基盤モデルと古典プランニングを用いた レシピ記述からの実世界調理計画認識実行ロボットシステム

金沢直晃 [1]　河原塚健人 [1]　大日方慶樹 [1]　岡田慧 [1]　稲葉雅幸 [1]
[1] 東京大学大学院 情報理工学系研究科
{kanazawa,kawaharazuka,obinata,k-okada,inaba}@jsk.imi.i.u-tokyo.ac.jp

概要

ロボットによる調理実行においては，自然言語のレシピ記述に基づいて作業を行えることが望ましいが，大規模言語モデルが登場した現在においても依然として課題が存在する．本研究では特に重要な 2 つの課題に着目し，大規模言語モデルと PDDL 記述の古典プランニングによる実世界で実行可能なロボット調理行動計画と，視覚-言語モデルを用いた少数データからの食材状態認識学習を行うロボットシステムを提案した．実世界でロボットが調理を実行する実機実験により提案システムの有効性を確認した．

1　はじめに

日常生活に欠かせない家事である調理においては，環境 · エージェント非依存なタスク記述であるレシピに従い，実行する環境の状況に合わせて調理作業を実行することが必要である．これまでも生活支援ロボットの文脈で，自然言語で記述されたレシピからロボットの行動計画及び実行を行う研究が数多く行われてきている [1, 2, 3, 4, 5, 6]．また，自然言語の指示からのロボットの行動計画については，大規模言語モデル (LLM) [7] が登場し，ロボットの行動計画への応用 [8, 9] も進められていて，これまでのルールベースの処理よりも高度に自然言語指示に対応可能な行動計画が行えるようになってきている．しかし，ロボットがレシピ記述に基づいて実世界で調理作業を実行するためには依然として課題が存在する．本研究では特に重要である「実世界で実行可能な行動計画」と「食材状態変化の認識」の 2 つの課題に着目する．

1 つ目の「実世界で実行可能な行動計画」の課題は，実際の環境の状態からレシピに書かれた調理行動を順に実行していくために必要な行動を適宜補完していく行動計画が必要であるというものである．[6] で指摘されている通り，「ボウルの中の卵を鍋に注ぐ」ためにはその前にボウルを持つ必要があるなどの，人が無意識に行うためにレシピには省かれて記述されないが実際には必要な行動が存在する．また，その時のキッチン環境の状況によって，目標の調理行動を行うための準備として必要な行動も存在する．ロボットが実世界で調理行動を実行するためには，それらを補完して実行可能な行動手順を計画する必要がある．そこで本研究では，LLM を用いて自然言語のレシピ記述をプログラムが解釈可能な調理関数シーケンスに変換し，変換された調理関数を PDDL [10] を用いた古典プランニングにより必要な行動を補完して実世界で実行可能にする行動計画 [11] を行う．

2 つ目の食材状態変化の認識の課題は，調理では作業中に食材が大きく状態変化し，「水が沸騰するまで加熱する」「バターが溶けたら卵液を注ぐ」など

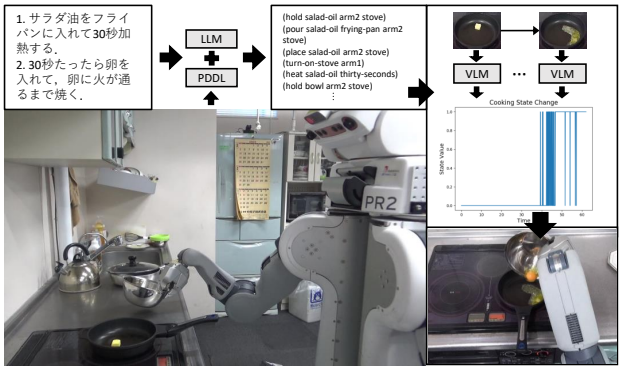

図 1: レシピからの調理実行ロボットシステム．大規模言語モデルによりレシピを関数シーケンスに変換，PDDL [10] 記述の古典プランニングにより実行可能なように補完した行動手順を計画．視覚-言語モデルを用いた少数データからの食材状態認識学習により調理中の食材の状態を認識しながらロボットが調理を実行．

