示す．GPT-3 には 7 つのプリミティブ動作があること，人のコマンドからプリミティブ関数を生成する例，次に入力されるコマンドの文章からプリミティブ関数を生成することを依頼する文章，実際のコマンド，を順番に記したプロンプトを入力する．人のコマンドからプリミティブ関数を生成する例には，

Q. Guide Alex from the entrance to the sink

A. [[“move_to”, “the entrance”], [“move_to”, “Alex”], [“speak”, “Please follow me.”], [“move_to”, “the sink”]]

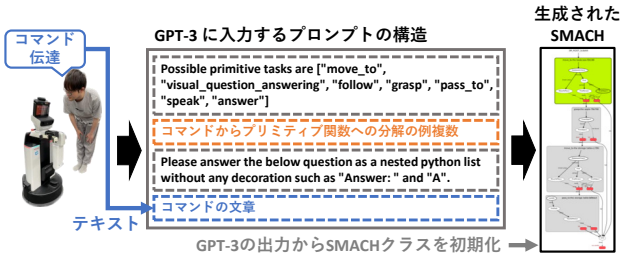

図 1 SMACH 生成のためのプロンプト

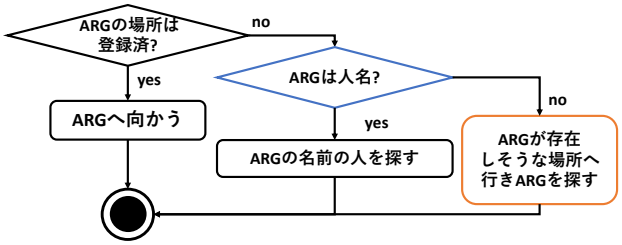

図 2 move_to の引数の種類に応じた動作分岐

Q. Could you Tell me the gender of the person at the dishwasher

A. [[“move_to”, “the dishwasher”], [“visual_question_answering”, “what is the gender of the person ?”]]

のように，あらかじめ人が Command Generator のコマンドを複数個プリミティブ関数列へ変換したものを入力する．GPT-3 の推論結果をパースし，ロボットの SMACH を初期化する．

2.2 プリミティブ関数のあいまいな引数の具体化

ロボットのプリミティブ関数は具体化の必要がある場合がある．例えば move_to 関数は様々な種類の引数をとりうる．この関数では図 2 のように，引数の種類に応じて動作を分岐させる．まず引数の場所が登録されている場合は，そこに直接移動する．もし引数が未登録かつ人名の場合は，人を探す行動に移る．人名か物体名であるかの判定は，引数を${arg} として，

Is the word (ARG) the name of a person or a thing? Please answer with PERSON or THING

というプロンプトを GPT-3 に入力して行う．もし引数が未登録かつ物体名の場合は，それがありそうな既知の場所をリストアップし訪問して探す．物体がありそうな場所のリストアップには，引数を${arg}，登

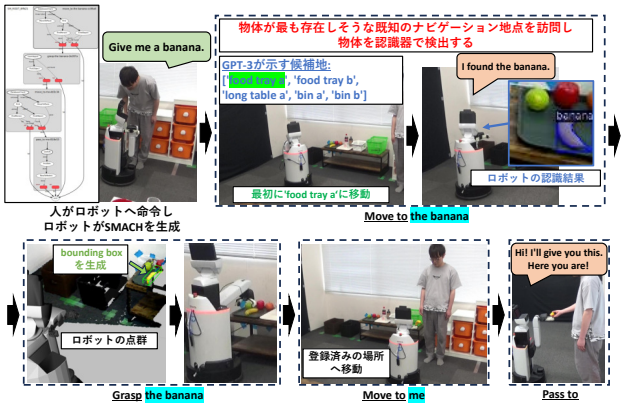

図 3 SMACH 生成とプリミティブ関数の具体化による
タスク実行実験

録済みの場所のリストを${registered place} として，

There are ${registered place} in navigable locations. I was told to look for ${arg}. Where is the most likely place to find ${arg}? Please tell me the ${arg} locations in order of likelihood among the navigable locations. Please answer the below question as a nested python list without any decoration such as “Answer:” and “A”.

というプロンプトを GPT-3 に入力して行う．

2.3 認識器の具体化

各プリミティブ動作の引数に応じて，認識器が何を認識するべきか逐次変更する必要がある．本研究ではオープンボキャブラリーな物体認識モデル Detic [5] を用いる．move_to や grasp 関数の引数が Detic の vocabulary に設定される．

3. 実験

提案システムの実験に HSR [6] を用いる．

3.1 SMACH 生成と move_to 推論を含む実験

図 3 に実験の様子を示す．人がロボットに “Give me a banana” と発話した．ロボットは move_to the banana, grasp the banana，move_to me，pass_to というプリミティブ関数列を生成した．まず move_to the banana について the banana という場所は登録されていないので，ありそうな場所を GPT-3 によって推論し，food tray a へ向かった．同時に Detic の Vocabulary を banana に変更して，banana を検出できるようにした．到着し，banana を発見したので grasp 動作に遷移し，banana の Bounding Box に腕を伸ばして把持した．その後 move_to me (コマンドを伝える人の場所) に移動し，バナナを渡してタスクを完了した．

3.2 RoboCup Japan Open 2022@Home GPSR

2023 年 3 月 7 日に東京大学で行われた RoboCup Japan Open 2022@Home DSPL リーグ GPSR タスクの様子を図 4，競技結果を表 1 に示す．3 種類のタスクが発話され，初期位置に言ってタスクを聞いて実行を 3 回繰り返した．3 回目の途中で時間切れになったが，最後まで完遂できたのは我々のチームのみであった．提案システムによって本種目で 1 位になった．

表 1 RoboCup Japan Open 2022@Home DSPL リーグ GPSR のスコア

順位	チーム名	点数
1	**Team JSK (ours)**	**130**
2	TRAIL	41.25
3	Hibikino-Musashi@Home	25
4	eR@sers	12.5
4	OIT-RITS	12.5
4	あばうたぁ゛ず	12.5
7	SOBITS	0

図 4 2 個目までのタスクの競技の様子．音声合成エンジンの発話を緑，ロボットの発話をオレンジで示している

4. 結論

本研究では大規模言語モデルベースのタスク具体化システムを構成した．提案手法は Coarse-to-Fine なタスクプランニングによってプリミティブ関数列生成，各関数の具体化，認識器の具体化を行い，実験によって本システムがオープンボキャブラリーなタスクを実行できることがわかった．用意するプリミティブ関数，各関数のさらなる具体化，認識器の改良などによって別タスクへの応用や動作精度の向上が見込まれる．

参 考 文 献

[1] A. Vaswani, et al. Attention is all you need. In *Advances in neural information processing systems*, Vol. 30, 2017.

[2] RoboCup@Home Command Generator. `https://github.com/kyordhel/GPSRCmdGen`. [Online; accessed 15-July-2023].

[3] J. Bohren and S. Cousins. The smach high-level executive [ros news]. *Robotics & Automation Magazine*, Vol. 17, No. 4, pp. 18–20, 2010.

[4] T. Brown, et al. Language models are few-shot learners. *Advances in neural information processing systems*, Vol. 33, pp. 1877–1901, 2020.

[5] X. Zhou, et al. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pp. 350–368. Springer, 2022.

[6] T. Yamamoto, et al. Human support robot (hsr). In *SIGGRAPH 2018 emerging technologies*, pp. 1–2. ACM, 2018.