基盤モデルを活用した自然言語による多様なタスク実現に向けたロボットシステムの統合

○辻知香葉　綱島颯志　保呂蒼威　生駒創　小武海大　大見謝恒和　白坂翠萌　和田輝　池田悠也　松嶋達也　松尾豊　岩澤有祐　（東京大学）

Executing various commands in various environments requires a robot system with high generalizability and adaptability. We report a case study of a robot system that integrates several foundation models (CLIP, Detic, GPT, Sentence BERT, and Whisper). In particular, we discuss the performance of the built system on the subject of General Purpose Service Robot, one of the competitions of RoboCup@Home, which is held for the technological development of home service robots. Our robot system won the competition with a perfect score in each category, verifying the system's effectiveness.

1. 序章

RoboCup@Home[1] はキッチンやリビングといった日常生活の場での人間との共同作業を追求する家庭内サービスロボットの技術発展を目指した競技会である．本稿では，音声認識・タスクプランニング・物体認識・人認識・ナビゲーション・ヒューマンロボットインタラクション・マニピュレーション等家庭内サービスロボットに必要な要素を網羅的に要求される General Purpose Service Robot (GPSR) という RoboCup@Home の競技に向けて，その解法を提案する．

提案した手法をトヨタ自動車株式会社が開発した Human Support Robot (HSR) [12] に実装し，2023 年 3 月に東京で開催された RoboCup Japan Open 2022，2023 年 5 月に滋賀で開催された RoboCup Japan Open 2023，および 2023 年 7 月に仏・ボルドーで開催された RoboCup 2023 の@ Home・Domestic Standard Platform League (DSPL) にて性能評価を行った．RoboCup Japan Open 2022 では GPSR 部門第 2 位，総合第 3 位，RoboCup Japan Open 2023 では GPSR 部門第 1 位，総合優勝，および RoboCup 2023 では GPSR 部門第 2 位，総合第 3 位を獲得し，本手法の有効性を示した．

2. 前提知識

2.1 General Purpose Service Robot

GPSR は，家庭内環境で人間から自然言語で与えられる様々なコマンドを実行するタスクである．このタスクでは，物体の運搬や人の案内，質問への回答など，多岐にわたるコマンドを実行する必要があり，これらのコマンドはコマンドジェネレータ [8] によってランダムに生成される．GPSR タスクの特徴的な点として，実行環境や使用されるオブジェクトは競技会直前に設けられるセットアップデーに公開されることが挙げられる．これにより，参加チームは，特定の環境に合わせてプログラムすることなく，実際の家庭環境に近い状況でロボットを動作させる必要がある．

2.2 基盤モデル

基盤モデル [2] とは，膨大なデータでの学習により高い汎化性と適応性を持ち，1 つのモデルで様々なタスクにチューニング可能な大規模 AI モデルのことである．

Table 1　Foundation models and their application

Foundation model	Application
Whisper	Speech-to-Text
GPT-3・GPT-4	Task planning
Detic	Object detection・Environment recognition
CLIP	Object classification・Environment recognition
Sentence BERT	Environment recognition

2.3 プロンプトチューニング

プロンプトチューニングとは，基盤モデルを特定のタスクに適応させるチューニング手法の 1 つである．モデルへの入力として与えるプロンプトと呼ばれるテキスト情報を操作する．必要な情報や文脈をモデルに提供できるよう，タスクの目的や要件に応じてプロンプトを適切に設計することで，モデルの応答や出力結果の改善を図る．

3. システム概要

様々な環境下で多様なコマンドを実行するには，異なる環境や条件でも対応できる汎化性と，環境や条件の変化に対して柔軟かつ容易にチューニングできる適応性が必要である．そこで本システムでは，高い汎化性と適応性を併せ持つ基盤モデルを複数活用することで対応を試みた．