の食材の状態変化に条件付けられた記述を含むレシピに対応するために，それらの状態変化をロボットが認識する必要があるというものである．食材の状態認識においても，これまでも専用のデータセットを用いて CNN を学習し状態分類を行う研究 [12, 6] などが行われてきている．しかし，調理においては様々な食材や状態変化が存在するためそれら全てのデータを大量に収集するのは困難であるため，必要なデータをなるべく少量にする工夫が必要であると考えられる．本研究では，視覚-言語モデル [13] を用いた少数データからの食材状態認識学習 [11] を行う．

以上を踏まえて本研究では，自然言語で書かれた料理レシピを基にロボットが実世界で実行可能な調理行動を計画し，食材の状態変化を認識しながら調理を実行するロボットシステム (図 1) を提案する．

2　レシピからの実世界で実行可能なロボット調理行動計画

本研究では，自然言語のレシピ記述からの実世界で実行可能なロボット調理行動の計画法 (図 2) を提案する．まず，大規模言語モデルを用いたレシピからの調理関数シーケンス生成 (2.1) により，レシピ記述をプログラムが解釈可能な調理関数のシーケンスに変換する．次に，変換された調理関数のシーケンスを実世界で実行可能にするために PDDL [10] の記述を用いた古典プランニングによりレシピに省かれている行動や現在の環境の状態から目標の調理行動を実行するために必要な行動を補完する (2.2).

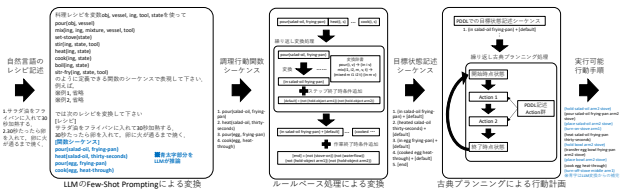

図 2: 自然言語のレシピ記述からの実世界で実行可能なロボット調理行動の計画法．大規模言語モデルを用いて自然言語のレシピから調理関数シーケンスを生成し，ルールベースの処理と PDDL [10] 記述を用いた古典プランニングにより行動を計画することで，大規模言語モデルによる変換結果を実世界で実行可能なように補完した行動手順を計画.

2.1　大規模言語モデルを用いたレシピからの調理関数シーケンス生成

調理では，食材の状態をレシピの手順に従って変化させていき目標の食材状態を実現することで最終的な料理を完成させることが目的となる．そのため，本研究では食材の状態を変化させる調理行動の関数については目標の食材状態の記述を引数として設定した行動関数記述を用いる．

本研究で対象とする目玉焼き，ポーチドエッグ，スクランブルエッグの 3 つの卵料理に含まれる調理関数表現として以下の 8 つの関数を考えた．

- pour(ingredient, vessel)
- mix(ingredient, ingredient, mixture, vessel, tool)
- set(stove, state)
- stir(ingredient, state, tool)
- heat(ingredient, state)
- cook(ingredient, state)
- boil(ingredient, state)
- sitr-fry(ingredient, state, tool)

ここで，ingredient は食材，mixture は複数食材の混合物であり ingredient の属性も持つ，vessel は容器，tool は道具を意味する変数である．

実験のために，目玉焼き，ポーチドエッグ，スクランブルエッグの 3 つの卵料理の日本語レシピ記述とその理想的な変換を人がアノテーションしたものを用意した．GPT-4 [7] を用いて，他の 2 つの料理についてレシピ記述とアノテーションを例示した Few-Shot Prompting によりレシピの変換を推論 (図 2) する実験を 3 つのレシピに実行した．実験の結果，人のアノテーションとの細かい表現の差異はあるが，概ね正しい関数シーケンス変換が行えることを確認した (3).

図 3: GPT-4 [7] によるレシピ記述の調理関数シーケンス変換結果．人のアノテーションと異なっている部分を赤字下線で示している．

2.2 PDDL を用いた実行可能な調理行動手順計画

本研究では，変換された調理関数シーケンスに対し，人は無意識に実行するためレシピには省かれてしまっている行動や，その時の環境の状態から目標の調理行動を実行するために必要な行動を補完し，目標とした調理行動を実世界で実行可能な行動手順計画を行う．

