実世界におけるインタラクティブな物体指示

羽鳥潤* 菊池悠太* 小林颯介* 高橋城志* 坪井祐太* 海野裕也* Wilson Ko Jethro Tan
Preferred Networks, Inc.
{hatori,kikuchi,sosk,takahashi,tsuboi,unno,wko,jettan}@preferred.jp

1 実世界における言語処理

これまでの自然言語処理は，主にソフトウェアやインターネット上の一種の仮想空間を対象として発展してきた．しかし，人工知能技術やロボットが社会に浸透し始めている今日，実世界において必要とされている言語処理技術はどのようなものだろうか？

本稿では，現実の物理世界において作業をするロボットに対して，人間同士のコミュニケーションと同じように指示することを考える．人間に作業指示を与えることを考えると，最初の大きな障壁となるのは，環境中に存在する様々な物体の中から作業対象となる物体を指し示す部分である．生活環境には，様々な物体が存在しており，また，同一の物体が多数あるかもしれない．こうした環境で意図通りに物体を指示するためには，物体の様態・形質・位置などに関する様々な言語表現を駆使するのと同時に，インタラクティブな双方向のやりとりによって正確かつ効率的に意思疎通する必要がある．しかし，ロボット工学の領域においてはそのような言語指示の問題は制限付き言語や単純化された環境を対象に行われるのが一般的であり [1–4]，自然言語処理の領域においても，言語の意味理解に関連する研究は質問応答に代表されるように，現実の物理世界から切り離された形で行われることが多かった [5]．

本研究では，現実の物理世界における物体の指示・作業の実行という文脈において，重要となる言語処理技術，必要なタスク設定とデータセット，そして，解かなければならない課題について考察したいと思う．

前述の通り，現実世界における言語指示理解においては，言語表現の多様性と曖昧性が大きな問題となる．例えば，図 1 の例文「ねえ，そこの茶色のふわふわの物体を右下に入れて．」のような抽象的・感覚的表現は人間同士のコミュニケーションにおいて頻出するが，既存のオントロジーなどを利用してあらゆる表現をカバーすることは難しい．また，図 1 では実際には「茶色のふわふわの物体」が 2 つ存在している．人間同士がこのような複雑な環境中でコミュニケーションをとる場合を考えると，1 回の発話で対象が一意に定まるようにまわりくどい表現を盛り込む（例えば「右上の箱の隅っこにある逆さのクマのぬいぐるみ」）よりは，

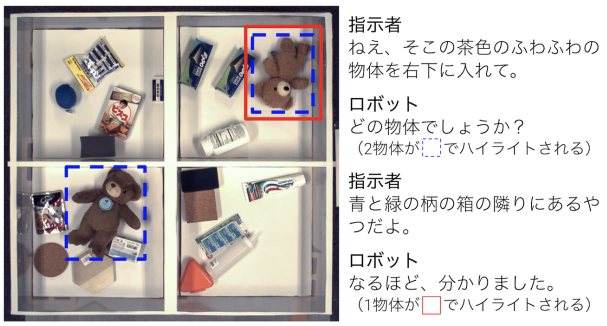

図1: 提案手法の概要．言語指示に曖昧性がある時には，聞き返しによって対象の物体をひとつに絞り込む．

インタラクティブなコミュニケーションを前提に情報を少しずつ追加していくことの方が多いかもしれない．また，そもそも自分の発話に解釈の曖昧性があることが，相手からのフィードバックによって初めて明らかになることも多いだろう．

本研究では，上述のような多様な実世界環境と言語表現に対応するため，深層学習をベースとした物体認識モデル [6,7] と参照表現解析モデル [8,9] を組み合わせて用いる．また，言語指示の曖昧性に対処するため，図 1 に描かれているような言語的・視覚的フィードバックを用いて，言語指示・対象物体の曖昧性を対話的に解決する方法を提案し，音声によるロボット実機への指示を行う実験により，聞き返しによって対象物体指示の精度が向上することを示す．

