Vol.2016-NL-227 No.10
2016/7/30

動画像情報と音声情報のシーケンス変換学習に基づく言語獲得

高渕健太† 岩橋直人† 國島丈生†

概要：本研究では，動画像情報と音声情報のシーケンス変換学習に基づくロボットの言語獲得手法を提案する．提案手法では，概念構造を表す記号列と音節列の相互変換を学習する．シーケンス変換学習として統計的機械翻訳手法である IBM Model4 とニューラルネットワークに基づく機械翻訳手法である Encoder-decoder モデルの 2 通りを試す．本提案手法の特徴は以下の2点である．1)動画像情報と音声情報の変換の学習を機械翻訳問題としてモデル化する．2) 形態素解析を必要とせず，音節列と概念構造を変換することができる．実験により，高い精度で動画像情報と音声情報が相互変換できることを確認した．

キーワード：言語獲得，シーケンス変換学習，機械翻訳

Language Acquisition Based on Sequence to Sequence Learning Between Video Information and Speech Information.

Kenta Takabuchi† Naoto Iwahashi† Takeo Kunishima†

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

Keywords: Language acquisition, Sequence to sequence learning, Machine translation

1. はじめに

ロボットと人の円滑な対話コミュニケーションのためには，ユーザごとに異なる環境の中でロボットが言語獲得をしなければならない．言語獲得において重要なことは，実世界の概念構造と言語の対応を正しく学習することである[1,2]．植田ら[3]の研究では，構文テンプレートの学習により，動作データから抽出した概念構造と命令文の対応の学習を可能にした（図 1）．しかし，従来の研究では，内容語として名詞，動詞からなる単純な構文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，格助詞を含む命令文を対象とし，動画像情報と音声情報のシーケンス変換学習によって言語獲得を行う手法を提案する．シーケンス変換学習として統計的機械翻訳手法である IBM Model 4[4]とニューラルネットワークに基づく機械翻訳手法である Encoder-decoder モデル[5]の 2 通りを試す．提案手法の特徴は以下の 2 点である．1）動作情報と音声情報の変換を機械翻訳問題としてモデル化する．2）形態素解析を必要とせず，音声情報と動作情報を相互変換する．形態素解析を必要としない手法によって，日本語，英語などの発話言語に依存しないロボットの言語獲得が可能であると考えている．

† 岡山県立大学
Okayama Prefectural University

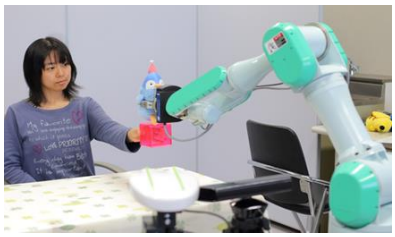

図 1 ロボットとのコミュニケーション

2. 提案手法

提案手法の概要を図 2 に示す．提案手法では，音声情報として，音声認識によって得られる音節列，動画像情報として動画像の動作に関連する概念構造（2.1 節）を扱う．これらの情報の相互変換をシーケンス変換学習によって学習する．従来のシーケンス変換学習は，原言語から目的言語への翻訳に用いられる．シーケンス変換学習の特徴として，1 対多，多対 1 の翻訳が可能であることがあげられる．本研究では，原言語と目的言語を音節列，概念構造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．

シーケンス変換学習には，IBM Model 4 を使った手法と Encoder-decoder 翻訳モデルを使った手法を提案する．

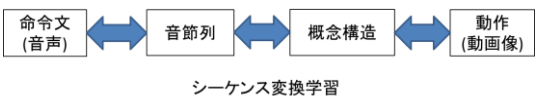

図 2 提案手法の概要

2.1 動画像からの概念構造の抽出

概念構造とは，動画像内の動作に関連した物体の深層格[6]を明らかにして記号列で書き表したものである．概念構造は複数の物体と，物体の動きの軌跡を記述した動画像から参照点に応じた HMM[7]によって抽出する．

