実世界での話し言葉指示による物体移動: 深層学習による画像・言語理解

○高橋 城志 * (PFN)　羽鳥 潤 * (PFN) 菊池 悠太 * (PFN)
小林 颯介 * (PFN) 坪井 祐太 * (PFN) 海野 裕也 * (PFN)
中島 統太郎 (PFN) 福田 昌昭 (PFN) Wilson Ko(PFN) Jethro Tan(PFN)

1. 実世界における話し言葉を用いた HRI

誰もが容易にロボットに指示を出せることは，社会への普及に欠かせない要素の 1 つである．指示方法にはプログラミング，タッチパネル，ジェスチャーなどが用いられてきたが，話し言葉による指示が最も直感的で様々な指示が可能である．しかし，ロボティクス分野における言語指示は，単純化された環境を対象とし，制限された単語や文法のみを扱うのが一般的であり [1, 2, 3, 4]，実用化には解決するべき課題が多い．

人間同士のコミュニケーションのように，話し言葉でロボットに指示するには主に多様性と曖昧性の 2 つの課題がある．例えば，図 1 の例文では，「ねぇ，そこのフワフワした茶色のものを右下の箱に入れて」とあるが，“フワフワした茶色のもの” 以外にも同じ物体を指すのに，“クマのぬいぐるみ”，“フワフワしたやつ”，“テディー・ベア” など多様な表現が考えられる（多様性）．動作の指示においても，“掴んで”，“動かしてください”，など表現方法は複数存在する．また，図 1 では，”フワフワした茶色のもの”に相当するぬいぐるみが 2 つ存在しており，この指示文だけではどちらを示しているのかを断定することはできないため，指示者に確認する必要がある（曖昧性）．

多様性の課題を解決するために，実環境における視覚情報とそれに対応した様々な言語表現を集めた訓練データ[1] を構築し，深層学習による物体認識と参照表現解析の組み合わせを適用した[2]．また，曖昧性に対応するため，音声によるロボットからの聞き返し，及び，拡張現実（Augmented Reality; AR）技術を用いた言語と視覚のフィードバックを用いる．

2. 言語指示による物体移動タスク

本研究では，人の指示者が音声入力を通じてロボットに物体移動を指示し，制御することを目指す．図 1 で示した 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．人はロボットに一つの物体を他の箱に箱に移動するように指示する（図 1 の指示例参照）．その指示に曖昧性がある場合，すなわち，1 つの物体に絞れない場合，ロボットから指示者に対して，「どの物体でしょうか」と聞き返しながら対象物体を視覚的に提示できるものとする．それに対して，「白板消しの箱の隣の方」と指示者が追加の指示により，ロボットは対象としている物体を絞ることができる．

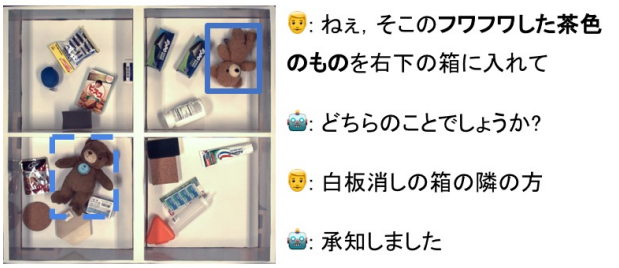

図 1　ロボット指示の多様性と曖昧性の例

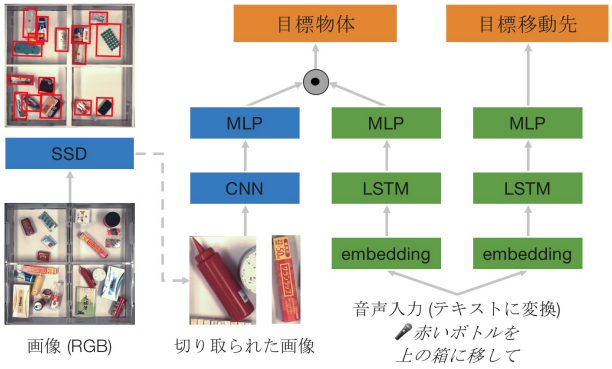

図 2　深層学習を用いた目標物体と移動先の検出

3. 提案手法

End-to-end の深層学習を用いた画像・言語理解により，単語や文法の事前設定，及び，ルールを必要としないモデルを提案する．本モデルの特徴として，言語による物体名とその物体の画像の関係性を学習するだけではなく，物体の色，質感，大きさ，位置関係などと言語表現とのマッピングも学習している．

提案手法は以下の順で処理を行う（図 2,3）．

1. 画像中の物体検出
2. 検出した物体と指示文からの目標物体選択
3. 指示文からの目標移動先の選択
4. 指示文の曖昧性の有無を検出
5. 曖昧性の提示と解消

3.1 物体候補検出

画像中にある物体検出には，CNN に基づいた Single Shot Multibox Detector(SSD) [5] を改良して用いた．SSD は，各物体の矩形領域を求めるものである．大量

