記号接地されたfetch-and-carryタスクの自動化と実行

神原元就，杉浦孔明（慶應義塾大学）

1. はじめに

日常タスクを支援するため，ユーザと自然にコミュニケーションを取ることが可能な生活支援ロボットの実用化は要支援者にとって有望な解決策の一つである．一方で，生活支援ロボットにおいて自然言語による指示を理解し適切に日常タスクを実行する能力は現状では不十分である．

本論文では，Fetch-and-Carry with Object Grounding (FCOG) タスクについての自然言語指示文が与えられたうえで，ロボットが視覚情報を基に指示を実行するための手法を構築することを目的とする．例えば，“Go to the bedroom, grasp the rabbit doll and send it to the corner sofa.” という指示文が与えられたうえで，ロボットが寝室へ移動し，ウサギの人形を特定したうえで把持し，最終的に部屋の隅にあるソファの上に人形を配置することが望ましい．

人間の発する指示はしばしば曖昧であり，対象となる物体やその目標領域をロボットが特定することは困難である．これは，既存システムにおいて，物体における記号接地に関する理解性能が不十分なためである．

既存の言語理解手法である Target–Dependent UNITER (TDU [1]) は物体操作指示文理解タスクにおいて高い精度を達成した手法である．一方で，この手法は複数の対象領域候補及び目標領域候補のペアを同時に入力とすることができない．従って，TDU は本タスクに適用することが難しい．理由として，fetch-and-carry タスクでは，環境にある複数の物体の中から，指示文について最も適切な候補ペアを特定することが求められる．そのため，同時に入力された候補ペア群の中から最も適切な候補ペアを予測できることが重要である．また，既存システムでは，FCOG タスクについての指示文を人手により生成していた．この場合，シミュレーションをオンザフライ方式で行うことができないため，ランダムに作成した多様なタスクで評価することができない．ゆえに，既存システムでは，固定されたタスクのみで評価を行っていた．

本論文では，TDU を拡張し，複数の対象領域及び目標領域候補のペアから最適な候補ペアを予測するタスクへ適用可能な手法を提案する．さらに，既存システムを拡張し自動指示文生成モデルを組み込むことで，シミュレーションをオンザフライ方式で行うことを可能とする．

本研究の独自性は以下である．

- TDU を拡張し，対象物体及び目標領域の候補群から最適な候補ペアの予測を行うことを可能にする．
- FCOG タスクにおいて，自動生成された指示文を利用することで，システム全体を自動化する．

2. 関連研究

HLSM–MAT [2] は Vision and Language Navigation (VLN) タスクのための手法の一つである．HLSM–MAT は，VLN タスクにおいて標準的なベンチマークである ALFRED ベンチマーク [3] において，既存手法を上回る性能を達成した．Target–Dependent UNITER [1] は物体操作指示文が与えられた上で，候補領域が指示文の示す領域であるか否かを予測するための手法の一つである．また，Case Relation Transformer [4] は対象物体を目標領域に移動するための物体操作指示文を生成する手法の一つである．これらの手法は，物体操作指示文についての Vision and Language タスクを扱っている手法であり，その点で本手法との関連が深い．

図 1 FCOG タスクにおけるシーン例．図中丸数字はシーンの順番を示す．

3. 問題設定

本論文で扱うタスクは，Fetch–and–Carry with Object Grounding (FCOG) タスクである．FCOG タスクでは，fetch–and–carry タスクについての自然言語指示文が与えられたうえで，ロボットが視覚情報を基に指示を実行する．本タスクでは，与えられた指示文に出現する対象物体及び目標領域について正確に記号接地を行うとともに，それらに関する fetch–and–carry タスクを実行することが望ましい．

図 1 に FCOG タスクの代表例を示す．この例は，指示文として “Go to the bedroom, grasp the rabbit doll and send it to the corner sofa.” が与えられた例を示している．ロボットはまず寝室に移動し，ウサギのぬいぐるみを把持する．続いて，ロボットは部屋の隅にあるソファの前まで移動し，ぬいぐるみをその上に乗せることが求められる．

