1O2-1

シーケンス変換モデルに基づくロボットによる言語獲得

Language Acquisition by Robots Based on Sequence to Sequence Learning

高渕 健太[*1] 坂本 伸次[*1] 植田 紗也佳[*1] 岩橋 直人[*1] 國島 丈生[*1]
Kenta Takabuchi Shinji Sakamoto Sayaka Ueda Naoto Iwahashi Takeo Kunishima

[*1] 岡山県立大学
Okayama Prefectural University

In this paper, we propose the method that enables robots to learn language based on sequence to sequence learning. The originalities of the proposed method is as follows: 1) Language acquisition problem is formulated by the statistical machine learning problem. 2) Morphological analysis is not necessary in the processes of learning and conversion. As sequence to sequence learning methods, we adopted the IBM Model 4 and encoder-decoder translation model. We got very promising results using IBM Model 4.

1. はじめに

人とロボットの円滑な言語コミュニケーションを実現するために重要なことは，言語と実世界の事物との対応付けが，ロボットによって適切に理解されることである．従来，ロボットが内部に持っている記号を実世界の事物にいかに対応させるかという問題は，記号接地問題[Harnad 90]として研究が行われてきた．植田らの研究[植田 15]では，構文テンプレートを獲得することにより，命令文と物体の動作データから抽出した深層格情報との対応付けを学習することを可能とした(図 1)．また，高野により，身体運動と言語との関連付けにより，動作を言語化する研究が行われた[高野 13]．しかし，従来の研究では，内容語として名詞，動詞のみからなる単純な構文構造しか扱われていなかった．

そこで，本研究では，内容語として名詞と動詞に加えて形容詞を含む命令文を対象にした，シーケンス変換モデル[Sutskever 14] に基づくロボットによる言語獲得手法を提案する．シーケンス変換モデルとして，IBM Model 4 [Brown 93]，および Encoder-decoder 翻訳モデル[Cho 14]を別々に用いる．

本提案手法の特徴は以下の 2 点である: 1) 言語獲得における記号接地問題を統計的機械翻訳の問題としてモデル化した．2) 形態素解析を必要とせず，言語情報として音節列を直接ロボットの内部記号に変換することが可能である．

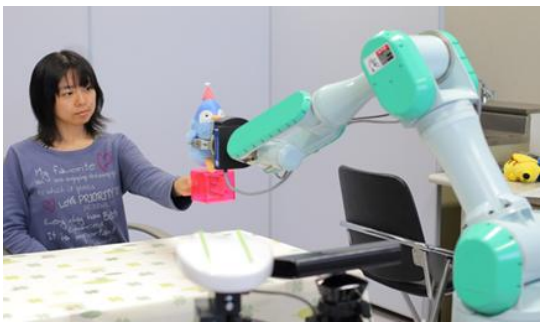

図1 ロボットとのインタラクション

2. 提案手法

提案手法の概要を図 2 に示す．まず，音声認識器を用いて，音声信号を音節記号列に変換する．また，動作の動画像データから，意味情報として，記号列で表現される深層格情報を抽出する．そして，これらの音節記号列と深層格情報をシーケンス変換モデルにより対応付ける．シーケンス変換モデルとは，入力記号列の長さと出力記号列の長さが異なる翻訳をモデル化するものである．つまり、翻訳において 1 対多，もしくは多対 1 の翻訳を扱うためのモデルである．従来は，原言語から目的言語への翻訳を扱う．一方，本研究では，原言語を音節記号列，目的言語を深層格情報，または逆に，原言語を深層格情報，目的言語を音節記号列に置き換えることで言語獲得を翻訳問題としてモデル化する．シーケンス変換モデルとして IBM Model 4 を使った手法と Encode-decoder 翻訳モデルを使った手法を考える．

図2 提案手法の概要

2.1 動作動画像からの深層格情報の抽出

深層格情報とは，「KIIRO KOPPU LND AO HAKO TRJ NOSERU」のように表記される，動画像内の物体の動きの深層格[Fillmore 75] を明らかにしてテキスト列で書き表した内部記号列で表現されるものである．深層格情報は複数の物体と，物体の動きの軌跡を記述した動画像から，参照点に依存した HMM[羽岡 00]を用いて抽出する．

提案手法で扱う深層格情報は,物体の名前，大きさ，色，物体の動作，動作の対象となるトラジェクタ，そして，動作の目標になる目標格、源泉格、場所格であるランドマークの合計 6 カテゴリの記号から構成される．それぞれのカテゴリには，物体の名前 10 種類，色 4 種類，大きさ 2 種類，動作の名前 6 種類，トラジェクタ 1 種類，ランドマーク 1 種類の記号が含まれる．