*最初の 6 人は全員筆頭著者であり貢献度に差はない

[1]英語と日本語の訓練データは Creative Commons Attribution 4.0 ライセンスで公開： https://github.com/pfnet-research/picking-instruction

[2]動画: https://www.youtube.com/watch?v=DGJazkywOWs (英語), https://www.youtube.com/watch?v=6ei_Dn-Uxqs (日本語)

の矩形領域から高いスコアの矩形領域のみ出力される．訓練データ中にない未知の物体でも検出可能である．

3.2 目標物体選択

検出された物体中から言語指示に該当する物体の選択を行うため，Listener モデル [6] を参考にした．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）で表現ベクトルに変換する．

- 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のそれぞれについて，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，言語指示は LSTM と MLP で表現ベクトルを得る，最後に，目標物体候補を各物体と言語指示の表現ベクトルのコサイン類似度によって計算し，スコアが最も高いものを目標物体として選択する．

3.3 目標移動先選択

物体をどの箱に動かすかを選択するために，LSTM と MLP で表現された言語指示に対して，softmax 関数を用いて最も確率が高い箱を目標移動先とした．

3.4 曖昧性判断

指示者からの指示が曖昧な場合，それを判断して聞き返す必要がある．本研究では，目標物体のスコアの大きさで聞き返しを行うか判断する手法を提案する．具体的には，検出された物体候補群のそれぞれに対して，指示文とのスコアを算出し，スコアが最も高い物体とその次に高い物体とを比較したときに，その差が閾値 m を超えた場合に曖昧であると判断して聞き返す（図 3）．

適切なタイミングで聞き返す場合には閾値 m の決定が重要になる．そこで， 正解の組と不正解の組のスコアが m 以上になるようにスコア関数を学習し，この m を聞き返しの判断に用いた．具体的には，正解事例の組（物体と指示文）から計算されるスコアが，乱択した負例の組から計算されるスコアよりも，m 以上に離れるように損失関数 L を設計し，スコア関数を訓練した．

$$L(\theta) = \mathbb{E}_{q,o}[\max\{0, m - f_\theta(q,o) + f_\theta(q,\hat{o})\} + \max\{0, m - f_\theta(q,o) + f_\theta(\hat{q},o)\}].$$

ただし，θ はスコア関数 f_θ のパラメータ，q と $\hat{q}$ はそれぞれ正解指示文と不正解指示文, o と $\hat{o}$ はそれぞれ正解物体・不正解物体を示す．

3.5 曖昧性の提示と解消

聞き返しが必要な場合，ディスプレイ，および拡張現実（AR）技術を用いて曖昧性のある物体に対して重ねて表示することで，指示者に直感的に分かるように提示した．AR による提示方法としては，図 4 左図に示すように，光の円で対象物体を囲んで点滅させる映像をヘッドマウントディスプレイ（Microsoft HoloLens）で表示した．人からの返答は元の指示文と結合して処理し，再度目的物体のスコアを計算した．また，聞き返す必要なしと判断した場合も，システムの判断をユーザに提示するために，図 4 右図のように目標物体と移動先を光の円と矢印で示した．

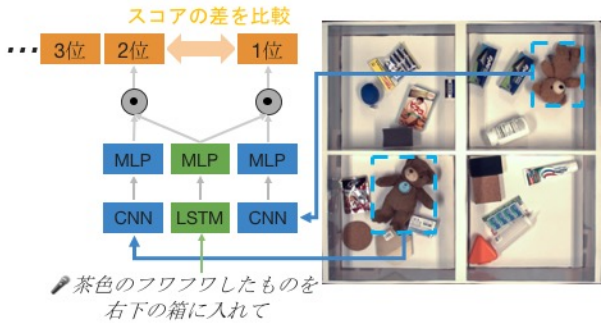

図 3 曖昧性判断の例：スコアの上位 2 物体の差が閾値以下を曖昧とした

図 4 AR による曖昧物体の提示．左図：曖昧性の提示（緑色と白色のボンドの曖昧性を表現）．右図：目標物体（白色のボンド）と移動先の提示．

4. 実験

4.1 訓練データ

図 1 に示すように，100 種類以上の日用品の中からランダムに取り出した 20〜30 種類の物体を 4 つの箱の中に散りばめた訓練データを新たに作成した．これらの物体の中には，同一，または，類似の物体は複数存在しているため，指示者は物体の形状，色，位置関係を含んだ様々な表現で指示を出すことになる．さらに，特定の名称を持たないもの，または，広く知られていないものが含まれるため，間接的・抽象的な表現を用いなければいけないものが複数存在する．物体の配置のされ方として，多数の物体が重なった状態と，比較的整理された状態の両方が含まれている．

訓練データ数は，全体で 1,180 画像，のべ 25,883 物体，英語が 91,590 文，日本語が 77,770 文含まれている．クラウドソーシングを利用し，各画像に対して最低 3 人の作業者により言語指示の文章が付与されている，作業者に対しては，対象とする物体を 1 つに特定でき，かつ人と話す時のような表現で指示文の作成を依頼している．なお，指示文に曖昧性がないかどうかの確認はしていない．