FCOG タスクは移動，物体把持，object grounding, 物体配置の 4 つのサブタスクに分割できる．本論文ではこれらのサブタスクのうち，特に object grounding タスクに着目する．

object grounding における入出力は以下で定義される．

- **入力:** 自然言語指示文及びロボットの視点からの複数画像
- **出力:** 候補領域が正解である確率の予測値

また，以下に他の各サブタスクにおける入力及びタスクの成功についての定義を示す．移動における入力は

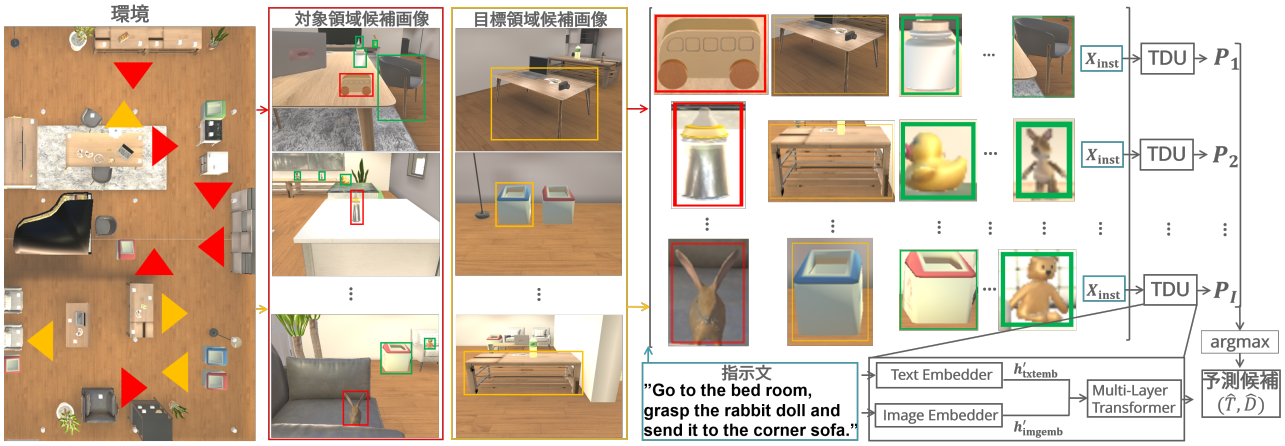


図 2 提案手法のネットワーク図．図中 TDU は拡張した Target–Dependent UNITER [1] を示す．また，$P_i = p(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}} | \boldsymbol{X}^{(i)}_{\text{targ}}, \boldsymbol{X}^{(i)}_{\text{dest}})$ である．

自然言語指示文であり，タスクの成功は指定された部屋への移動と定義する．物体把持における入力は，object grounding によって正解とされた対象物体候補の領域であり，タスクの成功は指定された物体の把持と定義する．物体配置における入力は，object grounding によって正解とされた配置領域候補の領域であり，タスクの成功は指定された領域への把持物体の配置と定義する．

本論文では以下のように用語を定義する．

- **対象物体:** 指示文が対象としている物体
- **目標領域:** 対象物体の移動先
- **候補領域:** 対象物体，または目標領域であるか否かの候補となる領域
- **コンテキスト領域:** 物体検出器で検出される物体

なお，対象領域，及びコンテキスト領域はそれぞれ対象物体，及びコンテキスト物体を囲む領域を指すものとする．また，FCOG タスクにおいて，移動，物体把持，及び物体配置の各サブタスクはルールベースにより行われるものとし，object grounding のみ学習ベースで行われるものとする．

4. 提案手法

提案手法は，Target–Dependent UNITER [1] を拡張した object grounding モデルである．提案手法の新規性は以下である．

- 対象物体及び目標領域の候補について，候補群から最適な候補の予測を行うことを可能にした．

図 2 に提案手法のネットワーク図を示す．ネットワークは 3 つのモジュールから構成され，それぞれ Image Embedder，Text Embedder，及び Multi-Layer Transformer である．図 2 に提案手法のネットワーク図を示す．