図 3に動作動画像とそこから抽出された深層格情報の例を示す。Kinect によって認識された物体の色，形，大きさといった特徴から得られる特徴ベクトルから物体クラスを得られる．また，認識された物体の座標の変位を記録することでランドマークやトラジェクタ，動作等の深層格情報を得ることができる．図 3中において，物体に重なる番号が物体クラスを示しており，物体の座標の変位は物体に連なる白い軌跡で示されている．軌跡が描かれている物体が動作の対象となるトラジェクタ(物体クラス 22，30)であり，動作の参照点となる物体がランドマーク(物体クラス 21，29)である．

連絡先：高渕健太，岡山県立大学大学院人工知能学研究室，
〒714-1197 岡山県総社市窪木 111，0866-94-2092，
E-mail:cd27028@c.oka-pu.ac.jp

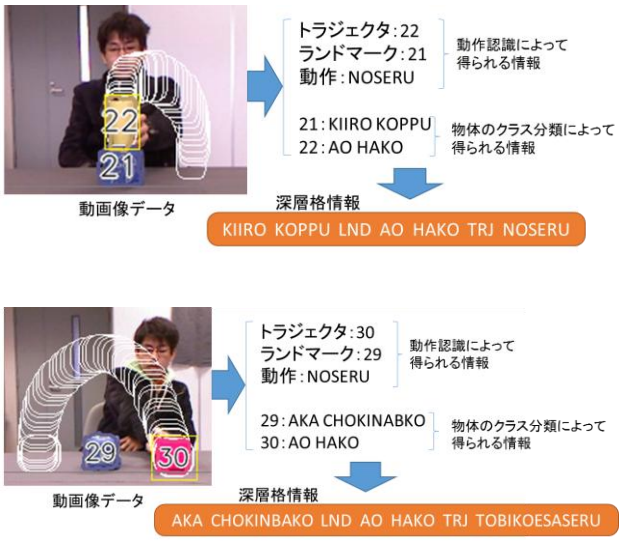

図3　深層格情報取得の流れ

2.2 IBM Model 4 を使ったシーケンス変換

命令文と深層格情報を統計的機械翻訳の問題であるとし，命令文 F が深層格情報Eに翻訳される確率を次式のように計算する．これを最大にする $\hat{E}$ が翻訳結果として出力される．

$$P(F \mid E)P(E) \tag{1}$$

$$\hat{E} = \arg\max_{E} P(F \mid E)P(E) \tag{2}$$

$P(E)$は言語モデル，$P(F|E)$は翻訳モデルと呼ばれ，言語モデルによって翻訳される命令文が決まり，言語モデルによって命令文の言語と深層格情報の対応関係が学習される．提案手法では言語モデルとして n-gram モデル，翻訳モデルとして IBM Model 4 を採用する．IBM Model 4 はグラフィカルモデルに基づいた翻訳モデルであり，IBM Model において 1 対多の翻訳を可能にし，翻訳における目的言の単語の相対位置を考慮したモデルである．

IBM Model 4 による翻訳の例を図 4に示す．「NULL」は二言語間で直訳できる単語がない場合にその代わりとして利用される．その例として日本語の助詞や数詞，英語の冠詞などが挙げられる．IBM Model 4 による翻訳を実装するために KenIM[KenLM]，Gizza++[Giza++]，Moses[Moses]を利用した．

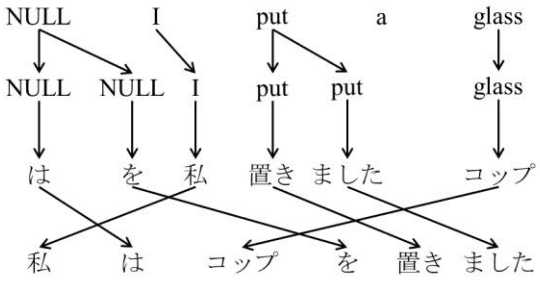

図4 IBM Model4 による翻訳の例

2.3 Encoder-decoder 翻訳モデルを使ったシーケンス変換

Encoder-decoder 翻訳モデル[Cho 14]はリカレントニューラルネットワーク(RNN)を組み合わせた機械翻訳モデルのことである．実装した Encoder-decoder 翻訳モデルを図 5に示す。

