

# 神経力学モデルによる文字列からの言語構造の自己組織化とロボット運動感覚との統合

○尾形哲也<sup>†‡</sup> 日下航<sup>†</sup> 奥乃博<sup>†</sup> (<sup>†</sup>京都大学大学院, <sup>‡</sup>科学技術振興機構)

## 1. はじめに

言語はコミュニケーションにおいて最も重要な役割を果たす“ツール”である。特にロボットと人間との言語インタラクションは、従来の音声対話研究が直接扱ってこなかった、実世界の状況に依存した言語の多義性や意味の補完などの問題が顕在化する。本来、実世界と記号世界は多様な対応を許しながら動的に変化する複雑な系であり、“記号接地問題[1]”との関連から、人工知能における最重要課題の一つとなっている。

これまでに特に工学的な立場から、確率モデル等を利用した文章と動作系列の変換等の試みがいくつか報告されている。本研究では、特に力学系を基盤とした、言語、感覚運動系の変換モデルに取り組む。具体的には人工神経回路モデルによる力学系構造を基盤とし、ロボットに自らの感覚-運動系と接地した言語を獲得させることを目的とする。このプロセスにおいて、脳という神経ネットワーク（力学系）による言語認知機構の理解を追求するとともに、ロボット自身に実世界と言語およびそれらの対応関係を、作り込みでなく学習によって獲得できる枠組を実現することを目的とする。

本稿では、我々が構築した再帰型神経回路モデルを利用した、一つのモデルを紹介する。具体的には、以下の3条件を実現するモデルを示す。(1)単語・文法の知識を与えない（文字列からの自己獲得）、(2)感覚運動の特徴量を与えない（自己組織化）、(3)言語・運動をともに力学系で記述する。

## 2. 従来研究と問題点

従来にも神経力学モデルを用いてロボットの感覚-運動系と言語を統合的に認知する試みがいくつか提案されている[2, 3]。これらの研究ではパラメータノードを共有する2つのRecurrent Neural Network (RNN) に、それぞれ感覚-運動系フローと単語列を学習させることで相互連想を実現する。学習対象文が2~3単語で構成される非常に単純な場合であればこの枠組は十分に機能するが、ある程度以上の複雑な文章では、この枠組は十分に働かない。具体的には、単語列が長くなると後半の単語は、RNNパラメータ空間でフラクタル階層にコーディングされてしまい、単語の意味を抽出し感覚-運動系フローに反映させることが困難となる事を我々は明らかにしている。

この原因は、従来モデルが語順などによる文の“構

造”と具体的な意味を担う“内容語”を同一のニューロン発火状態（パラメータ空間）に埋め込もうとしていることにある。一方、人間が言語を認知する場合、文の“構造”と“内容語”は脳の別ルートで処理されると考えられている[4]。例えば、“Marie broke window.”という文は、抽象化された要素（格、case）からなる“構造”([AGENT][ACTION][OBJECT])は主に上側頭回において処理され、個々の格のスロットに嵌る“内容語”(例. AGENT = “Marie”)は中側頭回で処理されると言われている。そこで、この構造を我々のモデルに反映することを考えた。

また従来のRNNによる言語獲得モデルのもう一つの問題としては、入力最小単位が単語であるという点が挙げられる。実際の言語では、文は単語から構成され、さらに単語はより小さな単位（音素・文字）から構成される。そこで本研究では、入力単位を従来から一段引き下げ文字列から、二重分節性の獲得を目指すこととした。

## 3. 提案する言語-運動統合認知モデル

### 3.1 全体モデル

本研究で提案する言語-運動統合認知モデルは、具体的にはモータ値の時系列からなるロボット動作パターンと文字系列からなる文の相互連想を行う。我々のモデルの概略を図1に示す。

動作パターン、言語ともにトップダウンのモデル化は行わず、データの汎化によって自己組織的に構造を獲得させることを目指す。具体的には、動作入力には各関節角度、視野入力は生画像を入力する。行為パターンを自己組織化する感覚-運動系RNN (Sensory-Motor RNN)と、言語を自己組織化する言語RNN群 (Language RNNs)が少数のニューロンを介して相互作用することで、相互連想を実現する。