4.1 入力

モデルの入力 $\boldsymbol{x}_i$ を以下のように定義する．

$$\boldsymbol{x}_i = \{\boldsymbol{X}^{(i)}_{\text{targ}}, \boldsymbol{X}^{(i)}_{\text{dest}}, \boldsymbol{X}^{(i)}_{\text{det}}, \boldsymbol{X}^{(i)}_{\text{inst}} | i = 1, ..., I\}$$

ここで，I は Faster R-CNN [5] により検出された対象領域及び目標領域の候補を組み合わせた，ペアの総数を示す．以降，可読性のため i については省略する．すなわち，対象物体及び目標領域の候補ペアについて，候補群の内の一つについて記述する．これを用いて，入力を次のように定義する．

- $\boldsymbol{X}_{\text{targ}} = \{\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{targloc}}\}$
- $\boldsymbol{X}_{\text{dest}} = \{\boldsymbol{x}_{\text{dest}}, \boldsymbol{x}_{\text{destloc}}\}$
- $\boldsymbol{X}_{\text{det}} = \{\boldsymbol{x}^{(j)}_{\text{det}}, \boldsymbol{x}^{(j)}_{\text{detloc}} | j = 1, ..., J\}$
- $\boldsymbol{X}_{\text{inst}} = \{\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{pos}}\}$

ここに，$\boldsymbol{x}_{\text{targ}}$ は対象物体候補の領域，$\boldsymbol{x}_{\text{dest}}$ は目標領域候補の領域，$\boldsymbol{x}^{(j)}_{\text{det}}$ は各コンテキスト領域，$\boldsymbol{x}_{\text{inst}}$ は指示文を表し，$\boldsymbol{x}_{\text{targloc}}$ は対象物体候補の領域位置，$\boldsymbol{x}_{\text{destloc}}$ は配置目標候補の領域位置，$\boldsymbol{x}^{(j)}_{\text{detloc}}$ は画像中の各物体または配置先の領域位置，$\boldsymbol{x}_{\text{pos}}$ は指示文中の各単語の位置を表す．また，J は Faster R-CNN [5] により検出されたコンテキスト領域の数を表す．ここで，検出に使用する画像は，対象領域候補及び目標領域候補を切り出した，それぞれの元画像とする．入力特徴については，[1] に示された手順で抽出した．具体的には，$\boldsymbol{x}_{\text{targ}}, \boldsymbol{x}_{\text{dest}}, \boldsymbol{x}^{(j)}_{\text{det}}$ についてはバックボーンネットワークの中間出力，$\boldsymbol{x}_{\text{targloc}}, \boldsymbol{x}_{\text{destloc}}, \boldsymbol{x}^{(j)}_{\text{detloc}}$ については各領域の座標情報から生成したベクトル，そして $\boldsymbol{x}_{\text{inst}}, \boldsymbol{x}_{\text{pos}}$ については WordPiece を利用し生成した one-hot ベクトルとした．バックボーンネットワークである ResNet50 [6] における fc6 層の出力を画像領域の特徴量として抽出した．各対象物体候補及び目標領域候補について，1024 次元の特徴量を得た．

4.2 ネットワーク構造

Image Embedder では，対象物体候補，目標領域候補，及び画像中の各物体または配置先の領域に対する埋め込み処理を行う．具体的には，$\boldsymbol{X}_{\text{targ}}$，$\boldsymbol{X}_{\text{dest}}$，$\boldsymbol{X}_{\text{det}}$ を埋め込み，中間特徴量 $\boldsymbol{h}'_{\text{imgemb}} = \{\boldsymbol{h}'_{\text{targ}}, \boldsymbol{h}'_{\text{dest}}, \boldsymbol{h}'^{(1)}_{\text{det}}, ..., \boldsymbol{h}'^{(J)}_{\text{det}}\}$ を得る．Text Embedder では，指示文に対する埋め込み処理を行う．Multi–Layer Transformer では，L 層の Multi–Layer Transformer [7] によりモデルの最終的な予測確率を得る．入力は $\boldsymbol{h}^{(1)}_{\text{in}} = \{\boldsymbol{h}'_{\text{imgemb}}, \boldsymbol{h}'_{\text{txtemb}}\}$ である．本モジュールでは，候補のペアが正解であるかについての予測確率 $p((\hat{T}, \hat{D})) = \text{argmax}_{i \in 1,..,I}\, p(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}} | \boldsymbol{X}^{(i)}_{\text{targ}}, \boldsymbol{X}^{(i)}_{\text{dest}})$ を獲得する．ここで，$p(\hat{y}_{\text{targ}}, \hat{y}_{\text{dest}} | \boldsymbol{X}^{(i)}_{\text{targ}}, \boldsymbol{X}^{(i)}_{\text{dest}})$ は，対象物体候補 $\boldsymbol{X}^{(i)}_{\text{targ}}$ 及び目標領域候補 $\boldsymbol{X}^{(i)}_{\text{dest}}$ のペアが正

