モバイルマニピュレーションタスクにおける曖昧な指示文からの対象物体選択

森下雅晴 [1]　長野匡隼 [1]　中村友昭 [1]
[1] 電気通信大学
m1910655@edu.cc.uec.ac.jp

概要

ロボットが行う基本的なタスクの一つであるモバイルマニピュレーションタスクは，移動ロボットがユーザが指示した物体を把持し，別の場所へと持っていくタスクである．その際に，ユーザが必ずしも対象物体を明示的に発話しない場合があり，そのような曖昧な指示発話の理解が必要となる．本稿では曖昧なユーザ発話文に加えて環境に存在する物体の情報を入力とした言語モデル BERT に基づく系列ラベリングにより，ユーザの意図に沿った物体を選択する手法を提案する．実験では，小規模なデータセットを作成しモデルの学習を行い，提案手法が学習データに含まれない未知の曖昧なユーザ発話に対応できることを示す．

1　はじめに

近年，日常生活を支援する家庭用サービスロボットへの期待が高まっている．そのようなロボットの基本的なタスクの一つがモバイルマニピュレーションタスクである．モバイルマニピュレーションタスクは，移動ロボットがユーザが指示した対象物体を把持し，別の場所へと持っていくタスクである．このようなタスクにおいて，ユーザの指示発話はロボットが理解しやすい定型文ではなく，日常的に使用している自然言語を用いる方がユーザにとっては望ましい．しかしそのような指示発話では，ユーザが必ずしも対象物体を明示的に発話しない場合があり，そのような曖昧な指示発話の理解が必要となる．

そこで，本稿では物体名が明示的に含まれないユーザの発話から，環境に存在する物体の情報を利用することで，対象物体を特定する手法を提案する．図 1 が想定するタスクであり，ユーザが環境の中にある物体を持ってくることをロボットに要求し，ロボットはその発話と環境にある物体から対象物体を特定しユーザに届ける．このようなタスクにおいて，対象物体を選択するためには，ユーザの発話と要求される物体の関係を正しく理解する必要がある．

図 1　想定しているタスクの例

提案手法では物体のリストを文字列として学習済み言語モデル BERT (Bidirectional Encoder Representations from Transformers)[1] に入力し，リスト内の各物体が対象物体であるか否かを表すラベル列を予測する系列ラベリング問題として扱う．これにより，ユーザが対象物体を明示的に発話しない場合や，学習データには含まれない新規の物体が存在する場合であっても対象物体を特定することができる．

関連研究としてユーザが明示的に対象物体を発話することを想定したモバイルマニピュレーションタスク [2] があるが，本手法のように曖昧な指示文を想定していない．また，学習済み言語モデルを用いて，行動や物体を特定する手法 [3][4] も提案されているが，これらの手法では新規物体を扱うためには，再学習（ファインチューニング）が必要となる．一方，本手法では，物体選択を系列ラベリング問題

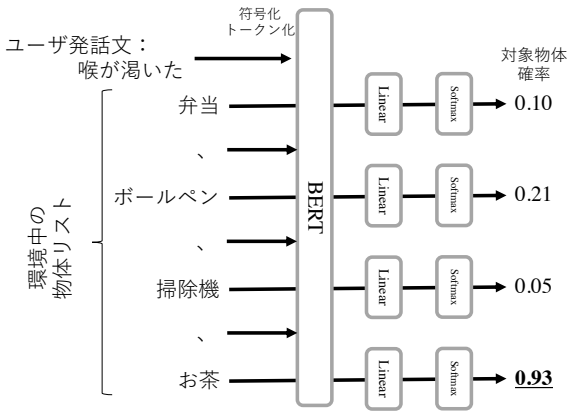

図 2　提案手法の概要

とすることで，新規物体に対しても再学習することなく，対応することができる．

2　提案手法

提案手法の概要を図 2 に示す．物体リストと発話文を BERT への入力とし，BERT からの出力を線形変換，softmax 関数を通して，各物体が対象物体であるかどうかを表す確率を出力する．