### 3.2 MTRNN

図1において各RNNのブロックにはMultiple Timescale RNN (MTRNN) [5]を利用している。異なる時定数の連続値ニューロン群からなるCTRNN (Continuous Time RNN)が、スパースに結合し全体のネットワークを形成する。用意するニューロン群、またその結合方法は任意に設計可能である。

本研究で用いたニューロン群は、入出力ノード(IO)と時定数の異なる2種類の文脈ノード (Context fast: Cf, Context slow: Cs)、および時定数が無限大のパラメータノードの4種類である。IO, Cf, Csの時定数( $\tau$ )はそれぞれ2; 5; 70と設定した。時定数が大きいほどニューロン状態の変化が緩やかになる。また、パラメ

ータノードは前向き計算の間は値を一定に保つ。MTRNN は、学習・認識・生成の3機能を実現する。

- (1) 学習：IO ノードに教師時系列データを入力し、Back Propagation Through Time (BPTT) によって、結合重みとパラメータ空間を更新する。
- (2) 認識：学習済みモデルの IO ノードに認識したい時系列データを入力し、BPTT によってパラメータノードの値のみを更新する。これにより、対象データを表現するパラメータを得られる。
- (3) 生成：パラメータノードに値をセットし、RNN の前向き計算を行うことで、IO ノードの発火状態の時系列データを得る。これが、与えたパラメータが表現する時系列パターンになっている。

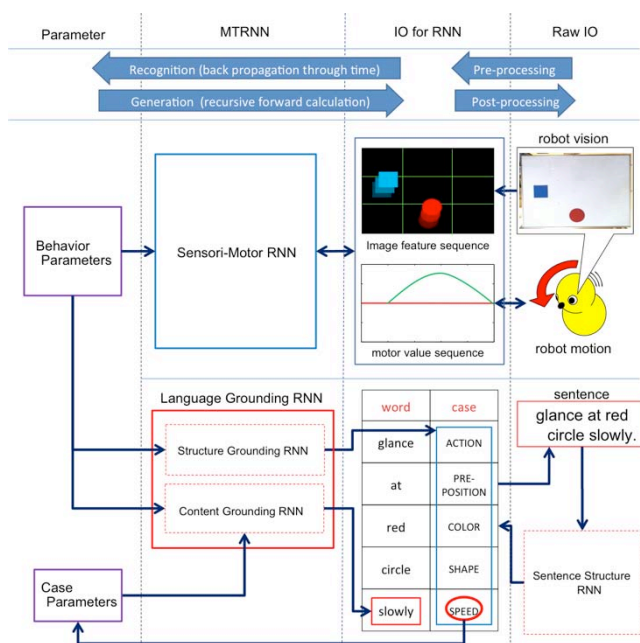


図1 提案する言語-感覚運動統合神経力学モデル

### 3.3 言語神経モデルと身体神経モデル

言語神経ネットワークは、前述した脳認知科学の見解に沿って、単語品詞種を出力する格 RNN (Structure Grounding RNN) と、単語綴りを出力する内容語 RNN (Content Grounding RNN) の2つを用意した。

我々の従来研究[6]を基に、まず IO, Cf, Cs の3つの RNN からなる MTRNN に、単独で言語(文字列)のみを学習させる。

[6]の知見からこの内容語 RNN の Cf ノードには、文字列から分節化された単語及びその格が発火状態として表現されるように自己組織化される。また Cs には文の文法的な複雑さ(目的語や副詞の有無)に応じた表現が獲得される(文字列から単語、文章への二重文節化構造)。そこで Cf ニューロンの発火状態を入力として、SOM (Self-Organizing Map) に学習を行わせる。すると SOM の認知ニューロンに独立した“格(品詞)”が自己組織的に割り当てられる。