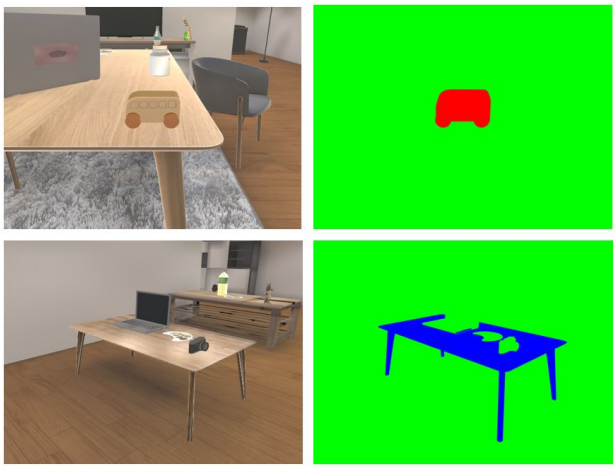


図 3 撮影された RGB 画像及び生成されたセグメンテーション画像．上段及び下段はそれぞれ対象物体及び目標領域についての例を示す．

図 4　CRT への入力領域例

表 1　提案手法のパラメータ設定

ヘッド数	12
レイヤ数	2
最適化関数	AdamW $(\beta_1 = 0.7, \beta_2 = 0.99999)$
学習率	8.0×10^{-5}
ステップ数	20000
バッチサイズ	8

解の領域であるかに関する予測確率を示す．最終的に予測対象物体 $\hat{T}$ 及び予測目標領域 $\hat{D}$ のペア $(\hat{T}, \hat{D})$ を出力とする．損失関数として交差エントロピー誤差を使用する．

5.　実験設定

5.1　データセット

事前訓練用データセットとして，WRS–UniALT2 データセット [8] を使用した．WRS–UniALT2 データセットは，画像と命令セットからなるシミュレーションベースのデータセットである．画像は World Robot Summit 2018 Partner Robot Challenge/Virtual Space Competition (WRS [9]) で使用された標準シミュレータにおいて収集されたものである．各画像には，部屋の中にある日常的な物体が 5 個程度含まれている．WRS-UniALT2 データセットにおいて，訓練，検証，及びテスト集合のサンプルサイズはそれぞれ 2048，210，及び 232 サンプルとした．訓練，検証，及びテスト集合はそれぞれモデルのパラメータの訓練，ハイパーパラメータの調整，及び MLU-FI タスク [1] におけるモデルの性能評価に使用した．

5.2　タスク及び指示文生成

実験では，指示文について，ルールベース及びクロスモーダル言語生成モデルの 2 通りで生成した．ルールベースによる指示文生成では，WRS で行われた手法に従った [9]．具体的には，まずランダムに対象物体及び目標領域を選択した．タスクにおいて，始めに移動する部屋は対象物体が置かれている部屋とした．指示文については，それらを示す単語を事前に作成したテンプレートに当てはめることで作成した．