本システムでは基盤モデルとして，音声認識での文字起こしには Whisper[6]，タスクプランニングには GPT-3[3]（text-davinci-003）および GPT-4[4]，物体認識での物体検出には Detic[13]，物体分類には CLIP[5] をそれぞれ使用した．また，環境情報の統合に用いる CLIP-Fields[9] の枠組みには 3 つの基盤モデルを使用しており，Detic は物体認識，CLIP は画像のエンコーディング，Sentence BERT[7] は画像ラベルのエンコーディングに用いた．使用した基盤モデルとその用途を Table 1 に示す．

コマンドから大規模言語モデル (Large Language Models, LLM) の言語知識のみでプランニングを行っても，そのプラン（以下 LLM プランと呼ぶ）は環境情報を考慮していないため必ずしも実行可能とはかぎらない．そこで，基盤モデルベースのモジュールで集めた環境情報を統合し，LLM のもつ言語知識と接地す

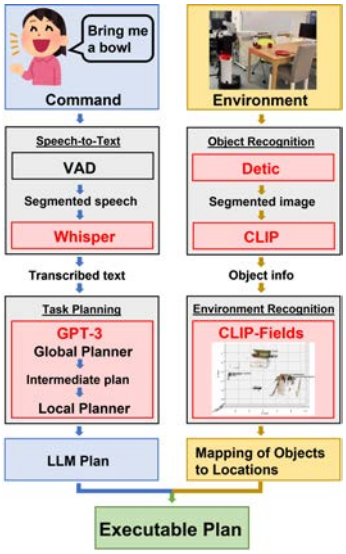

Fig. 1 System overview. Our system outputs an LLM plan from a command (blue) and an object-location mapping from the environmental information (yellow). The system finally outputs an executable plan by integrating the two (green).

ることで，LLM プランを実行可能なプランに変換する (Fig. 1)．例えば，”Bring me a bowl.”というコマンドに対し，LLM プランは「bowl がある場所にナビゲーションする」と言語的に解釈してプランニングをするが，実行環境における bowl の位置を考慮していない．そのため，この LLM プランのままでは実際に bowl までナビゲーションはできない．この LLM プランに環境情報からロボットが得た bowl の具体的な座標情報を加えることでナビゲーションが可能になり，ロボットはコマンドの遂行ができるようになる．また，タスク実行中に環境データを集めてロボットが保持する環境情報をアップデートすることで，環境の変化に対応したプランを出力する．具体的には，タスク実行中に RGB-D 画像を収集した後，3 つの基盤モデル（Detic・CLIP・Sentence BERT）を使ってデータセットを作成し CLIP-Fields を再学習することで，実行環境における物体とその位置情報の対応付けを更新する．これにより，例えば，”Bring me a bowl.”というコマンドに対し，bowl の位置移動前と後では物体の位置移動を考慮して bowl の座標が異なるプランを出力する．

4. 提案手法

本章では，GPSR タスクに向けて HSR に実装した機能のうち，基盤モデルを活用した 4 つの主要機能について述べる．

4.1 音声認識

人間とロボットのインタラクションは英語で行われる．音声認識は，発声区間認識モジュールと文字起こしモジュールの 2 つから構成される．発声区間認識には，DNN モデルである Silero VAD[10] を，文字起こしには基盤モデルである Whisper を活用した．この音声認識システムは，ノイズのある環境や様々なアクセントの英語を想定して開発された．

具体的には，HSR のマイクで拾った音源のうち，英語が発声されている確率が一定の閾値以上である発声区間をセグメント化し，それらのセグメントのみを Whisper に入力して文字起こしを行う．この発声区間認識により，ノイズに対してロバストな認識を目指した．

また，Whisper は，会話での発言などをプロンプトとして与えることにより文脈を把握して文字起こしを行う．そのため，GPSR のタスク設定情報や実行環境の説明をプロンプトとして Whisper に与えることで，単純なフレーズの認識にとどまらず，発話内容と GPSR タスクの関連性を考慮した文字起こしを行うことができた．具体的には，文字起こしのフェーズでもノイズ（文脈に沿わない単語など）を取り除くことや，タスク特有の聞き取りにくい単語（e.g., bamboo shoot) の認識精度の向上，さらにはアクセントに対してもよりロバストな認識を実現することができた．