格 RNN は、行為パラメータ (behavior parameter) を

バイアスとして、この SOM の勝者ニューロンの番号列を学習させたものとする。また内容語 RNN はこの格 RNN の出力と行為パラメータをバイアスとして文字列を学習させた物である。

身体 RNN (body model) は、複数の身体探索(バブリング)のための感覚(視覚)と動作(関節角)パターンデータを与え、身体モデルの結合重みおよび行為パラメータ (behavior parameter) を更新する。連想時にはこの行為パラメータ空間を介して、言語-感覚運動の変換が行われる。

### 3.4 言語モデルと身体モデルの変換

以下に、身体モデル、言語モデルの連想についてまとめる。

(a) 言語モデルから身体モデルへの変換：

外部より文字列が入力されると、その文字列予測誤差から内容語 RNN の Cf 及び Cs の表現が得られる。この Cf 表現が学習済みの SOM に送られ、そのニューロン番号列の予測誤差を利用して格 RNN の内部表現が得られる。さらに現時点での初期画像から得られる誤差をも利用して、行為パラメータを決定する。その後この行為パラメータを利用し、モータ指令列を動作の RNN の内部連想により取得、Keepon の動作を生成する。

(b) 身体モデルから言語モデルの連想：

指定した動作指令列の予測誤差、及び初期画像からの誤差を利用し、行為パラメータを決定する。この行為パラメータから、まず格 RNN が SOM の出力ニューロン番号列(品詞列)を出力する。内容語 RNN は行為パラメータと格 RNN の出力をバイアスとして文字列(単語)を生成し文章を生成する。

## 4. 言語-運動相互連想実験

小型ロボット Keepon[7]を用いて、首振り動作とそれを表現する文の相互連想学習を行った。Keepon は高さ 120[mm]の小型ロボットであり、首部に Pan 軸、Tilt 軸さらに垂直軸の稼働部を持つ。また頭部に2台の CCD カメラおよび1本のマイクロホンを搭載している。このロボット前面に複数の形状、色彩の異なるマーカーを配置したボードを設置し、頭部動作とそれを表現する動作を学習させた。実験風景を図2に示す。



図2 実験風景

文章は, [ACTION][DIRECTION] or [OBJ] ([SPEED]) の構造を持ち, 表 1 に示す 16 種類の単語からなる. 実際の学習時にはこれらの単語においてスペース等の入力は無く, 連続した文字列として入力される.

表 1 学習で用いた単語群

verb	look, glance
direction	up, down, left, right
sub-direction	upper, lower
color	blue, green, red
shape	square, circle
adverb	quickly, slowly
preposition	At

ここで”look”は指示方向を注視したまま静止する動作, ”glance”は指示方向を向いた後, 元姿勢に復帰する動作を意味する.

これらの単語を用いて **Keepon** 頭部動作を方向, マーカーの種類等で表現する全 102 文を学習用に準備した. **ACTION** は首振り動作の種類, **DIRECTION** は首振りの方向, **OBJ** はマーカーの色と形状, **SPEED** は速度をそれぞれ表現する. **SPEED** を省略した場合は, “slowly”と同じ速度となる. 動作パターンは 48 通り存在する. またマーカーの初期配置は 1080 通りを準備した. よって感覚運動パターンは全てで 48x1080 パターン存在する. その中の 400 の動作パターンを学習に用い, 100 パターンを評価用に用いた. この際, 各動作に対応する文章群 (102 文から選択. 一動作に複数の文章説明が可能) も用いた.

格 RNN が学習する Cf 発火状態を出力層に 9 つのニューロンを持つ SOM に学習させた. その結果, 既学習文, さらに未学習の文を認識させた際にも, 各単語が正しく分節化されるとともに, 同じ品詞ごとに SOM 出力ニューロンに割り当てられた. 表 2 に結果を示す. 表中の数字は各単語に対するニューロンの発火回数を表す. また括弧内は未学習文に含まれていた単語数である.

