Graduate School of Fundamental Science and Engineering
Waseda University

博士論文審査報告書

論　文　題　目

Learning of Language Grounding in Robot Behavior by Recurrent Neural Networks

リカレントニューラルネットワークによる言語と行動のグラウンディングの学習

申　請　者

Tatsuro	YAMADA
山田	竜郎

Department of Intermedia Studies,
Research on Intelligence Dynamics and Representation Systems

2019 年 2 月

ロボットが人間の要求にしたがって，あるいは人間とともに協調して働くためには，言語的にコミュニケーションをとる能力が求められる．ここでは雑談や質疑応答だけではなく，人間の言語指示を理解しタスクをおこなう，また逆に現在の状況や出来事を言語的に説明するといった能力が求められる．すなわち，実世界に接地（記号接地）された言語を使用することが最も重要となる．博士論文では，このような実世界にグラウンドされた形式で言語を理解し使用するロボットを実現する計算モデルを提案している．

本論文では特に，(1) ロボットが状況に依存して変化する言語表現の意味を理解し行動できる，(2) ロボットは内容語と機能語の両方からなる文章を理解することができる，(3) 言語表現とロボット自身の行動が双方向に変換できる，という３点に着目した研究を展開している．近年，深層学習の技術を用いて，言語表現からシミュレータ環境上エージェントの動作への変換を学習させる研究が盛んに行われている．しかし，実ロボットを用いた研究例，さらには動作から言語説明への変換も同時に実現している研究例はほとんど報告されていない．

これらの課題を解決するために，本研究では，Recurrent Neural Network (以下，RNN)を用いた機械学習モデルを使用している．本研究の中心的なアプローチは以下の三点である．

1. RNN 型のエンコーダデコーダモデル（Encoder- decoder model; 以下，EDM）によるロボット行動と言語表現の変換のボトムアップ学習．
2. 状況依存性を解決するための視覚情報の統合．
3. 双方向性を実現するための内部表現の共有．

RNN 型 EDM は元々自然言語処理の分野で対話や翻訳のために提案された学習器である．本研究では，このフレームワークを言語表現とロボット行動のマルチモーダル学習に応用している．言語と行動の関係性は全てデータからボトムアップに学習され人間の恣意的な設計を必要としない点に特徴がある．マルチモーダル情報（視覚，姿勢など）と統合されることで，現在の状況における意味にしたがって埋め込まれ，さらには，論理関係を示す「機能語」をも RNN の潜在表現の操作機能として埋め込まれることも確認している．また言語と行動の双方向変換を達成するため，行動用と言語用の 2 つの EDM を用意し，それらの内部表現を互いに共有するように訓練するモデルを提案し，その有効性を確認している．またその神経回路モデルの内部解析から，文法，運動，などの表現獲得の枠組みについても詳しく解析している．

本論文は，7 つの章で構成されている．以下に各章の概要について述べる．

1 章では本研究に至った背景とその目的，および課題を解決するためのアプローチについて概観し，そのアプローチの妥当性と論文全体の構成について述べている．

2 章では，本論文に関連する先行研究を紹介している．具体的には，ロボットや仮想エージェントにおいて，言語と運動を統合するための従来の記号

推論のアプローチ，また深層学習以降の学習型アプローチなどを紹介するとともに，本研究の立ち位置を明確化している．

3 章では，最初の二課題，すなわち（1）状況依存性の解決および（2）内容語・機能語双方の処理を達成するための手法であるマルチモーダル情報受容型の RNN 型 EDM を提案する．これは前述の第一，第二アプローチを統合したものである．まず一般的な Neural Network (以下，NN)，RNN，および EDM について順を追って説明したのち，提案手法についてその構造と挙動を詳しく説明する．提案モデルは，まずエンコーダ RNN が言語指示（ロボット行動）を固定次元のベクトルとして潜在空間に埋め込んだのち，それをデコーダ RNN が展開することで対応するロボット行動（言語説明文）を生成する．視覚とロボットの姿勢情報も共にエンコードすることで，言語表現の状況依存性を解決する．

4 章では，内容語のみを含む文章からロボットの行動への一方向変換タスクの学習のフレームワークと結果の解析を行なっている．先に述べた第一の課題に照らしてモデルを評価するため，生成すべき行動はその時の視覚入力（状況）に依存して異なるように設計している．ロボットが卓上の複数のベルを叩く，もしくは指さす，というタスクにおいて，提案モデルは言語，視覚，行動の関係を体系的に学習し，可能な状況の 1/3 の学習から未学習の状況でも適切な行動を生成する汎化能力を達成している．またその内部表現の解析も行なっている．

5 章では，内容語と機能語の両方を含む文章からロボットの行動への一方向変換について述べている．具体的には，先に述べた第二の課題である内容語と機能語双方からなる文章の理解に照らして評価している．機能語の一例として特に，“not”，“and”，“or”といった論理語が含まれるタスクを設定した結果，モデルは，これらの語を含む旗揚げタスクを学習することが可能となっている．学習後のモデルの内部表現について主成分分析を用いて解析した結果，内容語が視覚情報と統合された形で表現されること，また論理語はその語の示す論理操作に従った形で表現されることを確認している．例えば，指示動作の否定を表す“not”は，内部表現の非線形な変換として作用すること，そして異なる指示が寄与率の低い空間で，同じ動作となるように埋め込まれることを確認している．

6 章では，第三の課題である双方向変換能力を付与するため，表現共有法による拡張手法を提案し，実際に言語と行動の双方向変換の学習を行なった結果について述べている．学習後のモデルは，視覚情報として与えられる状況に即して，言語と行動を双方向に変換することに成功した．内部表現を主成分分析により解析したところ，文章と，それに意味的に対応する行動シーケンスが，共有空間において互いに近傍に埋め込まれていることが確認された．さらに大規模な運動データを利用した，スケーラビリティ拡張の可能性についても議論している．

7 章では本研究が，実世界にグラウンドされたロボットの言語使用をどのような側面において達成したか，またその貢献について総括し，残された課題と今後の展望について述べている．

以上をまとめると，本研究は記号接地問題という大問題を背景として，文脈依存する言語をロボットの運動への変換，さらにはその逆変換を，深層学習モデルである EDM をマルチモーダル学習にまで拡張することで実現した基礎研究である．この成果は，記号接地問題の一部の理解に対する学術的貢献のみならず，描画像からの動作連想や，行為経験を用いた画像認識などの人間の認知的側面の理解，さらにはロボットと人間との言語インタラクションへの応用の可能性を示すものである．これらの研究成果は複数の国際ジャーナルに採択されるとともに，国際会議などにおいて論文賞を受賞するなど国内外で高く評価をされており，高い学術的価値を持つとともに，表現工学の発展に寄与するものと評価できる．以上から本論文を博士（工学）の学位論文として相応しいものとして認める．

2019 年 2 月

審査員

主査　早稲田大学教授　博士（工学）　早稲田大学　尾形　哲也

早稲田大学教授　博士（工学）　早稲田大学　及川　靖広

早稲田大学教授　博士（人間科学）早稲田大学　河合　隆史