4.2 タスクプランニング

多様なコマンドを実行するため, 23 種類のアクション (e.g., navigation, pick) に一対一対応した 23 個のスキル関数 (e.g., navigation 関数, pick 関数) を用意した．各スキル関数は，それぞれ 1〜3 個の引数を指定することで対応するアクションを実行する．タスクプランニングでは，GPT-3（text-davinci-003）および GPT-4 を用いて，コマンドをスキル関数の組み合わせで達成できるようスキル関数の適切な選択・順序立て・引数指定（以下，プランニングと呼ぶ）を行う．プランニングの最終的な出力プランが LLM プランである．

プランニングは，多段階プロンプトによる Chain-of-Thought[11] の手法を採用し，2 段階に分けて行った．1 段階目では指示されたコマンドを入力とし，思考の過程を添えた大まかなプランを出力した．これを中間プランと呼ぶ．2 段階目では, 1 段階目の出力である中間プランを入力とし，引数が指定されたスキル関数の組み合わせである LLM プランを出力した．なお，2 段階目では，入力した文字列に応じて特定の関数を呼び出す機能である，Function calling を用いて，引数の指定及びスキル関数を呼び出した．中間プランを経ることにより，1 段階でダイレクトにプランニングする場合に比べ，実際の動作順序を考慮したプランニングが可能になった．特に, コマンドの語順と語句に対応するスキル関数の順序が異なる場合 (e.g., Tell me how many fruits are on the dining table.) には，本手法が有効であった．

4.3 物体認識

物体認識モジュールは，物体検出モジュールと物体分類モジュールの 2 つから構成される (Fig. 2)．物体検出モジュールには，基盤モデルである Detic を使用し，物体分類モジュールには同じく基盤モデルである CLIP を使用した．

物体検出モジュールは，検出結果に対して 3 つのフィルタリングを行い，見切れている物体・小さすぎる物体・包含されている物体を削除し，セグメントされた画像を出力する．また，従来の学習済みモデルはクラスが固定されているのに対し，Detic では任意の語彙（プロンプト）をクラスに指定することができるという特徴も利用した．従来の DNN ベースモデルの手法では，未知物体を検出するためには学習が必要であり，また，

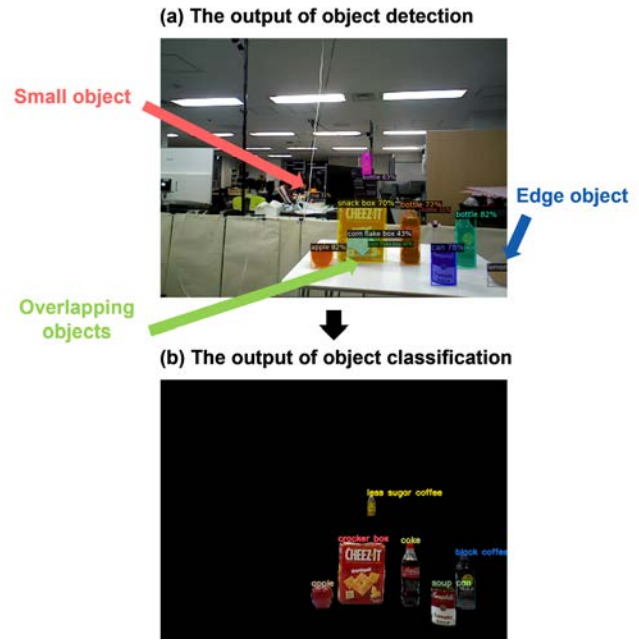

Fig. 2 Results of object recognition. Three filters are applied to (a) the object detection results, then (b) the object classification results are output.