言語指示の曖昧性解消の問題を扱った研究としては [3] などがあるが，実験は少数のラベル付き物体を用いた単純な環境で行われており，本稿で扱うような現実に即した設定で行われたものではなかった．視覚情報と言語情報を統合的に処理する研究は，特に近年の物体認識モデル [6] の性能向上に伴い，大きく注目されるようになった．画像説明文付与 [10]，参照表現解析 [8]，画像質問応答 [11]，画像付対話 [12] など様々なタスクが対象とされてきているが，これらはいずれも現実の物理世界に干渉して作業を行うことを目的としていない．物理世界における物体操作指示のタスクにおいて，制約のない生の口語表現を扱った研究としては，我々の提案手法が最初の統合的な手法と言える．

2 言語指示による物体移動タスク

本研究では，人間の指示者が，自然言語の音声入力を通じてロボットにタスクを指示し，制御することを目

* 最初の 6 人は全員筆頭著者であり貢献度に差はない

図2: 平均的な実験環境（左）と散らかった環境（右）の例．箱は 400mm×405mm である．

指す．まず，図 2 で示したような 4 つの箱を用意し，その中に様々な日用品が散りばめられた環境を作る．指示者は，ロボットに特定の一つの物体をつかみ，他の箱に移動するように指示することができる（例：「黄色いマスタードの容器を下に移して」）．

この際，指示の解釈に曖昧性があり対象の物体を一意に特定できない場合には，その曖昧性を解消するために対話的な聞き返しを用いる．例えば，図 2 の左の図では「黄色いマスタードの容器」が 2 つ作業空間中に存在しているので，システムは指示者に「どちらの物体でしょうか?」と聞き返す．これに対して，指示者は自然な口語指示を用いて意図通りの物体をつかむよう，追加の指示を与えることができる．図 2 の例では，「右の箱に入ってるやつをお願い」「熊さんと同じ箱の方」「壁際にある方です」といった具合である．

我々が実験に用いるデータセットも，現実的かつ認識・理解の点からは挑戦的な設定で作られている．データ中には 100 種類以上の物体が含まれており，一般的な特定の名称を持たないため，指示者が間接的・抽象的な表現を用いなければならないものが多く含まれている．更に，22 個の未知物体を評価時に用いることで，システムがどれだけ未見の物体に対して汎化して動作するか確認する．また，データセット中には，物体が比較的秩序立って配置された環境とひどく散らかった環境の両方が含まれている．散らかった環境では，多数の物体同士が重なっているため，物体認識の難易度が高くなる．一方で，比較的整理された環境においても，同一または類似の物体が複数存在するため，指示者が物体の色や形などの簡単な言語表現に頼ることができず，物体を一意に指示するために複雑な指示（例えば，他の物体からの相対位置，順序・状態・置かれ方などの表現）が必要となるケースも多い．

3 提案手法

移動対象物体と移動先の指示文の解析は 4 つのサブタスクに分けられる．各サブタスクと用いた手法を以下で述べる．人間の音声発話を入力とし，物体を示す領域を出力とする提案システム全体図を図 3 に示す．

■物体検出 指示の参照対象となる物体の情報を個別に解析するため，物体群を撮影した画像について，各物体の矩形領域を求める物体検出を行う．本研究では，

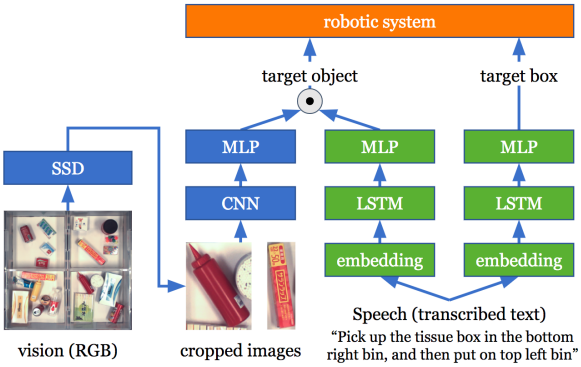

図3: 全体のネットワーク構成．入力画像は SSD によって検出した物体ごとに切り抜き，CNN によって特徴情報を抽出し，言語指示は LSTM によって同様に特徴情報を抽出する．両者を使って対象物と移動先を推定する．