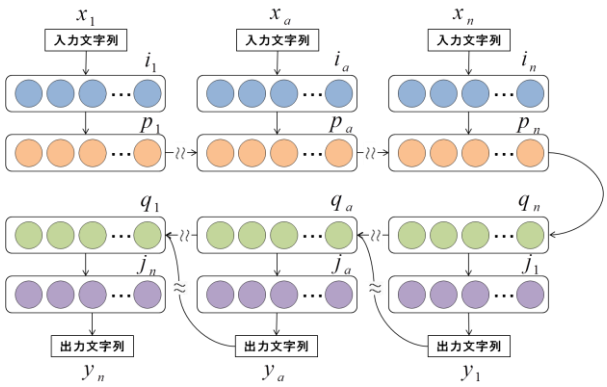

図5 Encoder-decoder 翻訳モデル

次に，Encoder-decoder 翻訳モデルの各層のリンクに関して記述する．以下に示すように各層から次の層へのリンクは W によって重み付けされる．

$$\mathbf{i}_n = \tanh(W_{xi} \cdot \mathbf{x}_n) \tag{2}$$

$$\mathbf{p}_n = \mathrm{LSTM}(W_{ip} \cdot \mathbf{i}_n + W_{pp} \cdot \mathbf{p}_{n-1}) \tag{3}$$

$$\mathbf{q}_1 = \mathrm{LSTM}(W_{pq} \cdot \mathbf{p}_{|\mathbf{w}|}) \tag{4}$$

$$\mathbf{q}_m = \mathrm{LSTM}(W_{yq} \cdot \mathbf{y}_{m-1} + W_{qq} \cdot \mathbf{q}_{m-1}) \tag{5}$$

$$\mathbf{j}_m = \tanh(W_{qj} \cdot \mathbf{q}_m) \tag{6}$$

$$\mathbf{y}_m = \mathrm{softmax}(\mathrm{W_{jy}} \cdot \mathbf{j}_\mathrm{m}) \tag{7}$$

i, j 層は埋め込み層と呼ばれ，単語情報を表す層で，p,q 層は隠れ層と呼ばれる．隠れ層に LSTM を用いた理由は，翻訳における注目単語とその前後に出現する単語の依存関係を記憶する必要があるためである．Encoder-decoder 翻訳モデルは Chainer[Chainer] で実装した．

3. 実験

提案手法の有用性を検証するために，3.1 節と 3.2 節で示す 4 つの条件（IBM-M, IBM-S, ED-M, ED-S）で，音節記号列と深層格情報との双方向変換の実験を行った．まず，動作動画像と，その動作を指示する命令文音声のペアを，学習用として 500 ペア，テスト用として 50 ペア用意した．テストデータは学習データのサンプルに含まれないものとした．動作動画像からの深層格情報の抽出は，2.1 節で記述した手法を用いた．音声から音節列を抽出のための音声認識には Julius[Julius]を使用した．表 1 に命令文と深層格情報のペアの例を示す．

表1　実験に使用した命令文の例

命令文(音節記号列)	深層格情報
あ お の と と ろ お も ち あ げ て	AO TOTORO TRJ MOCHIAGERU
み ど り の こ っ ぷ か ら あ か の は こ お は な し て	MIDORI KOPPU LND AKA HAKO TRJ HANASU
あ お の き ん ぎょ お も ち あ げ て	AO KINGYO TRJ NOSERU

3.1 IBM Model 4

(1)　人手で書き起こした音節記号列(IBM-M)

音節記号列は，人手によって書き起こした，認識誤りがないものとした．

(2)　音声認識により得られた音節記号列(IBM-S)

音節記号列は，音節認識器の出力結果を用いた．

3.2 Encoder-decoder 翻訳モデル

(1)　人手で書き起こした音節記号 (ED-M)

3.1で使用したデータを用いた．

(2)　音声認識により得られた音節記号列(ED-S)

3.1(2)で使用したデータを用いた．

4.　結果

4.1 評価尺度

実験結果の評価には，実験によって得られたデータと翻訳の正解データとの編集距離を用いる．また，認識精度の観点から以下の式であらわされる評価値を導入する．

$$A\_s = \frac{\text{編集距離が 0 の文章数}}{\text{文章数}} \times 100 \quad [\%] \qquad (8)$$

$$A\_w = \left(1 - \frac{\text{編集距離の合計}}{\text{音節数の合計}}\right) \times 100 \quad [\%] \qquad (9)$$

A_s と A_w をそれぞれ文精度，単語精度と呼ぶことにする．

4.2 音節認識精度

Julius による音声認識の結果は，文精度 0%，単語精度 64% であった．

4.3 IBM Model 4

(1)　人手で入力した音節記号列(IBM-M)