4.2 実験環境

評価実験用のハードウェアとして，FANUC 社のロボットアーム M10iA に，様々な物体の把持が可能なバキュームグリッパを取り付けたものを用いた（図 5）．物体認識の際の点群と画像には ENsenso N35 ステレオカメラと IDS uEye RGB カメラを用いた．なお，実装には Chainer [7] と ChainerCV [8] を利用した．音声

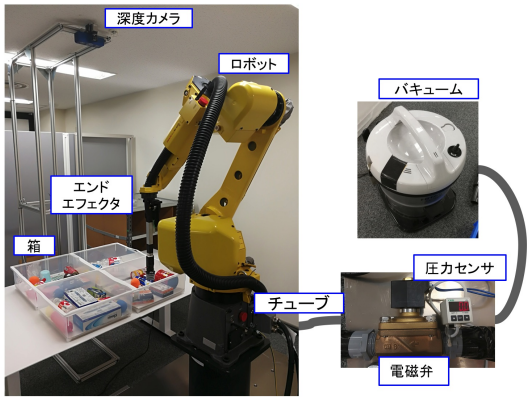

図 5　ロボットハードウェア概略

	聞き返しなし	聞き返しあり
英語指示	88.0%	92.7%
日本語指示	81.0%	84.1%

表 1　物体移動タスクの対象物体精度

認識には Google Chrome のウェブスピーチ API を用いた．

4.3　実験結果

表 1 に英語と日本語の指示での，聞き返しあり/なしのときの正答率を示す．実験には学習に使われなかった未知物体と未知語も含まれている．本実験ではシステムが曖昧だと判断したとき，1 度だけ追加指示を行うという条件で行った．その結果，英語指示では 4.7 ポイントの向上（39.2%の誤り削減），日本語指示では 3.1 ポイントの向上（16.3%の誤り削減）を実現した．

また，定量的な評価ではないが，文献 [9] と同様に，ヘッドマウントディスプレイによる曖昧性提示については，見えない部分も視点を変えて確認できる効果と，視野角の狭さによる制限などの課題が報告された．

追加実験として，箱の中の物体数を約 70 個まで増加させて実験を行った（図 6）．訓練データには平均約 20 個の物体が配置されていたが，物体が多い状況でも物体検出はある程度頑健に動くことが確かめられた．

本実験では日本語と英語の両方を用いた．言語ごとに訓練データを用意することでシステムを多言語対応できることが示された．なお，訓練データ中の表現には言語毎に作業者の国や文化の違いが見られた．例えば，図 1 の白板消しは欧米で販売されているもので英語では白板消しと呼ばれる一方，日本では一般的でないため日本語のデータでは形状と色を使って表現されていた．このように，物体の参照表現は一つの言語のデータだけを集めて翻訳しても得られないものがあることがわかった．

5.　結論

本論文では，人の口語指示を理解し，曖昧時には聞き返すインタラクション能力を持ったシステムの提案を行った．本モデルは画像処理と言語処理を深層学習で end-to-end で構築したものである．その結果，英語と日本語を同じシステムで実現し，英語指示では 92.7%，日本語指示では 84.1%と高い性能を示した．今後の展望として，異なる言語間での知識の再利用により，少ないデータ数での実現を目指す．

図 6 物体数約 70，指示文「左上のマスタードのボトルを右下に」の時の認識例

参　考　文　献

[1] R. Paul, J. Arkin, N. Roy, and T. M. Howard, “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] T. Yamada, S. Murata, H. Arie, and T. Ogata, “Representation learning of logic words by an rnn: From word sequences to robot actions,” *Frontiers in neurorobotics*, vol. 11, p. 70, 2017.

[4] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh, “Interactive text2pickup network for natural language based human-robot collaboration,” *arXiv preprint arXiv:1805.10799*, 2018.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, and A. Berg, “SSD: Single shot multibox detector,” in *ECCV*, 2016.

[6] Y. Licheng, T. Hao, B. Mohit, and T. L. Berg, “A joint speaker-listener-reinforcer model for referring expressions,” in *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[7] S. Tokui, K. Oono, S. Hido, and J. Clayton, “Chainer: a next-generation open source framework for deep learning,” in *Workshop on machine learning systems on Neural Information Processing Systems*, 2015.

[8] Y. Niitani, T. Ogawa, S. Saito, and M. Saito, “ChainerCV: a library for deep learning in computer vision,” in *Proceedings of ACM Multimedia Workshop*, 2017.

[9] E. Sibirtseva, D. Kontogiorgos, O. Nykvist, H. Karaoguz, I. Leite, J. Gustafson, and D. Kragic, “A comparison of visualisation methods for disambiguating verbal requests in human-robot interaction,” *arXiv preprint arXiv:1801.08760*, 2018.