畳み込みニューラルネットワーク（CNN）に基づく物体検出モデルである Single Shot Multibox Detector (SSD) [7] を改良して用いた．SSD は，画像を細分化する大量の矩形領域についてスコアを算出し，高いスコアを持つ領域のみを選択的に出力する．

■物体選択 物体検出モデルの出力領域のそれぞれが実世界の物体に 1 対 1 で対応付いていると仮定する．その領域の集合から，指示文の参照物を含むような領域を一つ選ぶことを目指す．この物体選択モデルとして，Yu ら [9] の参照表現解析モデルを改良して用いた．

実世界の各物体の情報は，以下の種類の情報を結合し，多層パーセプトロン（MLP）でエンコードした出力として表現する．

- 視覚情報ベクトル: 領域の画像及び全体画像を CNN を用いてエンコードして結合する
- 領域素性ベクトル: 領域の縦位置・横位置・面積を全体画像に対する比率で表す
- 比較ベクトル: 視覚情報ベクトルと領域素性ベクトルそれぞれについて，他の候補の同ベクトルとの差ベクトル集合を求め，max/min/average pooling を適用し結合する

一方，言語指示文の情報は，文を分かち書きした後に単語埋め込みベクトルと LSTM 及び MLP でエンコードを行う．そして，各物体の参照先としての当てはまりの良さを表すスコアを，各物体情報と言語情報のベクトルのコサイン類似度によって計算する．予測には，類似度最大となる物体を選択する．モデルの訓練では，正解事例のペア（物体と指示文）から計算されるスコアが，乱択で作成した負例のペアから計算されるスコアよりも，十分なマージン以上に離れるように損失関数を設計した．

■移動先予測 「何を」「どこに」動かすかを予測する必要があるが，今回は物体の予測選択を確定させた後に，移動先予測問題を解くことにした．まず，言語指示文を単語埋め込みベクトルと LSTM によって文ベクトルへとエンコードする．そして，先に予測された物体の矩形領域について，領域の全体画像内での縦位置

・横位置を素性として，文ベクトルに結合し，それを入力とする MLP により予測を行う．今回のタスクでは，移動先は 4 種類の区画に限られるため，MLP の最終層ではソフトマックス層を用いて 4 値分類の確率分布を算出した．モデルの訓練時には，正解の物体領域で素性を取り出して学習を行った．

■指示の曖昧性判定と解消　章 1 で述べたように，人間からの指示文は，本質的に不明瞭で曖昧にもなりうる．また，実世界で物理的な作業を行うシステムにおいては，意図しない行動による大幅な時間経過・物体破損・人身事故の可能性をはらむため，誤った判断による損害が大きい．そのため，本研究では指示の解析に関して確信度を測定し，十分に予測の確信度が大きい場合にはそのまま物理的行動に移り，確信度が小さい場合には図 1 のように人間に対して追加の情報提供を求めるフィードバックを行うようにした．

今回は単純な方法として，1 位と 2 位の物体のスコアの差が, 一定の閾値以上の場合に「確信が大きく」, 閾値未満のときに「確信が小さい」とする．なお，移動先の予測については，指示の曖昧性が低く，また予測精度も十分に高かったため，今回は確信度に関わる処理を行っていない．また，確信度が小さくフィードバック行動を行う際には，上記の閾値差の範囲内に含まれるスコアを持つ物体を UI 上でハイライトして表示することで，指示者がより自然に追加の情報提供を行えるようにした．閾値は，モデル学習時に用いた正例と負例の間のマージンの値をそのまま用いた．