IBM Model 4 におけるテキストデータを使った実験では,双方向の変換で，テストデータ 50 ペア全てにおいて編集距離が 0 になる変換結果を得る事ができた(図 6, 7, 8, 9)．

(2)　音声認識により得られた音節記号列(IBM-S)

音節記号列から深層格情報への変換結果と正解データを比較すると，A_s=80，A_w=87 が得られる(図 6, 7)．これらのことより，Julius による音声認識の精度があまり高くない場合でも IBM Model 4 による変換は有効であると言える．また，深層格情報から音節記号列への変換では，A_s=2，A_w=64 が得られた(図 8, 9)．

4.4 Encoder-decoder 翻訳モデル

(1)　人手で入力した音節記号(ED-M)

音節記号列から深層格情報への変換では A_s=46.0，A_w=73.1 となった(図 6, 7)．また，深層格情報から音節記号列への変換では A_s=46，A_w=63 が得られた(図 8, 9)．

(2)　音声認識により得られた音節記号列(ED-S)

音声認識を行ったテストデータ 50 文の命令文から深層格情報の変換結果より， A_s=36、A_w=78 となり(図 6, 7)， IBM Model 4 を使った場合に比べ，精度は低くなった．また，深層格情報から音節記号列への変換では A_s=0，A_w=11 が得られた(図 8, 9)．

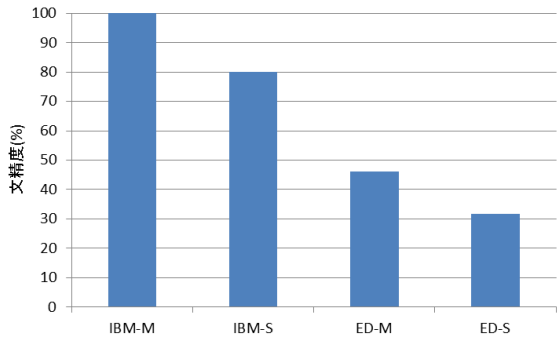

図 6　音節記号列から深層格情報への変換　(文精度)

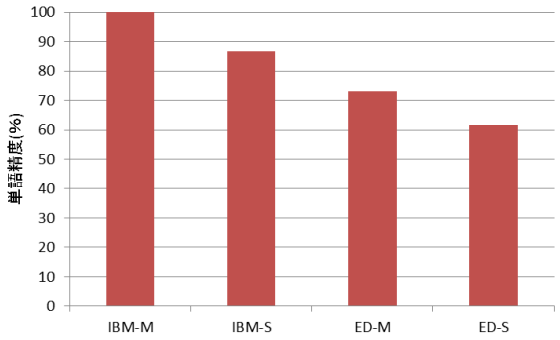

図 7　音節記号列から深層格情報への変換
(単語精度)

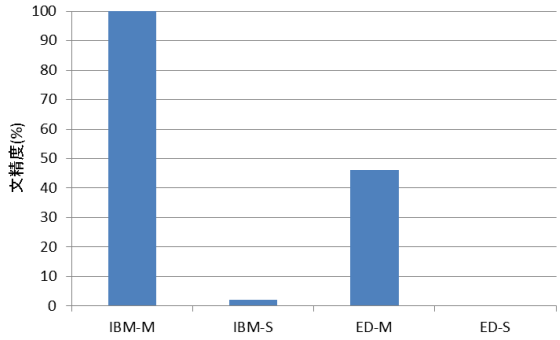

図 8　深層格情報から音節記号列への変換(文精度)

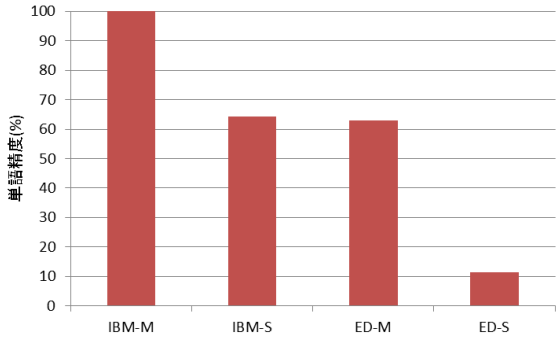

図 9　深層格情報から音節記号列への変換

(単語精度)