対象とする調理行動計画の問題を PDDL [10] の問題記述に落とし込み，PDDL の action として 2.1 で提案した調理関数の 8 つに加えて，直接食材の状態を変化させない行動も含めた調理に必要な行動として hold, place, move-to, open-tap, close-tap, turn-on-stove, turn-off-stove, transfer, fetch-water の 9 つを追加した合計 17 の action を考える．これらの action と対応する precondition と effect の predicates をそれぞれ定義し，調理行動計画の domain を用意する．実環境での実行可能性を考慮した計画にするために，これらの action や predicates には arm や spot などの調理関数には含まれていない要素も追加して定義する．これにより，双腕ロボットで実行するのに腕が 3 本必要な調理行動を実行する，その場所に存在していない物体を使おうとするなどの実行不可能な行動が計画されることを防ぐことができる．

2.1 で変換された調理関数シーケンスを，それぞれの関数に対応する PDDL の action の effect のシーケンスに変換することで，目標調理状態のシーケンスを取得する．シーケンスの各ステップにおいて，開始時点での状態と目標状態を繋げるように PDDL 記述での古典プランニングを解き，目標状態の条件を満たした時点での状態を得る．それを次のステップの開始時点での状態として，古典プランニングを繰り返すことで実行可能な行動手順を計画する (図 2)．また，人が調理する際と同じような自然な行動になるように，各ステップの目標状態には，変換された effect に追加してロボットは何も持っていない状態であるという条件を追加し，全てのシーケンス終了後はロボットは何も持っておらず，コンロも水道も off の状態であるという条件を追加している．目玉焼き，ポーチドエッグ，スクランブルエッグの 3 つの料理に対して関数表現シーケンスから行動手順計画を行い，実行可能な計画が行われていることを定性的に確認した．なお，今回の実験では最初の開始時点の状態である初期条件はそれぞれの料理に必要な材料がキッチンにセットされている状態から計画を行ったが，それぞれの材料や道具が決められた置き場にセットされているという条件や，ロボットが何らかの方法で認識した初期条件から計画することも可能である．

3 視覚-言語モデルを用いた少数データからの食材状態認識学習

本研究では，CLIP の論文 [13] でも議論されている Linear-Probe を用いた少数データからの食材状態認識学習法 (図 4) を提案する．この Linear-Probe では，CLIP の Image-Encoder が出力する画像特徴量に対して線形識別器を学習し，画像の分類を行う．食材の状態認識としては，過去の食材状態変化時系列データについての人のアノテーション時刻より前の画像を変化前 (0)，アノテーション時刻以降の画像を変化後 (1) とラベルをつけて Linear-Probe を学習する．この方法では多次元の画像特徴量に対する識別器の学習を行っており，実験を行う度に獲得される新規データを認識器の学習データに追加することが可能になりロバスト性を向上させていけるため，安定した食材状態認識が可能になる．

卵料理に含まれる「水の沸騰」，「バターの融解」，「卵のタンパク質の熱変性」の 3 つの状態変化のデータを収集し，提案手法を適用する実験を行った (図 5)．それぞれのデータの最初の画像に対して Open-Vocabulary な物体検出器である GroundingDINO [14] を用いて自然言語のプロンプトからフライパンや鍋の中身もしくは全体の領域を取得した．以降はこの取得した領域を注視領域として状態変化認識の処理を行い，(a) 鍋やフライパンの中身領域と，(b) 鍋やフライパンの全体領域のどちらを注視領域としたほうが認識器の性能が良くなるかを比較した．

それぞれの状態変化について 2 つの時系列データを取得し，1 つのデータを学習用データとして利用して提案手法により状態変化認識器を学習し，もう 1 つの時系列データに対して認識を行うことで提案手法の有効性を検証した (図 5)．時系列データの最初の時刻から順に画像の状態分類を実行し，変化後のラベルである 1 と最初に判定された時刻を状態変化の推論結果とした．この推論結果の時刻と，人のアノテーション時刻を比較して性能を評価した．実験の結果，フライパンや鍋の全体領域よりもフライパンや鍋の中身領域を注視領域として利用した条件の方が認識結果が良くなり，妥当な状態変化認識を