表 2 単語生成時の Cf 発火に対する SOM 出力

SOM neuron	Words
1	look 51(9), glance 49(11)
2	upper 9(4), lower 9(1)
3	at 58(9)
4	-
5	blue 18(2), green 15(2), red 15(2)
6	-
7	right 15(5), left 15(3), up 6(1), down 6(1)
8	square 18(3), circle 22(4)
9	quickly 34(4), slowly 28(7)

#### 4.1 結果 : 動作からの文連想

感覚運動系情報を与えて, 文を連想させた. まず格 RNN における, 非文章 (文法的に正しくない文) の生成率は, 既学習データにおいて 0.75% (3/400), 未学習データにおいて 16% (16/100), 全体では 3.8%であった. また内容語 RNN において, 感覚運動情報と格パターンを与えた際の, 単語連想失敗率は既学習データにおいて 0.34% (5/1485), 未学習データで 5.4% (19/355), 全体では 1.3%であった.

連想が適切かどうか, つまり文法及びビスペルが正確で誤った情報を含まない, という評価においては, 既学習動作では 96.0% (384/400)と高い正解率であったが, 未学習動作では 37.0% (37/100)にとどまった. しかし, この未学習データでは, “もっともらしい (人間からみて解釈可能)”と思われる文章が生成された. 以下に事例を示す.

図 3 に既学習動作の例を示す. 右上の Green Circle を Look (視線を移動後維持) するという動作である. このとき格 RNN は動作系列と画像から, ‘1’, ‘2’, ‘7’, ‘9’を出力し, この格 RNN 出力及び動作画像情報から, 内容語 RNN が‘look’, ‘upper’, ‘right’, ‘slowly’を出力した. 正確に動作を表現した文章が生成されていることが確認できる.

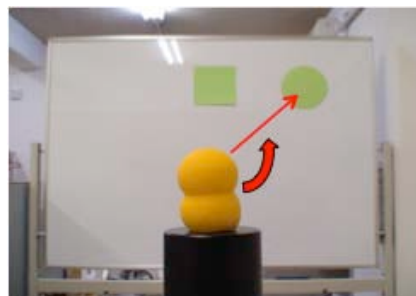


図 3 適切な文章の生成の例 (既学習データ)

対して図 4 に未学習動作の例を示す. 左下の何も対象が無い領域に視線を移動後, 元姿勢に戻る動作である. このとき格 RNN は, ‘1’, ‘3’, ‘8’, ‘9’を出力した. ここで‘8’は対象形状の品詞番号であり, この動作を表現するには適切ではない. しかしこの格出力に対応して, 内容語 RNN は‘glance’, ‘at’, ‘circle’, ‘quickly’を出力した.

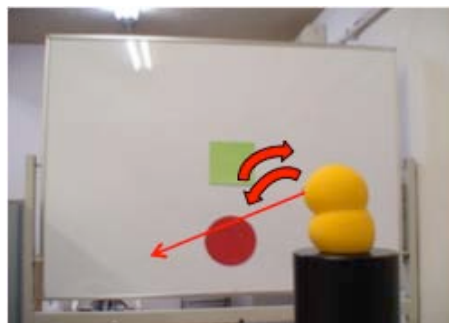


図 4 不適切な文章の生成の例 (未学習データ)

注視対象はこの動作には存在しないが、格 RNN の出力に対して内容語 RNN は、その領域に最も近い物体である Circle を出力として選択した。このように最終的には誤りと判断される文章出力であっても、本モデルでは現実世界において類似していると判断できる文章が生成される。

このように提案モデルは、全体的に十分な能力を有しているものの、格 RNN の未学習のパターンへの汎化が困難であることが確認された。これは同一感覚運動状況を表現しうる各シーケンス候補が複数ありうる為で、何らかの文脈情報（例えば行為の意図など）による拘束が別途必要である可能性を示唆していると思われる。

