生活支援ロボットにおけるマルチモーダル言語処理

杉浦 孔明
国立研究開発法人　情報通信研究機構
komei.sugiura@nict.go.jp

1　はじめに

少子高齢化社会のなかで，1 人の要支援者を物理的･経済的に支える生産年齢人口は減少しており，要支援者の家族が離職を余儀なくされるケースが発生するなど，社会全体の生産性向上を妨げている．この解決策として，介助犬（育成に 2 年必要）レベルのタスクを行う生活支援ロボットの研究開発が活発に進められている．生活支援ロボットのハードウェアは整備されつつあるものの，現状では，曖昧な音声命令を理解し，タスクを実行する精度が不十分である．例として，日常環境で「シリアルと牛乳を取ってきて」という音声命令をロボットが実行するタスク（Fetch and Carry タスク）を考える．人間同士の場合はこのような省略された命令文で通じる場合が多いが，ロボットが行動を開始するために十分な情報を含んでいない．一方，十分な情報を含む指示文[1]をユーザに強いることは著しく不便である．

上記の問題への単純なアプローチとして，スロット値がすべて確定するまで聞き返す戦略が考えられる．実際に，世界最大の生活支援ロボットのベンチマークテストであるロボカップ@ホーム [5] においても，このアプローチが支配的である．しかしながら，このアプローチでは「どのペットボトルですか？」「キッチンのどの棚ですか？」「棚の何番目の段ですか？」など多くの確認発話が生成されるため，動作実行するまでに必要な時間が長く，不便である．

本稿では，上記の問題に対して，実世界知識を利用して曖昧性解消を行うロボットの関連研究と我々の取り組みについて紹介する．

[1]例えば,「キッチンの一番高い棚の 3 段目の右側にあるシリアルと冷蔵庫のドアポケットにある牛乳を取ってきて」

2　ロボティクスにおけるマルチモーダル言語処理

実世界知識にグラウンドしたコミュニケーションを行うロボットに関しては，ロボティクス・音声対話・自然言語処理・画像処理等の分野において，以前から多くの研究が行われてきた．Kollar らは，ロボットに与える移動指示に関して，ランドマークオブジェクトや動作にグラウンドした言語表現を学習する手法を提案した [6]．本分野に関連するプロジェクトとしては，DARPA BOLT [10]，Robo Earth [14]，RoboBrain [12]，長井らによる CREST プロジェクト [18]，などがある．本分野と関連が深いベンチマークテストとしては，ロボカップ@ホーム [5] がある．ロボカップ@ホームは世界最大の生活支援ロボットのコンペティションであり，日用品の探索，棚からユーザに言われたものを取ってくる，などの移動マニピュレーションとヒューマンロボットインタラクションを統合したタスクが設定されている．

近年の代表的なアプローチとしては，visual semantic embedding (VSE) [4, 9, 11, 13, 16]，visual question answering(VQA) [2], キャプション生成 [15] などがある．これらのアプローチは，視覚的特徴と言語的特徴を共通の潜在空間に埋め込むものが多い．[16] では，参照表現の生成と理解を同時に行うモデルが提案されている．[4] は，Pick and Place タスクに関する言語理解および音声対話を扱うとともに，PFN-PIC データセットを公開している．移動タスクを扱った研究としては，[1] が挙げられる．また，[7] では，Attention Branch Network [3] を利用し，Fetch and Carry 指示文に関するマルチモーダル言語生成を行う手法が提案されている．

一方，我々はマルチモーダル言語理解手法 Multimodal Classifier Generative Adversarial Network (MMC-GAN) を提案し，Carry and Place タスクに適用した．また，[8, 9] では，Multimodal Target-Source

Classification Model with Attention Branches (MTCM-AB) を提案した．本手法を Pick and Place タスクに適用し，PFN-PIC データセットに対して言語理解精度 90.1%を得た．これは，人間による言語理解精度 90.3%と遜色ない精度といえる．

3 おわりに

生活支援ロボットによる曖昧な音声命令の理解は，多くの関連課題を有する挑戦的な分野である [17]．本稿では，本分野に関連するタスク概要と代表的な手法を紹介した．

謝辞

本研究の一部は，JST CREST，NEDO，総務省 SCOPE の助成を受けて実施されたものである．

参考文献

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proc. CVPR*, pp. 3674–3683, 2018.