BERT[1] は Transformers の encoder を使用した言語モデルであり，本稿では事前学習済みの BERT をファインチューニングして使用した．事前学習済みのモデルを利用することで，言語モデル内で学習されている常識的な知識により，指示文と対象物体の関係を捉えることが期待できる．

提案手法では BERT の出力を利用した系列ラベリングにより物体を選択する．一方，分類タスクにより物体の選択をすることも可能であるが，対象物体が複数になった場合への対応が難しいといった問題がある．また分類タスクでは，分類対象の物体数だけの出力ノードを用いて学習するため，事前に決めた物体のみにしか対応できず，新規物体へ対応するためには出力ノード数を変え再学習する必要があるといった問題がある．これに対し，系列ラベリングではリスト内の各物体に対して対象物体か否かのラベルを出力するため，複数の物体を選択することができ，言語モデルに含まれる単語であれば新規物体であっても対応することができる．モデルは，IO 法に基づいた対象物体を表すラベル（“I”）と非対象物体を表すラベル（“O”）を出力するよう学習する．これにより，「オレンジジュース」のように「オレンジ」と「ジュース」という 2 つのトークンに分かれてしまう単語であっても，それぞれ “I” が出力されるため，トークンを結合して「オレンジジュース」という単語を選択することができる．なお，物体リストは「、」で区切ることで，BIO 法を使わずとも異なる単語を結合してしまうことはない．

学習時には，出力された対象物体/非対称物体である確率と正解ラベルとのクロスエントロピーを損失関数として用い学習する．テスト時には，それぞれのラベルの確率を出力し，確率が高いラベルを予測ラベルとする．

本稿では東北大学で作成された事前学習済みモデル [5] をファインチューニングして使用した．このモデルでは日本語版ウィキペディアを用いて事前学習されている．ファインチューニングに使用したデータセットについては 3.1 節で述べる．

表 1　発話文の種類

ユーザの要求	物体例	発話文の例
食事	弁当, お茶	ご飯の用意をして
食べ物	弁当	食べ物を持ってきて
飲み物	お茶	飲み物を持ってきて
遊具	ボール	遊び道具が欲しい
筆記用具	ペン	筆記用具が欲しい
書物	本	何か読みたい

3　実験

本実験では，食べ物・飲み物・遊具・筆記用具・書物，また食べ物と飲み物で構成される食事，に関連する物体を選択することを想定し，小規模なデータセットを作成してファインチューニングを行い，提案手法の有効性を確認した．また，より曖昧な表現を含む発話文を用いた場合の結果も確認した．

3.1　データセット

モデルのファインチューニングに用いるデータセットとして，ユーザの発話文と物体のリストの組を作成した．まずユーザの発話文として具体的な物体名を含まない「食べ物を持ってきて」，「食事を用意して」といったロボットにユーザの要求に沿った物体を持ってくることを指示する文を 50 個用意した．今回想定した，発話文の種類と正解となる対象物体の例を表 1 に示す．この 50 文を学習用 40 文，テスト用 10 文に分割した．すなわち，テスト用の文は学習用とは異なる文で構成されている．

次に，学習用物体リストとテスト用物体リストの

表 2　「食べ物」に対応した物体リスト

学習用物体リスト	テスト用物体リスト
パン，カップ麺，米， ラーメン，白ごはん， フライドチキン， パスタ，スープ， 味噌汁，うどん	白米，お弁当， インスタントラーメン， 麺，そば，食パン， カップ焼きそば，弁当， サラダチキン，メロンパン