提案手法で用いる概念構造は，物体の名前，大きさ，色，人が行う動作，深層格情報の 6 カテゴリで構成される．それぞれのカテゴリには，物体の名前 10 種類，大きさ 2 種類，色 4 種類，動作 6 種類，深層格情報 2 種類の情報が含まれる．深層格情報として本研究では，トラジェクタ，ランドマークを用いる，トラジェクタは，動画像内で動作の対象になる物体に付与され，ランドマークは動作の参照点になる物体に付与される．

動画像と，そこから抽出される概念構造の例を図 3 に示す．Kinect から得られる深度情報を含む動画像から，物体の色，形，大きさなどの情報を用いて物体認識を行い，物体クラスを得る．また，認識された物体の座標の変位を記録することでトラジェクタ，ランドマーク，動作の概念構造を得ることができる．図中において，物体に重なる番号が物体クラスを示しており，物体に連なる白い線がその物体の座標の変位を示している．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），動作の参照点になる物体がランドマーク（物体クラス 21, 29）である．図 3 (1) の動画像から得られる概念構造は「KIIRO KOPPU LAN AO HAKO TRJ NOSERU」，(2) の動画像から得られる概念構造は「AKA CHOKINBAKO LND AO HAKO TRJ TOBIKOESASERU」となる．「LND」，「TRJ」はそれぞれランドマーク，トラジェクタを表す記号で，それぞれ物体を表す直前の記号がランドマーク，トラジェクタであることを表す．

図 3　概念構造抽出の流れ

2.2 IBM Model 4 を使ったシーケンス変換学習

概念構造と音節列の変換を機械翻訳問題であるとし，音節列 F が概念構造 E に変換される確率を以下の式（1）のように計算する．音節列 F が与えられたときの変換結果 $\hat{E}$ は式（1）を最大化する E が出力される．

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

$$\hat{E} = \underset{E}{\operatorname{argmax}}\, P(F \mid E)P(E) \tag{2}$$

$P(E)$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 によるシーケンス変換学習を実装するために，KenLM[8]，Giza++[9]，Moses[10]を利用した．

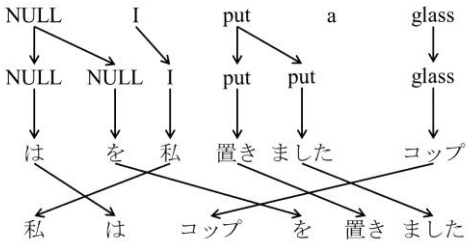

図 4　IBM Model 4 による翻訳の例

2.3 Encoder-decoder 翻訳モデルを使ったシーケンス変換学習

Encoder-decoder 翻訳モデル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．

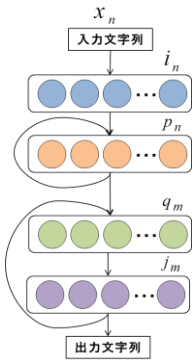

図 5　Encoder-decoder 翻訳モデル

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．

$$\mathbf{i}_n = \tanh(W_{xi} \cdot \mathbf{x}_n) \tag{3}$$

$$\mathbf{p}_n = \mathrm{LSTM}(W_{ip} \cdot \mathbf{i}_n + W_{pp} \cdot \mathbf{p}_{n-1}) \tag{4}$$

$$\mathbf{q}_1 = \mathrm{LSTM}(W_{pq} \cdot \mathbf{p}_{|\mathbf{w}|}) \quad (5)$$

$$\mathbf{q}_m = \mathrm{LSTM}(W_{yq} \cdot \mathbf{y}_{m-1} + W_{qq} \cdot \mathbf{q}_{m-1}) \quad (6)$$

$$\mathbf{j}_m = \tanh(W_{qj} \cdot \mathbf{q}_m) \quad (7)$$

$$\mathbf{y}_m = \mathrm{softmax}(\mathrm{W_{jy}} \cdot \mathbf{j}_m) \quad (8)$$

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[11]で実装した．

3. 実験

提案手法の検証をおこなうために，3.1 節で示す条件で概念構造と音節列の相互変換を行った．まず，動画像と音声データ 1,000 ペアを用意した．動画像から概念構造を抽出し，音声データは音声認識器を使って音節列にテキスト化する．この概念構造と音節列 1,000 ペアに対して 10 分割交差検定を行った．行った実験の条件を表 1 に示す．また，概念構造と音節列の例を表 2 に示す．音声データから音節列を得るための音声認識には Julius[12]を使用した．

