ロボットへの教示手段としての言語処理の課題

海野 裕也　　坪井 祐太
株式会社 Preferred Networks
{unno, tsuboi}@preferred.jp

1　はじめに

家庭や職場で誰もが自由にロボットを使う．高齢化による労働力不足を背景に，ロボットに対する期待は高まる一方で，未だに身近な環境でロボットは活躍していない．本稿では，ロボットの技術的，産業的な面の課題意識から，言語処理をの発展によってどのように貢献できると考えられるかを明らかにしていく．

2　ロボット産業の現状

そもそも世の中にはどのようなロボットがあるだろうか．大別すると，スピーカーやマイクを備えて人と対話する，コミュニケーション型のロボットと，アーム（図 1）や台車を備えて加工や搬送などを行う作業型のロボットに分かれる．前者が音声処理やソフトウェアに重きが置かれる一方，後者はモーターの制御や各種センサーを使った低遅延な処理に重きが置かれる．ここで扱いたいのは後者である．

現在，世の中で作業型のロボットが使われているのは，製造業の生産現場である．例えば自動車産業では大量のアーム型ロボットが使われている．一方で，製造業に限っても，ロボットが十分に普及しているのは自動車産業や電子デバイス産業などの一部の業種に限られている．ましてや家庭内で活躍するには，まだ技術的に大きなギャップが存在する．

一般的に産業用ロボットは「半完成品」と言われており，単体では機能しない．まず，ロボットは「腕」に相当する関節部分だけしか持たず，エンドエフェクタと呼ばれる「手」に相当する部分はシステム構築者が要求に合わせて設計するのが一般的である．加えてロボットがどのように動くかは，システム構築者が設定する必要がある．この作業は「教示」と呼ばれている．現在ロボット産業の足かせになっているのは，こうしたシステム構築を行うロボットシステムインテグレーター（以下，SIer）の不足と，業種ごとの経験が

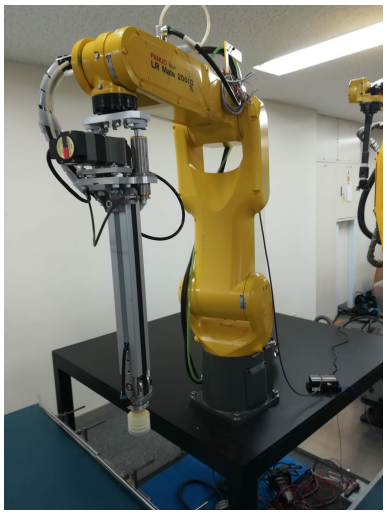

図 1: 産業用ロボットの例．ロボット本体は黄色の部分のみで，吸引式のグリッパを別途設計する必要がある．

足りないことと言われている [7]．言語処理の技術によってこの SIer の作業を肩代わりすることが期待されるのは，後者の教示作業であろう．

実際に，教示作業はどのように行われるのだろう．ロボットアームは複数の関節から構成されており，各関節にはモーターが内蔵されている．ロボットに仕事をさせたいときは，手先位置をどこに持っていきたいのかをティーチングペンダントと呼ばれる専用の端末で入力し，その時の関節角を記録する．現場でロボットを動かすときはこの記録された関節角を再現するように動作させる．これが一般的なロボットの利用形態であり，ティーチングプレイバックと呼ばれている．

現状ではこの過程は人手で行われているため，広く応用される上でのボトルネックになりやすい．例えば，ロボットアームを個人が利用しようとすると，こうした教示作業を個人が行う必要がある．教示作業を簡素化するために，近年ではロボットアームを直接動かして教示させるダイレクトティーチや，シミュレーション画面で教示を行うオフラインティーチなどの方式が

積極的に製品化されている．しかし，これらの手段を取ったとしても，やってほしい仕事を直接的にロボットに指示する世界からは大きくかけ離れている．

3　指示のための言語処理

SIer の行っている教示作業を分解してみよう．ロボットシステムで行って欲しい要求仕様から，実際のロボットの動作を考案する．そこから，実際にロボットが経由して欲しい経由点を入力する．実環境で問題ないことを確認して，実際に動かす．

例えば,「段ボール箱を一つずつ掴んで，ベルトコンベアに載せる」という要求があったときに，段ボール箱がどれか，掴むために手先をどの位置に動かすか，掴んだ後にどこまで運ぶか，そのときにどのような経路でアームを動かすか，これらは全て座標位置や手先の向きを人手で入力する必要がある．