CRT による生成では，入力として対象領域及び目標領域の画像ペアを使用した．ペアの抽出は以下のように行った．まず，対象物体及び目標領域とする物体について，環境の中からランダムに選択し情報を取得する．続いて，対象物体及び目標領域が含まれる画像をそれぞれ取得する．このとき，対象物体について，カメラ位置は正面方向かつ一定の距離，及び高さの地点とした．また，目標領域について，カメラ位置は目標領域となりうる家具ごとに予め定めておくこととした．以上の手順により獲得した画像群から，対象領域及び目標領域の情報を抽出した．具体的には，それぞれの画像について，シミュレータ内にある情報を基に対象物体及び目標領域のセグメンテーション画像を生成した．図 3 に，画像の例を示す．この図において，左列及び右列はそれぞれ RGB 画像及びセグメンテーション画像である．また，上段及び下段はそれぞれ対象物体及び目標領域についての例を示す．セグメンテーション画像において，赤色及び青色のセグメントがそれぞれ対象物体及び目標領域を示す．各領域はこのセグメンテーション画像を基に決定する．最終的に，得られた各画像及び領域情報を利用する．

指示文生成モデルとして，Case Relation Transformer (CRT [4]) を使用した．CRT への入力は，対象領域，目標領域，及びコンテキスト領域とそれらの座標情報であり，出力は FCOG タスクについての自然言語指示文である．ここで，コンテキスト領域として，対象領域が含まれる画像から，Faster R-CNN [5] によって抽出した領域群を使用した．図 4 に入力される領域例を示す．これらの領域は，図 3 の画像から抽出された領域を示す．この入力が与えられた時,「白い瓶の手前にあるおもちゃの木の車をテーブルの上に移動して」のような指示文が出力されることが望ましい．

CRT については，ALFRED–fetch データセット [8] を用いて訓練した．ALFRED–fetch データセットは，ALFRED ベンチマーク [3] を利用し構築したデータセットである．このベンチマークは，シミュレーション環境において，自然言語による指示文とロボットのカメラ画像から家事タスクにおけるロボットの行動を学習させるためのベンチマークである．ベンチマークは複数の逐次的なサブゴールから構成され，各サブゴールごとに指示文が存在する．ALFRED–fetch データセットでは，Fetch–and–carry タスクに関するサブゴールにおける指示文，把持直前及び配置直後のロボットの視覚画像を抽出した．ALFRED–fetch データセットは 47862 サンプルを含む．なお，本論文では，CRT を利用した指示文生成の場合においても，対象物体のある部屋へ移動するための指示文はルールベースで生成した．

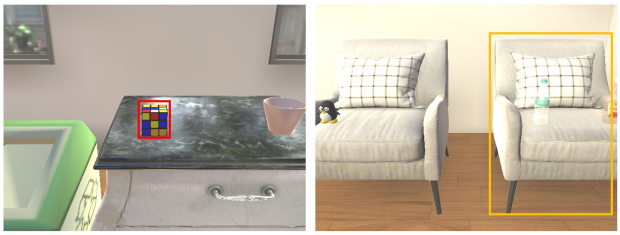
5.3　パラメータ設定

実験では WRS で使用されたシミュレーション環境を使用した．パラメータ設定を表 1 に示す．ここで，ヘッド数およびレイヤ数はそれぞれ transformer におけるヘッド数及びレイヤ数を表す．本提案手法の訓練ではメモリ 11GB 搭載 GeForce RTX 2080 Ti，64GBRAM，及び Intel Corei9 9900K を使用した．なお，Faster R-CNN による特徴抽出のみメモリ 24GB 搭載 TITAN RTX,

表 2 各サブタスクにおける評価結果．表中丸括弧内の数字は成功回数/試行回数を示す．

手法	object grounding 精度 [%]	物体把持 成功率 [%]	物体配置 成功率 [%]
ルールベース理解	92.9 (13/14)	100.0 (13/13)	38.5 (5/13)
提案手法	7.1 (1/14)	100.0 (1/1)	0.0 (0/1)

“Go to the kitchen, put a bottle in the trash.”



“Go to the living room, put a soap on a toilet.”

図 5　CRT による指示文生成例

“Go to the living room, grasp the rubik’s cube and send it to the trash box for burnable.”

