再帰型神経回路モデルの内部表現の共有による ロボットの行動と言語の双方向の生成

Bi-directional Generation of Robot's Behavior and Language by Shared Representation of Recurrent Neural Network

5116E021-6　松永　寛之　　指導教員　尾形　哲也　教授

MATSUNAGA Hiroyuki　　Prof. OGATA Tetsuya

概要：　本研究の目的はロボットの行動から言語，言語から行動の双方向の生成である．双方向生成のために，言語と行動を生成するための Sequence Autoencoder (SA) と，言語と行動を対応付けるための内部表現の共有の２つの手法を組み合わせた学習モデルを提案する．提案モデルが状況に応じてロボットが言語と行動を対応付け，双方向の生成が可能か検証を行い，双方向の生成が可能であることを確認した．

キーワード：内部表現の共有，再帰型神経回路モデル，Sequence Autoencoder，双方向生成，ロボット

Key Words：shared representation，recurrent neural Network，sequence autoencoder，bi-directional generation，robot

1. はじめに

人間が生活している実環境下において，人間とロボットが協調して働くためには，言語表現を用いたコミュニケーション能力が必要である．しかし，ロボットにとって，実世界を言語表現と対応付けて理解することは困難である（記号接地問題 [1]）．この課題の解決を目的として，ロボットによる言語と行動の接地を扱った研究が行われてきた．以前は作りこまれたルールを基に言語理解を行うアプローチ [3]が行われていたが，実世界のあらゆる環境をルールの中に落とし込むことは困難であった．近年ではロボットにデータを与えルールを学習するアプローチが行われている．このアプローチを選択した際の学習器の１つが再帰型神経回路モデル (Recurrent Neural Network; RNN) である．RNN はシーケンスデータを学習することが可能な神経回路モデルである．Ogata ら [2]は，言語用と行動用の RNN を用い，言語と行動を対応付ける PB と呼ばれるパラメータを設定して双方向の生成を実現した．しかし，この手法は生成時に PB を求める反復計算を必要とすることが課題であった．

本研究では，状況に応じてロボットの行動から言語，言語から行動の双方向生成を行うことを目的とする．生成時に反復計算を必要としない，RNN を用いた双方向生成モデルを提案し，双方向の生成が可能か実際にロボットを用いて検証を行う．

2. 提案モデル

本研究では，以下の２つの手法を組み合わせて，言語から行動，行動から言語の双方向の生成を行うことを提案する．

①Sequence Autoencoder (SA)

②内部表現の共有

①は言語と行動の各シーケンスの生成のために必要となる．SA は RNN を用いて，シーケンスデータを符号化（エンコーディング）して圧縮された内部表現ベクトルにし，その内部表現から復号化（デコーディング）して元のシーケンスデータを生成することが可能な学習モデルである [4]．②は言語と行動を対応付けるために必要となる．言語と行動の各シーケンスの内部表現に学習時に拘束をかけることで，対応する言語と行動の表現を共有する手法である [5]．図１に提案モデルの概要を示す．

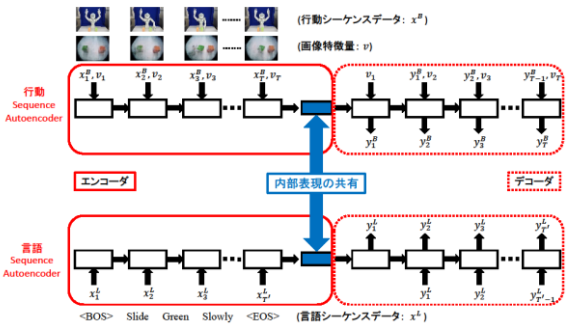

図１　提案モデルの概要

提案モデルでは，言語の SA（図１下側）と行動の SA（図１上側）の２つの SA を用いる．各 SA はエンコーダ（赤い実線で囲まれた部分）とデコーダ（赤い点線で囲まれた部分）の２つの RNN で構成されている．

言語の SA では，エンコーダに言語シーケンスデータ$\boldsymbol{x}^L = (\boldsymbol{x}_1^L, \boldsymbol{x}_2^L, \ldots, \boldsymbol{x}_{T'}^L)$を入力し，内部表現空間に埋め込む．最後の入力$\boldsymbol{x}_{T'}^L$を埋め込んだ内部表現が，入力データを表現するベクトルである．行動の内部表現と共有するためには，各内部表現の次元が等しい必要がある．各エンコーダとデコーダの間に共有層（図１の青い層）を設け，次元を等しくするためにアフィン変換を行う．共有層から元の次元に戻すアフィン変換を行ったベクトルをデコーダの最初の入力として，最初の出力$\boldsymbol{y}_1^L$を出力する．次の時刻では$\boldsymbol{y}_1^L$を入力として，$\boldsymbol{y}_2^L$を出力する．最後の時刻までこれを繰り返す．出力シーケンス$\boldsymbol{y}^L$が目標とするのは，入力した$\boldsymbol{x}^L$自身であり，$\boldsymbol{x}^L$と$\boldsymbol{y}^L$の誤差を最小化する学習を行う．

行動の SA では言語の SA と同様に行動シーケンスを扱うが，加えて画像特徴量 $\mathbf{v}$ をエンコーダ，デコーダの入力として用いる．行動の出力は画像に条件付けられる意味的な表現であること