#### 4.2 結果：文からの動作連想

提案モデルに既学習文章もしくは未学習文章と初期視覚情報を与えて、動作を連想生成させた。理想的な動作時系列と生成された時系列データの二乗誤差を運動軸とステップ数で平均化したもので、その精度を評価した。その結果、全体の約 70% 程度が平均誤差以内にとどまることが確認された。

この平均誤差程度の例としては、モデルに “glancelowerleftquickly.” (glance lower left quickly.) という未学習文字列を与えた場合に、“glance lower left slowly.” に近い動作が生成される等の結果が観測された。また glance の入力に対し、動作が look になる等の出力も観察された。

このように完全な正解動作になる例は多くはないものの、人間からみた時に “それらしい” 文章解釈が実現されている判断できる。今後、これらの動作の評価法についても検討していく予定である。

### 5. まとめ

本稿では、神経力学モデルを備えたロボットによる言語獲得について報告した。本研究の課題は、単独では意味を成さない文字が二重分節構造を持って文を形成し、感覚運動系を通して知覚される実世界へと接地されるようなモデルを実現することである。具体的には、神経力学モデル Multiple Timescale Recurrent Neural Network (MTRNN) を用いて、モータ値および視野画像の時系列からなる行為パターンと文字系列からなる文の相互連想器を学習させた。

一連の実験から、我々のモデルにおいて単独では意味を成さない文字が二重分節構造を持って文を形成し、感覚運動系を通して知覚される実世界へと接地される過程が、限定的ではあるが確認できた。

今後は、神経力学モデルにおいて、文字列から文を学習する場合の、文章構造の複雑さの限界、単語数のスケーラビリティ等について詳細な実験行っていく。また本モデルでは、格構造を文章から獲得したが、本来、実世界の行為現象自体が [誰が][何を][どうした] といったように組み合わせ構造を持っている。この実世界の組み合わせ機能が、まず先に構造化され言語に転用される機構の方が、発達のモデルとして自然だと

も考えられる[8]。

これらの考察をふまえて、さらに実世界認知と言語認知の力学系によるカップリング研究を展開していく予定である。

謝辞：

本研究は、JST さきがけ「情報環境と人」、科研費学術創成研究 (19GS0208)、科研費基盤研究 (B) (21300076)、科研費基盤研究 (S) (19100003) 及び GCOE の支援を受けた。

### 参 考 文 献

- [1] Harnad, S.: The symbol grounding problem, *Physica D: Nonlinear Phenomena*, Vol. 42, pp. 335–346 (1990).
- [2] Sugita, Y. and Tani, J.: Learning semantic combinatoriality from the interaction between linguistic and behavioral processes, *Adaptive Behavior*, Vol. 13, No. 1, pp. 33–52 (2005).
- [3] Ogata, T., Murase, M., Tani, J., Komatani, K. and Okuno, H. G.: Two-way Translation of Compound Sentences and Arm Motions by Recurrent Neural Networks, *Proc. of IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS-2007)*, pp. 1858–1863 (2007).
- [4] Dominey, P. F. and Rumus, F.: Neural network processing of natural language: I. Sensitivity to serial, temporal and abstract structure of language in the infant, *Language and cognitive process*, Vol. 15, pp. 87–127 (2000).
- [5] Y. Yamashita and J. Tani, “Emergence of Functional Hierarchy in a Multiple Timescale Neural Network Model: a Humanoid Robot Experiment,” *PLoS Comput. Biol.*, vol.4, 2008.
- [6] W. Hinoshita, H. Arie, J. Tani, H. G. Okuno, T. Ogata: Emergence of Hierarchical Structure mirroring Linguistic Composition in a Recurrent Neural Network, *Neural Networks 12*, pp.311-320, Jan. 12. 2011.
- [7] H. Kozima, C. Nakagawa, and H. Yano, “Using robots for the study of human social development,” *AAAI Spring Symposium on Developmental Robotics*, 2005.
- [8] 乾敏郎: 言語獲得と理解の脳内メカニズム, *The Japanese Journal of Animal Psychology*, Vol. 60, No. 1, pp. 59–72 (2010).