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, Lawrence Z.C., and D. Parikh. VQA: Visual question answering. In *ICCV*, pp. 2425–2433, 2015.

[3] H. Fukui, T. Hirakawa, T. Yamashita, and H. Fujiyoshi. Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. In *CVPR*, pp. 10705–10714, 2019.

[4] J. Hatori, et al. Interactively Picking Real-World Objects with Unconstrained Spoken Language Instructions. In *IEEE ICRA*, pp. 3774–3781, 2018.

[5] Luca Iocchi, Dirk Holz, Javier Ruiz-del Solar, Komei Sugiura, and Tijn van der Zant. RoboCup@Home: Analysis and Results of Evolving Competitions for Domestic and Service Robots. *Artificial Intelligence*, Vol. 229, pp. 258–281, 2015.

[6] T. Kollar, S. Tellex, D. Roy, and N. Roy. Toward Understanding Natural Language Directions. In *Proc. ACM/IEEE International Conference on Human-Robot Interaction*, pp. 259–266, 2010.

[7] A. Magassouba, K. Sugiura, and H. Kawai. Multimodal Attention Branch Network for Perspective-Free Sentence Generation. *Conference on Robot Learning (CoRL)*, 2019.

[8] A. Magassouba, K. Sugiura, and H. Kawai. A Multimodal Target-Source Classifier with Attention Branches to Understand Ambiguous Instructions for Fetching Daily Objects. *IEEE Robotics and Automation Letters*, Vol. 5, No. 2, pp. 532–539, 2020.

[9] A. Magassouba, K. Sugiura, A. Trinh Quoc, and H. Kawai. Understanding Natural Language Instructions for Fetching Daily Objects Using GAN-Based Multimodal Target-Source Classification. *IEEE RA-L*, Vol. 4, No. 4, pp. 3884–3891, 2019.

[10] Shiwali Mohan, Aaron Mininger, James Kirk, and John E Laird. Learning Grounded Language through Situated Interactive Instruction. In *AAAI Fall Symposium: Robots Learning Interactively from Human Teachers*, pp. 30–37, 2012.

[11] V K. Nagaraja, V I. Morariu, and L S. Davis. Modeling Context between Objects for Referring Expression Understanding. In *ECCV*, pp. 792–807, 2016.

[12] Ashutosh Saxena, Ashesh Jain, Ozan Sener, Aditya Jami, Dipendra K Misra, and Hema S Koppula. RoboBrain: Large-Scale Knowledge Engine for Robots, 2014.

[13] M. Shridhar and D. Hsu. Interactive Visual Grounding of Referring Expressions for Human-Robot Interaction. In *RSS*, 2018.

[14] Moritz Tenorth, Alexander Clifford Perzylo, Reinhard Lafrenz, and Michael Beetz. The RoboEarth Language: Representing and Exchanging Knowledge about Actions, Objects, and Environments. In *Proc. ICRA*, pp. 1284–1289, 2012.

[15] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and Tell: A Neural Image Caption Generator. In *Proc. CVPR*, pp. 3156–3164, 2015.

[16] L. Yu, H. Tan, M. Bansal, and T L. Berg. A joint Speaker Listener-Reinforcer Model for Referring Expressions. In *CVPR*, Vol. 2, 2017.

[17] 杉浦孔明. ロボットによる大規模言語学習に向けて-実世界知識の利活用とクラウドロボティクス基盤の構築-. 計測と制御, Vol. 55, No. 10, pp. 884–889, 2016.

[18] 長井隆行, 谷口忠大, 尾形哲也, 岩橋直人, 杉浦孔明, 稲邑哲也, 岡田浩之. 記号創発ロボティクスによる人間機械コラボレーション基盤創成. 第 19 回クラウドネットワークロボット研究会, pp. 23–27, 2015.