表 3　テストデータによる物体選択の評価

Precision	0.81
Recall	0.78
F-measure	0.79

中から 6〜7 個の物体をランダムに取り出し，入力物体リストとした．この入力物体リストはユーザ発話文に合わせて正解となる物体を 1 つ以上含み，その他は正解ではない物体で構成されている．この時，リスト中の単語の位置は物体選択に関係ないため，単語の並びがランダムになるよう留意した．また，学習用物体リストとテスト用物体リストはそれぞれ異なる物体で構成されている．表 2 が，食べ物に対応した学習用とテスト用の物体リストである．また同時に，IO 法によって入力物体リスト内の各物体に対応した正解ラベルを作成した．この手順を学習用とテスト用それぞれのユーザ発話文に対して 200 回行い，合計で学習用 8000 個，テスト用 2000 個のデータセットを作成した．

このように，テストデータには学習データには含まれない物体，発話文を使用した．すなわち，テストデータで正しく物体を選択するためには，発話文と物体の関係を正しく学習する必要がある．

3.2　テストデータによる評価

表 3 が，テストデータを用いた物体選択の Precision, Recall, F-measure である．テストデータは学習データに含まれていない物体と発話文で構成されていたが，各指標で 0.7 を超えるスコアを得られた．すなわち，提案手法は学習していない未知の表現・物体に対しても，ある程度対応できると考えられる．

3.3　より曖昧な発話文による物体選択

学習データは「○○ を持ってきて」，「○○ が欲しい」などの物を要求する文で構成されている．本実験では，直接的に物を要求しない，より抽象的な表現から物体を選択できるか確認した．学習に使用した表現と比べてより曖昧で抽象的な表現を含む発話文を入力としたときの結果が，表 4 である．1 つ目では曖昧な表現として「何か飲みたいな」という発話から，2 つ目の例では「ちょっとメモしたいな」という発話から物体を選択した．どちらの場合も物体リスト内の太字で示した正解となる対象物体を正しく選択できている．この結果は，事前学習済みの言語モデルにより，ファインチューニング時にはない「飲む」や「メモ」といった単語と各物体の関連性を捉えられているためだと考えられる．

表 4　より曖昧な発話文からの物体選択

物体リスト	発話文	選択結果
オレンジジュース，鉛筆， カレー，電話，消しゴム， ボール，**水**，ペン	何か飲みたいな	オレンジジュース， 水
カップ麺，**ペン**，服， ボール，新聞紙，水	ちょっとメモしたいな	ペン

4　おわりに

本稿では，家庭用サービスロボットがユーザの曖昧な発話から，環境内にある物体をユーザの意図に合わせて選択する手法を提案した．実験では，小規模なデータセットを用いた予備的な実験であったにも関わらず，未知の表現・物体であっても正しく物体選択できることを確認した．データセットを拡張することで，より曖昧な指示から物体を選択できるようになると考えられる．そこで今後，クラウドソーシングを利用してデータセットを拡充し，より曖昧な指示やより広いドメインへの対応を考えている．

また，現時点では複数の正解が物体リストにあるとき，それら全てが抽出されるようになっている．そのような場合には，ユーザとの対話を通して物体を選択するような方法を導入することを考えている．

謝辞

本研究は，JST ムーンショット型研究開発事業 JPMJMS2011 の支援を受けたものである．

参考文献

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language” arXiv:1810.04805, 2019

[2] Takahiro Kobori, Tomoaki Nakamura, Mikio Nakano, Takayuki Nagai, Naoto Iwahashi, Kotaro Funakoshi, and Masahide Kaneko, “Robust Comprehension of Natural Language Instructions by a Domestic Service Robot”, Advanced Robotics, Vol. 30, Issue 24, pp. 1530-1543, 2016

[3] 田中翔平，湯口彰重，河野誠也，中村哲，吉野幸一

郎，“気の利いた家庭内ロボット開発のための曖昧なユーザ要求と周囲の状況の収集”，情報処理学会研究報告，NL-253, 2022
[4] Michael Ahn, et al., “Do as I can, not as I say: Grounding language in robotic affordances.” arXiv preprint arXiv:2204.01691, 2022
[5] “Pretrained Japanese BERT models”, https://github.com/cl-tohoku/bert-japanese