表 1　実験条件

	IBM Model 4	Encoder-decoder 翻訳モデル
人手	IBM-M	ED-M
音声認識	IBM-S	ED-S

表 2　実験に使用したデータペアの例

音節列	概念構造
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

4. 実験

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．

$$\text{文精度} = \frac{\text{編集距離が 0 の文章数}}{\text{文章数}} \times 100 \quad [\%] \quad (9)$$

$$\text{単語精度} = \left(1 - \frac{\text{編集距離の合計}}{\text{単語数の合計}}\right) \times 100 \quad [\%] \quad (10)$$

ただし，概念構造から音節列への変換では，単語精度の評価は音節になるので音節精度と言い換える．

4.2 音声認識精度

Julius による音声認識の精度は，文精度 7.1%，単語精度 67.5%であった．

4.3 IBM Model 4

(1) 人手で書き起こした音節列（IBM-M）

IBM Model 4 による人手で書き起こしたデータを使った実験では音節列から概念構造への変換において文精度 92.5%，単語精度 97.1%という非常に良い結果が得られた．概念構造から音節列への変換では，文精度 71.9%，音節精度 91.8%となった．

(2) 音声認識によって得られた音節列（IBM-S）

音声認識結果を使った音節列から概念構造への変換では，文精度 64%，単語精度 76.7%となり，概念構造から音節列への変換では，文精度 10.6%，音節精度 73.7%となった．

4.4 Encoder-decoder 翻訳モデル

(1) 人手で書き起こした音節列（ED-M）

人手で書き起こした音節列から概念構造への変換では文精度 82.6%，単語精度 93.1%となり，比較的良好な結果が得られた．概念構造から音節列への変換では，文精度 73%，音節精度 90%となった．

(2) 音声認識によって得られた音節列（ED-S）

音声認識結果を使った音節列から概念構造への変換では，文精度 55.6%，単語精度 78.1%となり，概念構造から音節列への変換では，文精度 6.4%，音節精度 34.1%となった．