4　実験

4.1　学習データセット

学習用のデータセットは人手で作成した（今後公開予定）．およそ 100 種類の日用品を用意し，ランダムにおよそ 20 個ずつ選び，4 つの箱に散りばめた画像を収集した．各画像中の物体を，別の箱に移動する指示をアノテートした．この際，なるべく自然な指示文になるように指示した．この中には名前のわかりにくいものも含まれているため，単に物体の名前を示すだけではなく説明的に指示をする必要も出てくる．指示文の多様性を持たせるために，各画像につき最低 3 人のアノテータに作業を依頼した．全体では 1,180 画像，のべ 25,883 物体，77,700 文が収集された．このうちの 76,551 文（25,500 物体）を訓練データ，1149 文（383 物体）を開発データとして使用した．

4.2　実験環境と設定

実環境における実験のため，図 4 のような Fanuc 社製ロボットアーム M10*i*A を用いた．様々な物体の把持を可能とするため，ロボットアームの先端にはバキュームグリッパを取り付けた．物体認識器のための画像と点群の撮影には，それぞれ Ensenso N35 ステレオカメラと IDS uEye RGB カメラを用いた．Google Chrome の Web Speech API の音声認識システムを利用して，音声からの書き起こし文字列を言語指示文とした．なお，実験結果に音声認識誤りの場合は含めていない．実装には Chainer [13] を利用し，SSD の実装は，ChainerCV [14] のサンプルを元にした．

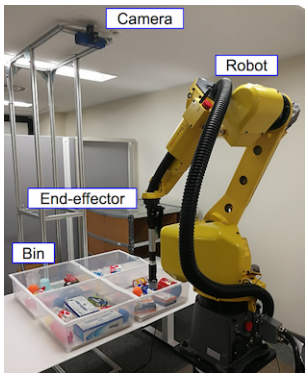

図**4:** ロボットを使った実験環境

1. “左上の茶色い丸いもの 1. “右上の箱に入ってる箱
を右に移動させて下さい” を左に”
2. “左上の缶を移動させて 2. “小さい方です”
下さい”

図**5:** 実環境における本システムの試行例．図中の赤い実線は，本システムが対象物体であると予測した物体である．青い破線は，1 つ目の指示に曖昧性があり追加の指示がを要求した際の対象物体の候補となっている物体を示している．小さな緑の円は，指示者が意図した真の対象物体を示している．

4.3　実環境における評価実験

評価環境として，作成したデータセット（図 2）と同様の環境を用意した．5 名の日本語母語話者が，環境中の 1 つの移動対象物体とその移動先を口頭でロボットに指示する実験を行った．この際，人間に対してお願いするように指示してもらった．本実験では指示が曖昧で一意に対象物体を同定できないとシステムが判定した場合には，システムは一度だけ追加の指示を請うことができる．実験のはじめに 16 個の物体を無作為に選び環境中に設置した．評価環境の多様性を担保するため，5 回の試行ののち新たに 6 個の物体を他の物体と入れ替え全物体の場所も無作為に入れ替えた．5 名の指示者にはおよそ 25 回の試行を行い，総試行回数は 126 回であった．126 試行のうち 86 試行については，訓練データに一度以上出現した既知物体のみで実験が行われ，残りの 40 試行は 25% の物体を訓練データに一度も出現していない未知物体に置き換えて行われた．

表 1 に実験結果を示す．指示者の意図通りの移動先の箱を選択できた率（精度）は 90.5%, 対象物体を選べた精度は 84.1% であった．訓練データに含まれていな

	移動先精度 (聞き返しあり)	対象物体精度 (聞き返しあり)	対象物体精度 (聞き返しなし)	聞き返し 比率
既知物体のみ	89.5% (77/86)	90.7% (78/86)	88.4% (76/86)	22.1% (19/86)
未知物体あり	92.5% (37/40)	70.0% (28/40)	65.0% (26/40)	57.5% (23/40)
総合	90.5% (114/126)	84.1% (106/126)	81.0% (102/126)	33.3% (42/126)

表**1:** ロボットアームを使った音声による物体移動タスクの実験結果.