図 6 ルールベースで生成した指示文を用いた際の提案手法によるタスク実行例．図中丸数字はシーンの順番を示す．

256GBRAM，及び Intel Corei9 9820X を使用した．また，FCOG タスクの試行には，358 秒/回を要した．

6.　実験結果

6.1　定量的結果

表 2 に，ルールベースで作成した指示文を生成した利用した場合 [9] の性能を示す．表にはルールベースによる理解を利用した手法及び提案手法の各サブタスクにおける結果を示した．実験は 1 回行い，各数値はその結果を示す．ただし，移動タスクの結果については，既存手法から変更を行っていないため省略した．

6.2　定性的結果

図 5 に，CRT により生成した指示文についての定性的結果を示す．図中の各サンプルにおいて，左列及び右列はそれぞれ対象物体及び目標領域を撮影した画像である．また，赤色の領域及びオレンジ色の領域はそれぞれ対象領域および目標領域を示す．また，画像の下にある文が，各サンプルにおける生成文である．

上側のサンプルにおいて，対象物体及び目標領域はそれぞれ哺乳瓶及びゴミ箱である．このサンプルに対して，CRT の生成文は “put a bottle in the trash” であった．これより，CRT は対象物体及び目標領域について適切な表現を生成できたと考えられる．下側のサンプルにおいて，対象物体及び目標領域はそれぞれルービックキューブ及び椅子である．このサンプルに対して，CRT の生成文は “put a soap on a toilet” であり，対象物体及び目標領域ともに不適切な表現であった．この原因の一つとして，実験で利用した環境において，CRT を訓練したデータセットの中に出現しない物体が含まれていたことが挙げられる．例えば，ルービックキューブは ALFRED–fetch データセットに出現しない．これにより，CRT はルービックキューブについて適切に表現することができなかったと考えられる．

図 6 にルールベースで生成した指示文でタスクを行った際の，提案手法が成功した例を示す．この例において，指示文は “Go to the living room, grasp the rubik’s cube and send it ti the trash box for burnable” であった．ロボットはリビングルームに移動し，object grounding においてルービックキューブを特定し，そしてルービックキューブの把持に成功することが求められる．この例において，提案手法はこれらのサブタスクに成功した．

7.　結論

本論文では，fetch–and–carry タスクについての自然言語指示文が与えられたうえで，ロボットが視覚情報を基に指示を実行する，Fetch–and–Carry with Object Grounding (FCOG) タスクを扱った．本研究の貢献は以下である．

- Target–Dependent UNITER [1] を拡張し，対象物体候補及び目標領域候補について，複数の候補群から最適な候補の予測を行うことを可能にした．
- FCOG タスクにおいて，自動生成された指示文を利用することによって，フレームワーク全体の自動化を達成した．

謝辞

本研究の一部は，JSPS 科研費 20H04269，JST ムーンショット，NEDO の助成を受けて実施されたものである．

参　考　文　献

[1] S. Ishikawa and K. Sugiura, “Target-dependent UNITER: A Transformer-Based Multimodal Language Comprehension Model for Domestic Service Robots,” IEEE RA-L, vol.6, no.4, pp.8401–8408, 2021.

[2] S. Ishikawa and K. Sugiura, “Moment-based Adversarial Training for Embodied Language Comprehension,” ICPR, 2022 to appear.

[3] M. Shridhar, J. Thomason, D. Gordon, et al., “ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks,” CVPR, pp.10740–10749, 2020.

[4] M. Kambara and K. Sugiura, “Case Relation Transformer: A Crossmodal Language Generation Model for Fetching Instructions,” IEEE RA-L, vol.6, no.4, pp.8371–8378, 2021.

[5] S. Ren, K. He, R. Girshick, et al., “Faster R-CNN: Towards real-time object detection with region proposal networks,” IEEE Trans. PAMI, vol.39, no.6, pp.1137–1149, 2016.

[6] K. He, X. Zhang, S. Ren, et al., “Deep Residual Learning for Image Recognition,” CVPR, pp.770–778, 2016.

[7] A. Vaswani, N. Shazeer, N. Parmar, et al., “Attention Is All You Need,” NeurIPS, pp.5998–6008, 2017.

[8] 吉田悠，石川慎太朗，杉浦孔明，“生活支援ロボットによる物体操作タスクにおける Funnel UNITER に基づく指示文理解，” JSAI2022，p.2O4GS7，2022.

[9] H. Okada, T. Inamura, and K. Wada, “What competitions were conducted in the service categories of the world robot summit?,” AR, vol.33, no.17, pp.900–910, 2019.