を学習可能なモデルになっている．

言語と行動のシーケンスを対応づけるために，言語 SA 中の共有層と行動 SA 中の共有層に拘束をかけて，内部表現の共有を行う．拘束は，言語と行動が対応している場合は，表現が近づくように，逆に対応していない場合は，表現が遠ざかるように損失関数を定義して行う．

提案モデルでは，言語を入力しエンコーディングした表現を行動のデコーダに入力することで，行動を生成することが可能である．また，その逆も然りである．

3. 実験

3.1 タスクデザイン

状況に応じてロボットが言語と行動を対応付け，双方向の生成可能か検証するタスク設定を説明する．状況を変化させるために，3 色（赤，緑，黄）の同じ形状の物体の内 2 つをロボットの前方に左右に並べる．並べ方は 6 通りである．言語として，動詞，目的語，副詞の 3 単語で構成される文章を用いる．表 1 に文章と運動の種類を示す．言語は 18 通り，行動は 12 通りの運動である．

表 1　文章と運動の種類

文章	動詞	目的語	副詞
	push pull slide	red green yellow	slowly fast
運動	動作	対象物の位置	動作の速さ
	PULL PUSH SLIDE	RIGHT LEFT	SLOW FAST

言語から行動の生成では，ロボットは文章指示された色の物体が左右どちらにあるか画像を用いて判断し，適切な運動の選択を行う．逆に行動から言語の生成では，動作対象物が何色かを画像を用いて判断し，適切な文章を生成する．

3.2 学習データ

行動のシーケンスデータとして，ロボットの左右 5 次元ずつ腕の関節角度の値を用いる．言語は，全単語数 8 にシーケンスの最初と最後の合図の記号を加えた 9 次元の one-hot-vector 表現を用いる．画像のデータは，ロボットで取得した生画像から事前に抽出した 10 次元の特徴量を用いる．学習データは全 72 パターン（運動 12 通りのそれぞれに環境 6 通り）を 1 セットとし，6 セットを学習用に収集した．学習は 2 通り行った．学習 1 では全 72 パターンを学習させた．学習 2 では各文章につき 1 つの運動をランダムに抜いて，計 54 パターンのみを学習させた．

4. 結果

学習したモデルを実装したロボットを用いて双方向生成が可能か検証した．行動から言語の生成では，学習 1 と学習 2 ともに全パターンで適切な言語を生成することができた．言語から行動の生成は，物体を正しい方向に移動させることができたか否かで判定した．結果は学習 1 と学習 2 ともに，言語が”push fast”と”slide fast”の 24 パターンを除いた 48 パターンで正しい方向に物体を移動させることができた．ここで，物体を動かせなかったパターンは，腕がわずかに内側にずれることで物体を押しつぶしてしまったことが理由であり，対象物体や動作の速さは正しかった．そのため，ロボットが生成した行動がどの運動の軌道と類似度しているか判定したところ，全パターンで正しい運動を生成していたことが分かった．

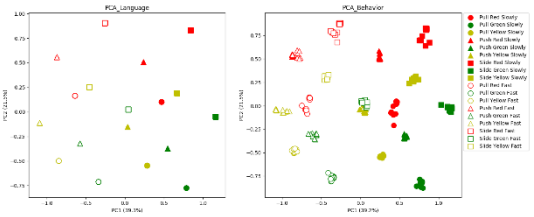

図 2　内部表現の PCA

共有した内部表現の分析を行った．学習 2 の内部表現を主成分分析 (PCA)によって 2 次元に圧縮し図 2 に示した．図 2 の左が言語，右が行動である．図 2 の点は，色は物体の色に対応し，形は動詞に対応，白抜きが“fast”，塗りつぶしが“slow”に分けて示している．左右見比べるとどちらも文章ごとに点が密集している．これは，行動のシーケンスが，運動毎ではなく言語の体系に対応付けられた表現として獲得されたことを意味する．

5. まとめと今後の展望

本研究では，双方向生成が可能なモデルとして，2 つの SA の内部表現を共有する手法を提案し，有効性を検証した．本実験では言語と行動が，画像を用いる条件下で 1 対 1 に対応するタスクに限定したが，今後は言語と行動が多義的に対応するタスクへの応用が考えられる．また提案モデルは，SA を増やせばより多くのモダリティを扱うことが可能であり，他のマルチモーダル統合研究への応用が考えられる．

参考文献

[1]S.Harnad, “The symbol grounding problem”, Physica D, Vol.42, pp.335-346, 1990

[2]T.Ogata, M.Murase, J.tani, K.Komatani, H.G.Okuno, “Two-way Translation of Compound Sentences and Arm Motions by Recurrent Neural Networks”, IEEE Int. Conf. on Intelligent Robots and Systems, pp.1858-1863, 2007

[3]Winograd, Terry. “Understanding natural language”, *Cognitive psychology*, 1972, 3.1: 1-191.

[4]Dai, Andrew M., and Quoc V. Le. "Semi-supervised sequence learning." *Advances in Neural Information Processing Systems* (2015): 3079-3087.

[5]Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "See, Hear, and Read: Deep Aligned Representations." *arXiv preprint arXiv:1706.00932* (2017).