この操作をソフトウェアで代替することを考える．すると，自然言語で書かれた要求仕様を理解する言語処理技術，環境を認識するための画像認識とセンサーの技術，認識環境と指示内容から経由点を生成するロボティクスの技術などが必要になる．現状では SIer が人力でこれらの処理を行い，判断し，教示をしていると言っても良い．これまで計算機によってこうした判断を行う事が技術的にできなかったため，最初から諦めていたという点はあるかもしれない．一方でこの数年のセンサー性能の向上や，画像処理と言語処理の革新によって，これらをソフトウェア的に実現することは現実味を帯びてきている．この上で，どのような技術的な課題が出てきそうか，いくつか具体的に上げてみよう．

- 操作対象物を特定する必要がある．対象物が 1 種類であれば，特別に指示する必要性は無いが，環境に複数の物体が存在する場合に一意に物体を指定する必要がある [5] ．
- 操作のためのロボットの姿勢を指定する必要がある．例えば持ち上げたネジを所定のネジ穴に入れることを考えてみよう．掴んだ姿勢によっては，ネジ穴に差し込めない．そのため，どこをどのように把持するか，どのような姿勢で置くかというレベルの情報が必要になる．
- 操作の指示には，必ずし物体が存在せず，場所を指定する必要があることがある [3]．例えば，テレビの横にリモコンを置いて，と言われたときの「横」とはどこのことを指すだろう．横に置いて，と，横に掛けて，では指し示す場所のイメージが大きく異るだろう．
- 「置く」「挿す」「回す」など，動作を示す表現がロボットの動きを指示する必要がある [6]．対象物のサイズや形状，姿勢によっても実際の動作は異なってくる．例えば，コネクタを挿す動作を考えると，挿すものと挿す対象の姿勢や形状によって行うべき動作は変わってくるだろう．
- 「ゆっくり」「少しだけ」「ぴったりと」「揃えて」など，動作や状態に対する修飾表現を扱える必要がある．これらの表現は多くの場合曖昧であったり，言語表現だけでなく状況によって意図する内容が変わってくることが想定される．
- 言語による指示のみならず，曖昧なフィードバックを元に意図を汲んだり [1]，指差しと言語指示を組み合わせたり [2]，デモンストレーションと言語による指示を組み合わせるなど [4]，他の種類の情報をうまく活用するなどの方向性もある．

このように，動作や仕事を指示する表現は，置かれた状況によって意味合いが変わったり，曖昧性が生じることがある．これらは純粋に言語で記述された文の解析のみではなく，環境やロボットの状態の解析も含めた，複合的な解析が必要となる．

4　おわりに

本稿ではロボット産業の現状から，ロボットへの教示作業が大きなボトルネックの一つとなっていることを紹介した．ロボットへの指示のためのインターフェースとして，言葉による指示は可能性があり，言語処理が果たす役割は大きい．加えて，ここにはこれまでの課題設定とは異なる様々な問題が存在する．今後，ロボットの文脈での言語処理の研究が発展することが望まれる．

参考文献

[1] Riku Arakawa, Sosuke Kobayashi, Yuya Unno, Yuta Tsuboi, and Shinichi Maeda. DQN-TAMER: Human-in-the-loop reinforcement

learning with intractable feedback. In *Proceedings of Workshop on Human-Robot Teaming Beyond Human Operational Speeds and Robot Teammates Operating in Dynamic, Unstructured Environments (RT-DUNE)*, 2019.

[2] Richard A. Bolt. "Put-That-There": Voice and gesture at the graphics interface. *SIGGRAPH Computer Graphics*, Vol. 14, No. 3, p. 262–270, 1980.

[3] Ozan Arkan Can, Pedro Zuidberg Dos Martires, Andreas Persson, Julian Gaal, Amy Loutfi, Luc De Raedt, Deniz Yuret, and Alessandro Saffiotti. Learning from implicit information in natural language instructions for robotic manipulations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)*, 2019.

[4] Tsu-Jui Fu, Yuta Tsuboi, Sosuke Kobayashi, and Yuta Kikuchi. Learning from observation-only demonstration for task-oriented language grounding via self-examination. In *Proceedings of Workshop on Visually Grounded Interaction and Language (ViGIL)*, 2019.

[5] Jun Hatori, Yuta Kikuchi, Sosuke Kobayashi, Kuniyuki Takahashi, Yuta Tsuboi, Yuya Unno, Wilson Ko, and Jethro Tan. Interactively picking real-world objects with unconstrained spoken language instructions. In *Proceedings of International Conference on Robotics and Automation (ICRA)*, 2018.

[6] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. ALFRED: A Benchmark for Interpreting Grounded Instructions for Everyday Tasks. Dec 2019.

[7] 日本経済再生本部. ロボット新戦略, 2015.