検出対象でない物体が誤検出されることもある．しかし，プロンプトチューニングの手法を用いて部屋の中に置かれる物体に合わせてプロンプトを調整することで，学習せずに未知物体の検出が可能になり，また，検出対象でない物体を除外することができた．具体的には，対象のオブジェクトのみがすべて検出されるプロンプトチューニング済みのプロンプトのリストを Detic に適応した．

物体分類モジュールでは，物体検出モジュールからのセグメントされた画像を CLIP を用いて分類する．プロンプトチューニングだけでは分類が難しい未知物体に対しては学習を行った．各物体につき 30〜60 秒程度で 100〜500 枚の画像を収集し，データオーグメンテーションを行って，学習済み CLIP モデルに続く全結合層のみをチューニングした．

4.4 環境認識

コマンドを実行するには，物体とその位置情報を結び付ける必要がある．本システムでは，基盤モデルである CLIP とニューラル場表現を組み合わせることでセマンティックな表現を空間に保持する CLIP-Fields を活用し，環境データ（RGB-D 画像）から物体と位置の対応付けを学習する．

まず，コントローラーを用いてロボットを部屋内で移動し，RGB-D 画像を収集して CLIP-Fields を学習する．その後，タスク実行中にデータを収集し，CLIP-Fields を再学習して環境情報を更新することで，環境の変化に対応する．

Fig. 3 の例では，最初のコマンド（Take a bowl and place it on the shelf.）は，bowl と shelf の位置を CLIP-Fields を用いて推論し，ナビゲーションを行った．そして，実行中に人間が介入し lemon を shelf に移動した．HSR は最初のコマンドの実行中にデータ収集と再学習

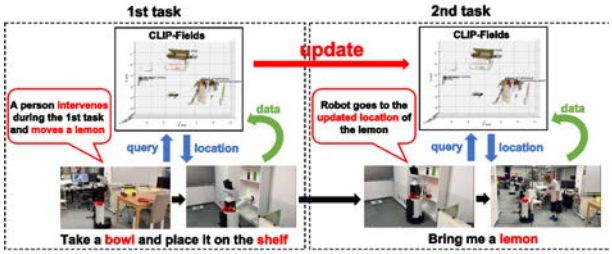

Fig. 3 Executing GPSR task. A human moved the location of a lemon during the first command execution. HSR updated the environment information and executed the second command considering the lemon's shift in location.

Table 2 Accuracy of foundation model-based modules and command completion rate

Foundation model-based module	Number of trials	Accuracy
Speech recognition	25	92.0%
Task planning	23	91.3%
Object recognition	17	94.1%
Command completion rate	23	52.2%

を行うため，2 回目のコマンド（Bring me a lemon.）では，lemon の位置を正しく推論し，移動後の位置である shelf にナビゲーションした．

5. 結果

我々は 2023 年 3 月に東京で開催された RoboCup Japan Open 2022, 2023 年 5 月に滋賀で開催された RoboCup Japan Open 2023, および 2023 年 7 月に仏・ボルドーで開催された RoboCup 2023 の@ Home・DSPL に参加し，本システムの性能評価を行った．

3 大会通じて 23 コマンドに挑戦した結果を Table 2 に示す．集計にあたり，音声認識の分母にはコマンドの聞き取りとコマンド実行中の聞き取りを含め，物体認識の分母には，物体認識を必要とするコマンドのみを含めた．また，タスクプランニングとコマンド完遂率の分母は挑戦した全てのコマンドであり，タスクプランニングに失敗したケースもコマンド完遂率の分母に含めた．成功判定は，音声認識および物体認識は競技会の審査員が正しく認識したと見なした場合を成功とした．タスクプランニングは，人間が妥当なプランだと判断したものを成功とし，コマンド完遂率は，最後までコマンドを遂行しきれた場合のみを成功とした．尚，HSR が自律的に人間の手助けを要求してコマンド実行を継続した場合も成功とした．

特に，RoboCup Japan Open 2023 の 5 トライアル目では，10 分の制限時間内に 3 コマンド（1 トライアルあたりの上限は 3 コマンド）全てを完遂することができ，カテゴリ 2（難易度別に 4 カテゴリあり，易しい順にカテゴリ 0，カテゴリ 1，カテゴリ 2，カテゴリ 3 となる）の満点である 170 点を獲得した．また，競技結果は，RoboCup Japan Open 2022 では GPSR 部門第 2 位，総合第 3 位，RoboCup Japan Open 2023 では GPSR 部門第 1 位，総合優勝，および RoboCup 2023 では GPSR 部門第 2 位，総合第 3 位であった．