行えることが確認できた．

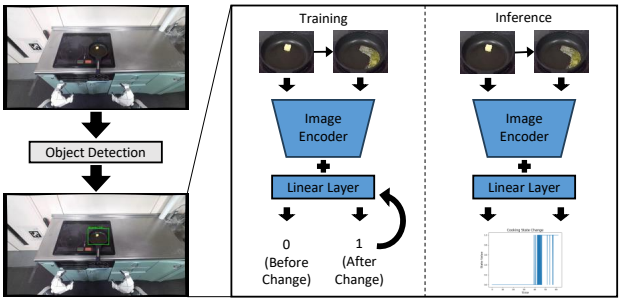

図 4: 視覚-言語モデルを用いた少数データからの食材状態認識学習．GroundingDINO [14] を用いてフライパンや鍋の中身もしくは全体の領域を取得し注視領域とし，その領域の画像について CLIP [13] の Linear-Probe を用いた少数データからの食材状態認識学習によって状態変化を認識．

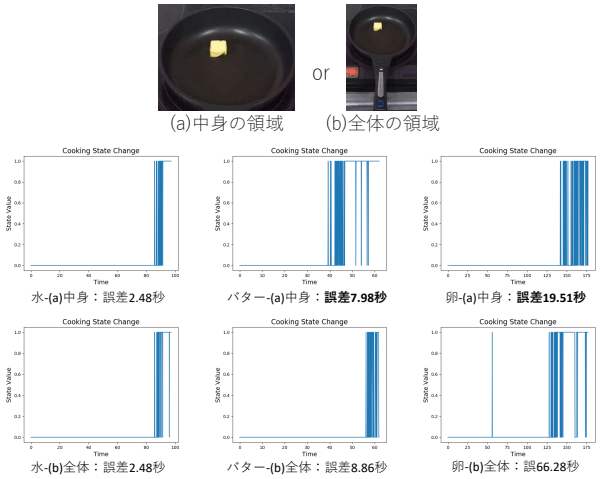

図 5: 食材状態変化認識の実験結果．変化後ラベルの 1 と最初に判定された時刻を状態変化の推論時刻とし，人のアノテーション時刻と比較して性能を評価．フライパンや鍋の全体領域よりも中身領域を注視領域として利用した条件の方が認識結果が良くなり，妥当な状態変化認識を行えることを確認．

4 ロボットによる実世界での調理実行実験

提案したシステムを用いて，自然言語のレシピ記述から実世界で実行可能な調理行動手順を計画 (2) し，加熱中の食材の状態変化を視覚-言語モデルを用いた食材の状態認識 (3) をしながら順に調理行動を実行し，実世界でロボットが目玉焼きを調理する実機実験 (図 6) を行い，提案システムの有効性を確認した．今回の実験では動作部分については，人がダイレクトティーチにより教示した動作を実行している．

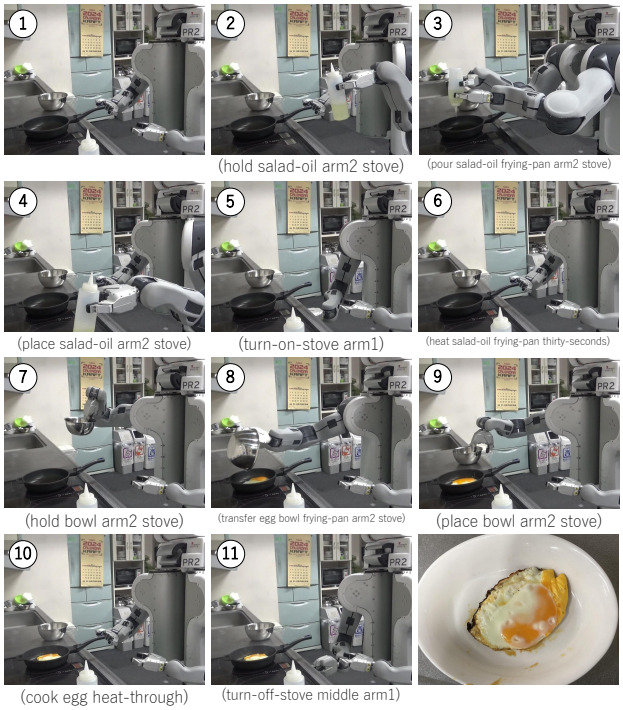

図 6: 実世界でのレシピ記述からのロボットの目玉焼き調理実行実験．自然言語のレシピ記述に基づいて提案システムにより，行動計画及び食材状態認識を行って順に調理を実行し，実際に目玉焼きを調理できることを確認．