5. 考察

人手で入力した音節記号列を用いて，IBM Model 4 を学習した実験結果では，極めて良好な実験結果を得ることができた．音声認識結果を用いて，IBM Model 4 を学習した場合でも，音節記号列から深層格情報への変換において良好な結果が得られた．本研究での Julius による音声認識精度は 64%程度であったにもかかわらず良好な結果が得られたことから，音声認識精度に左右されない翻訳が可能であると言える．深層格情報から音節列への翻訳で得られた精度は Julius による音声認識の精度とほぼ同程度の値になっていることから，IBM model 4 を使った翻訳精度は音声認識率の改善によってより良い翻訳結果を得られると期待できる．

人手で入力した音節記号を用いた Encoder-decoder 翻訳モデルの実験結果より，単語精度と文精度の差が大きく開いている．したがって，単語の変換に比べて文章構造を正しく学習できていなかったと考えられる．音声認識結果を用いた Encoder-decoder 翻訳モデルの実験結果では，双方向の翻訳で良い結果は得られなかった．また，深層格情報から音節列への単語精度は Julius の認識率と比べて極めて低い．

実験全体を通して Encoder-decoder 翻訳モデルでの精度が IBM Model 4 による実験と比べて低い．これはニューラルネットワーク特有の長期記憶を保持させることの難しさから発生している現象と考えられる．1 つの深層格情報に対応する音節は多いもので 7 音節ある．これに加えて音節記号列で表現された文章を長期記憶として保持して学習すると考えると，翻訳モデルは複雑になることが予想される．本研究では LSTM を使って長期記憶を保持させようとした 1 つの深層格情報と複数の音節対応と音節列で表現された文章構造の学習で，LSTM がうまく機能しなかったと考えられる．また，翻訳モデルの学習中に過学習が起きていることが考えられる．従来，ニューラルネットワークを使った研究では学習データは膨大な数になることが多い．その点で，本研究の学習データが十分な数であったかは今後の課題として取り組んでいきたい．過学習の低減のためには，学習データを増やすことと，より音節記号列と深層格情報の変換に適した翻訳モデルを構築する必要があると考える．

これらのことから，音節記号列と深層格構造の双方向の翻訳で IBM Model 4 を使った方が有効であると言える．

6. まとめ

言語獲得の問題を，シーケンス変換モデルに基づく翻訳問題でモデル化し，形態素解析を必要としない手法を提案した．IBM Model 4 を使った手法では比較的良好な結果を得る事ができた．音声認識誤りに左右されない翻訳を行える可能性を示した．Encoder-decoder 翻訳モデルでは音節列にから深層格情報への翻訳では良い結果を得られなかった．

今後はこのようなことが起きた原因の検証，改善を行い，より汎用的な命令にも対応できる手法を開発する予定である．

謝辞

本研究は，JSPS 科研費 15K00244，および，JST CREST「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」の助成を受けたものです．

参考文献

[Brown 93] Petet F. Brown, Stephen A Della, Vincent J, Della Pietra, Robert L. Mercer : The Mathematics of Statistical Machine Translation, Parameter Estimation, Computational Linguistics Vol.19 No.2, pp.263-311, 1993.

[Chainer] http://chainer.org/

[Cho 14] Kyunghyun Cho, Bart van Merrienboerr, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio.: Learning phrase representations using RNN encoder-decoder for statistical machine translation, In Proceedings of the 2014 Conference on EMNLP, 2014.

[Fillmore 75] Charles J. Fillmore: TOWARD A MODERN THEORY OF CASE & OTHER ARTICLES, (邦訳:田中春美, 船城道雄: 格文法の原理 言語の意味と構造, 三省堂, 1975).

[Giza++] https//code.google.com/p/giza-pp/

[羽岡 00] 羽岡哲郎, 岩橋直人: 言語獲得のための参照点に依存した空間的移動の概念の学習, 信学技報 TECHNICAL REPORT OF IEICE, PRMU2000-105, pp.49-58, 2000.

[Harnad 90] Stevan Harnad: The Symbol Grounding Problem, Pysica D, Vol.42,pp335-346, 1990.

[Julius] http://julius.osdn.jp

[KenLM] https://kheafield.com/code/kenlm/

[Moses] http://www.statmt.org/moses/

[Sutskever 14] Ilya Sutskever, Oriol Vinyals, Quoc Le: Sequence to Sequence Learning with Neural Networks, NIPS, 2014.

[高野 13] 高野渉, 中村仁彦: 全身運動から言語空間の構築と運動認識への応用, 人工知能学会論文誌 28 巻 4 号 A, 2013.

[植田 15] 植田紗也佳, 岩橋直人, 國島丈生: ロボットによる実世界情報を用いた付属語の獲得, JSAI2015, 2D3-OS-12b-3, 2015.