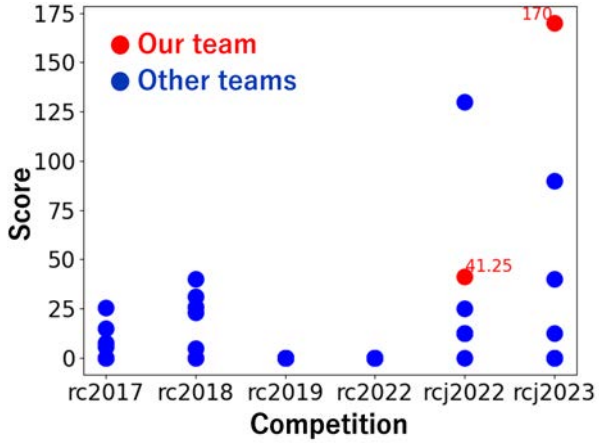

Fig. 4 Comparison with other teams in GPSR. Blue dots show other teams' scores, and red dots show our team's.

Table 3 Causes of command completion failure and percentage of failed commands

Causes of command completion failure	Number of failed commands	Percentage of failed commands
Failure to execute skill functions using foundation model-based modules	11	27.3%
Failure to execute other skill functions	11	63.6%
Time-out	11	9.1%

6. 考察

RoboCup@Home の競技会における GPSR タスクの得点の遷移を Fig. 4 に示す．我々は基盤モデルベースのモジュールを多用し統合することで，基盤モデル台頭前の最高点である 40 点 (2018 年) を 425 ポイント上回った 170 点を獲得し，GPSR タスクの得点を大幅に引き上げることができた．これは基盤モデルの有効な活用により，GPSR タスクの改善が実現できたことを示していると考えられる．

しかし，基盤モデルベースのモジュール別の精度は 90 %以上と高かった一方，コマンド完遂率は約 50 %と低かった．コマンド完遂の失敗原因別にみた，11 個の失敗コマンドに対する各割合を Table 3 に示す．これより，コマンド完遂率の低さの一因として，follow_person 関数での人追従失敗といった，基盤モデルを使わないスキル関数自体の精度の低さが挙げられる．これらのスキル関数の改善によって，コマンド完遂率の向上が見込まれると推察できる．

また，獲得していないアクションを必要とするコマンドの実行ができない課題もある．今後は人間の教示による模倣学習を通じて新しいアクションを学習し，実行可能なコマンドの種類を増やすことも視野に入れたい．これにより，より幅広いコマンドの遂行が可能になり，総合的なパフォーマンスの向上が期待される．

7. 結論

本稿では複数の基盤モデルを統合したロボットシステムの構築事例を報告し，特に，RoboCup@Home の競技の 1 つである GSPR を題材に，構築したシステムの性能について議論した．競技会では過去最高点を獲得し，本システムの有効性を示した．一方で，スキル関数の改善や新たなアクションの獲得により，GPSR タスクのさらなる向上とコマンド完遂率の改善を目指す必要があることが示唆されたため，今後の研究ではこれらの課題に取り組み，ロボットの性能向上を図っていくことが重要である．

参 考 文 献

[1] RoboCup@Home, Accessed 2023-07-01. `https://www.robocup.org/domains/3`.

[2] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[3] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[4] OpenAI. GPT-4 Technical Report. *arXiv e-prints*, page arXiv:2303.08774, Mar. 2023.

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[7] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

[8] RoboCup@Home. Robocup@home command generator. `https://github.com/kyordhel/GPSRCmdGen.git`, 2015.

[9] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022.

[10] S. Team. Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. `https://github.com/snakers4/silero-vad`, 2021.

[11] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[12] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase. Development of human support robot as the research platform of a domestic mobile manipulator. *ROBOMECH journal*, 6(1):1–15, 2019.

[13] X. Zhou, R. Girdhar, A. Joulin, P. Krähenbühl, and I. Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022.