5 おわりに

本研究では自然言語で書かれた料理レシピを基にロボットが実世界で調理を実行可能にするために，大規模言語モデルと PDDL 記述の古典プランニングによる実世界で実行可能なロボット調理行動計画と，視覚-言語モデルを用いた少数データからの食材状態認識学習を行うシステムを提案した．実際に実世界で目玉焼きの調理を実行することで提案システムの有効性を確認した．

今後は，動作計画部分においてもレシピ記述に基づいた動作計画を行い，全体の調理実行ロボットシステムとして統合していくことで，より多くの料理やレシピ記述を多様な環境で実行できるように全体のシステムを拡張していく．調理関数表現をより多くの料理品目に対応可能なように拡張することで多くの料理やレシピに対応できるようになると考えられるが，その際にも本研究で提案した行動計画と食材状態認識が有効であると考えられる．

参考文献

[1] Michael Beetz, Ulrich Klank, Ingo Kresse, Alexis Maldonado, Lorenz Mösenlechner, Dejan Pangercic, Thomas Rühr, and Moritz Tenorth. Robotic roommates making pancakes. In **2011 11th IEEE-RAS International Conference on Humanoid Robots**, pp. 529–536, 2011.

[2] Mario Bollini, Stefanie Tellex, Tyler Thompson, Nicholas Roy, and Daniela Rus. Interpreting and executing recipes with a cooking robot. In **Experimental Robotics**, pp. 481–495. Springer, 2013.

[3] Gayane Kazhoyan and Michael Beetz. Programming robotic agents with action descriptions. In **2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**, pp. 103–108. IEEE, 2017.

[4] Masahiro Inagawa, Toshinobu Takei, and Etsujiro Imanishi. Analysis of cooking recipes written in japanese and motion planning for cooking robot. **ROBOMECH Journal**, Vol. 8, No. 1, pp. 1–13, 2021.

[5] David Paulius, Kelvin Sheng Pei Dong, and Yu Sun. Task planning with a weighted functional object-oriented network. In **2021 IEEE International Conference on Robotics and Automation (ICRA)**, pp. 3904–3910. IEEE, 2021.

[6] Kota Takata, Takuya Kiyokawa, Ixchel G Ramirez-Alpizar, Natsuki Yamanobe, Weiwei Wan, and Kensuke Harada. Efficient task/motion planning for a dual-arm robot from language instructions and cooking images. In **2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)**, pp. 12058–12065. IEEE, 2022.

[7] OpenAI. Gpt-4 technical report. **arXiv preprint arXiv:2303.08774**, 2023.

[8] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, et al. Do as i can, not as i say: Grounding language in robotic affordances. **arXiv preprint arXiv:2204.01691**, 2022.

[9] Marta Skreta, Naruki Yoshikawa, Sebastian Arellano-Rubach, Zhi Ji, Lasse Bjørn Kristensen, Kourosh Darvish, Alán Aspuru-Guzik, Florian Shkurti, and Animesh Garg. Errors are useful prompts: Instruction guided task programming with verifier-assisted iterative prompting. **arXiv preprint arXiv:2303.14100**, 2023.

[10] Constructions Aeronautiques, Adele Howe, Craig Knoblock, ISI Drew McDermott, Ashwin Ram, Manuela Veloso, Daniel Weld, David Wilkins SRI, Anthony Barrett, Dave Christianson, et al. Pddl— the planning domain definition language. **Technical Report, Tech. Rep.**, 1998.

[11] 金沢直晃, 河原塚健人, 大日方慶樹, 岡田慧, 稲葉雅幸. 対象物状態中心の調理行動記述に基づくレシピからの卵料理の実世界調理実行ロボットシステム. 第 24 回 SICE システムインテグレーション部門講演会講演概要集, pp. 3G2–08, Dec 2023.

[12] Rahul Paul. Classifying cooking object's state using a tuned vgg convolutional neural network. **arXiv preprint arXiv:1805.09391**, 2018.

[13] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **International Conference on Machine Learning**, pp. 8748–8763, 2021.

[14] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. **arXiv preprint arXiv:2303.05499**, 2023.