い未知物体が配置された場合では対象物体精度は 9 割から 7 割程度に低下したが，移動先精度は変わらなかった．なお，システムが曖昧と判定した率も未知物体がある場合に増えた．また，移動先精度（聞き返しあり）84.1% と移動先精度（聞き返しなし）81.0% との比較から対象物体の聞き返しの効果も確認できた．

図 5 に，実験中に出現した実例を示す．どちらも，一度目の指示をシステムが曖昧と判断し，追加の指示を要求したものである．それぞれ“左上”や“箱”など，意図しない曖昧な指示に対し聞き返すことで正しい物体を選択することに成功している．このような事例は他にもあり，指示の厳密性を要求するよりも，このように能動的に確信度を向上させる技術は重要である．

その他，把持に成功した事例から次のような傾向が見られた．

- 「黒い丸いもの」など，形状や色，サイズなどの概念による指示の認識.[*1]
- 「斜め下に持って行って」「右に動かして」「割り箸の下にあるものを上に持っていってください」など物体の相対的な位置関係の考慮.

このように，物体そのものの一般的な名称に限らない様々な指示形態の認識や，追加指示の要求による確信度の向上など, 実世界での作業において重要であるとかんがえられる能力が獲得できていることが確認できた．

5 結論

本稿では，現実世界における制限なし音声言語指示に対し，人間の指示者の意図を聞き返しによって正確かつ効率的に理解することができるシステムを提案した．ロボットを用いた物体移動タスクの評価において，対象物体の選択で 84.1%，移動先の選択で 90.5% の精度を達成した．また，聞き返しによって対象物体認識の誤りを削減できることを示した．

ロボットに対するインターフェースとしての言語は今後ますます重要になってくるだろう．スマートフォンやスマートスピーカーといった機器が主に情報端末としての役割を果たすのと比較して, ロボットは物理的に環境に干渉するという本質的に大きく異なる特徴がある．そのため，ロボットのための自然言語処理はこれまで取り組んでこなかった新しい課題の宝庫であり，魅力的な研究分野であると考える．

[*1] 指示文中の物体の表現と画像中の領域との直接の対応は教師情報として与えていない.

参考文献

[1] R. Paul, *et al.*, “Efficient grounding of abstract spatial concepts for natural language interaction with robot manipulators.” in *Robotics: Science and Systems*, 2016.

[2] M. Shridhar and D. Hsu, “Grounding spatio-semantic referring expressions for human-robot interaction,” *arXiv preprint arXiv:1707.05720*, 2017.

[3] D. Whitney, *et al.*, “Reducing errors in object-fetching interactions through social feedback,” in *IEEE International Conference on Robotics and Automation.*, 2017, pp. 1006–1013.

[4] B.-A. Dang-Vu, *et al.*, “Interpreting manipulation actions: From language to execution,” in *Robot 2015: Second Iberian Robotics Conference*, vol. 417. Springer, 2016, pp. 175–187.

[5] D. Ferrucci, *et al.*, “Building Watson: An overview of the DeepQA project,” *AI Magazine*, vol. 31, no. 3, pp. 59–79, 2010. [Online]. Available: http://www.aaai.org/ojs/index.php/aimagazine/article/view/2303

[6] K. He, *et al.*, “Deep residual learning for image recognition,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

[7] W. Liu, *et al.*, “SSD: Single shot multibox detector,” in *European Conference on Computer Vision*, 2016.

[8] L. Yu, *et al.*, “Modeling context in referring expressions,” in *European Conference on Computer Vision*, 2016.

[9] Y. Licheng, *et al.*, “A joint speaker-listener-reinforcer model for referring expressions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[10] O. Vinyals, *et al.*, “Show and tell: A neural image caption generator,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3156–3164.

[11] S. Antol, *et al.*, “Vqa: Visual question answering,” in *IEEE International Conference on Computer Vision.*, December 2015.

[12] A. Das, *et al.*, “Visual Dialog,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] S. Tokui, *et al.*, “Chainer: a next-generation open source framework for deep learning,” in *Workshop on machine learning systems on Neural Information Processing Systems*, 2015.

[14] Y. Niitani, *et al.*, “ChainerCV: a library for deep learning in computer vision,” in *Proceedings of ACM Multimedia Workshop*, 2017.