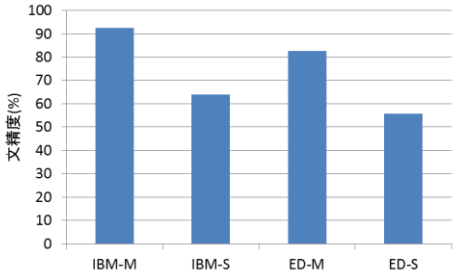

図 6　音節列から概念構造への変換（文精度）

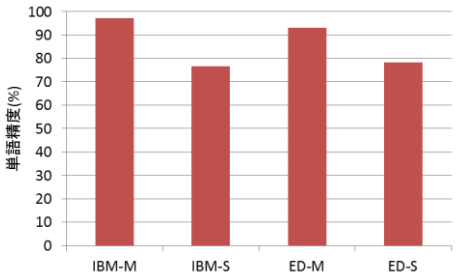

図 7　音節列から概念構造への変換（単語精度）

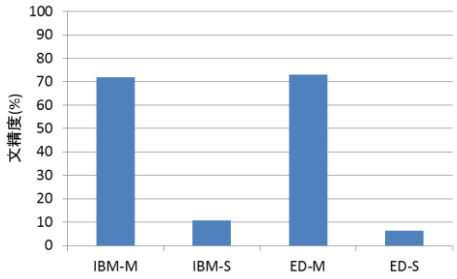

図 8　概念構造から音節列への変換（文精度）

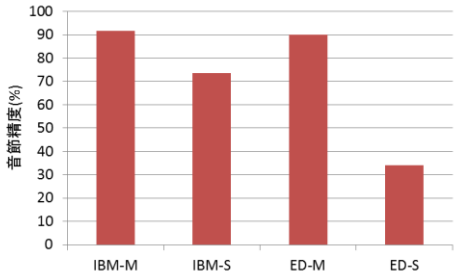

図 9　概念構造から音節列への変換（音節精度）

5. 考察

人手で書き起こした音節列を使った実験では，IBM Model 4, Encoder-decoder 翻訳モデルでそれぞれ良好な結果を得られた．特に，音節列から概念構造列への変換では，文精度，単語精度ともに良い結果を得られた．概念構造から音節列への変換では，音節精度では良い結果を得られたが，文精度では IBM Model 4，Encoder-decoder 翻訳モデルともに 70%程度の精度しか得られなかった．これらのことから，概念構造から音節列への変換の学習が音節列から概念構造への変換に比べて複雑な学習をしなければならなかったと考えられる．

音声認識結果による音節列から概念構造への変換では，IBM Model 4，Encoder-decoder 翻訳モデルで単語精度にはあまり差が見られないが文精度では IBM Model 4 を用いた手法の方が，10%程精度が高いことが分かる．Julius の文精度，単語精度の値を見ると，文精度 7.1%，単語精度 67.5%であったので，それぞれの手法で音節列に音声認識誤りがある場合でも正しく概念構造に変換することが可能であると考える．また，概念構造から音節列への変換では，音節精度で IBM Model 4 を用いた手法が Encoder-decoder 翻訳モデルの手法に比べて高くなっている．このことから，音声認識誤りのある音節列に対して，IBM Model 4 を用いた手法が Encoder-decoder 翻訳モデルによる手法に比べてより頑健であると考えられる．

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くなる結果となった．また，音声認識誤りに大きく影響されることも分かった．これらの原因として，学習データが不足していることが考えられる．これらは今後の課題として取り組んでいきたい．

また，本研究の音声データは，構文が限定されているものしか扱っていない．本来の言語獲得では，構文に依存しない複雑なデータを扱うべきであると考えるが，本研究の動画像からの概念構造の抽出の複雑さの点から実験では限られた構文に従うデータのみを扱った．

本稿の実験結果から，音声認識誤り，データ量の観点を考慮すると，概念構造と音節列の相互変換で IBM Model 4 を使った手法が有効であると言える．

6. まとめ

言語獲得の問題を，概念構造と音節列のシーケンス変換学習に基づく機械翻訳問題でモデル化し，形態素解析を必要としない言語獲得手法を提案した．IBM model 4 と Encoder-decoder 翻訳モデルそれぞれによる実験から，特に IBM Model 4 による手法で良い結果を得られた．また，音節列の一部に音声認識誤りがあった場合でも正しく概念構造に変換することができた．

今後は，実験データの充実による変換精度の向上，概念構造抽出の改善を行う．また，他言語での実験や，より複雑な構文を用いた実験を行う予定である．

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．また，研究に協力していただいた植田紗也佳氏, 坂本伸次氏に感謝する．

参考文献

[1] Iwahashi, N.. Robots That Learn Language. Development Approach to Human-Machine Conversation. Human Robot Interaction (N. Snaker,Ed.), I-tech Education and Publishing. 2007, p.95-118.
[2] Taniguchi, T., Nagai, T., Nakamura, T., Iwahashi, N., Ogata, T. and Asoh, H.. Symbol emergence in robotics: a survey. Advanced Robotics, 2016, Vol 30, Issue 11-12, p.706-728.
[3] 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.
[4] Brown, P. E, Pietra, V. J. D., Pietra, S. A. D. and Mercer, R. L.. The Mathematics of Statistical Machine Translation，Parameter Estimation, Computational Linguistics. 1993, Vol.19, no.2, p.263-311.
[5] Cho, K., Merrienboerr, V. B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H. and Bengio, Y.. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Proceedings of the 2014 Conference on EMNLP.
[6] Fillmore, C. J.. 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)
[7] 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 TECH-NICAL REPORT OF IEICE, PRMU2000-105, 2000, p.49-58.
[8] KenLM. https://kheafield.com/code/kenlm
[9] Giza++. https://code.google.com/archive/p/giza-pp/
[10] Moses. http://www.statmt.org/moses/
[11] Chainer. http://chainer.org/
[12] Julius. http://julius.osdn.jp