DNN-based object manipulation to syllable sequence for language acquisition

Kenta Takabuchi, Naoto Iwahashi, Ye Kyaw Thu, and Takeo Kunishima

Okayama Prefectural University, Japan
(Tel: 81-866-94-2001, Fax: 81-866-94-2199)

{cd27028h, iwahashi, ye, kunishi}@c.oka-pu.ac.jp

Abstract: A deep neural network (DNN)-based language acquisition method that can convert object manipulation videos to their descriptions in syllable sequences is presented. The proposed method is novel in that the conversion is achieved without any information on words and explicit clustering of object images. It enables language learning with only a small amount of training data by combining a convolutional neural network (CNN), reference-point-dependent hidden Markov models (RPD-HMM), and a recurrent neural network (RNN). In experiments conducted using only five hundred pairs, each consisting of a video and its spoken description as training data, a bilingual evaluation understudy (BLEU) score of 96 was obtained.

Keywords: DNN, CNN, RNN, language acquisition, motion-to-syllable learning

1 INTRODUCTION

Language acquisition is a challenging topic in the artificial intelligence research area, but is essential for practical communication between robots and humans. Living assistance robots that communicate with humans need to have mutual understanding of the surrounding environment with humans. For example, if a human gives a command such as 'put the stuffed toy on a box', a robot should be able to differentiate the stuffed toy from the box, determine the physical location of each, and also understand the meaning of the word put. Here, the meaning of the word put includes knowledge of the kind of action that should be taken. To this end, Iwahashi [1] and Iwahashi et al. [2] proposed a developmental approach that enables robots to learn linguistic communication capabilities from scratch based on little verbal and nonverbal interaction, such as behavioural information, with humans. Takabuchi et al. [3] proposed a language acquisition method for robots that utilises phrase-based statistical machine translation without any information on words. Further, language acquisition by robots has been attracting interest in various research fields [4][5].

In this study, we applied deep neural networks (DNN) for language acquisition from object manipulation video. The proposed method made it possible to convert the motion and the image features of objects directly to syllable sequences which describe them without any information on words and explicit clustering the image features. Our approach is theoretically sequence to sequence learning [6][7]. We evaluated the bilingual evaluation understudy (BLEU) [8] score and the rank-based intuitive bilingual evaluation measure (RIBES) [9], obtained for conversion from object manipulation videos to Japanese syllable sequences.

2 PROPOSED METHOD

2.1 Overview

An overview of our proposed method is given in Fig. 1. First, we use an in-house online machine learning toolkit called L-Core [2] to detect image objects, annotate them with individual IDs, trace moving objects, and recognise trajector and landmark from object manipulation videos created using Microsoft Kinect v1. Second, feature extraction is carried out to extract image features from the image of the segmented objects using a convolutional neural network (CNN) [10]. We use the CAFFE deep learning framework for image features extraction [11] with an open trained network model of IMAGENET [12]. However, because features from 4096 dimensions are too large to learn with a recurrent neural network (RNN) using only a small amount of training data, we compress them to 256 dimensions with principal component analysis (PCA). Third, the object motion trajectories are classified into motion classes with the pre-trained reference-point-dependent hidden Markov models (RPD-HMM) proposed by Haoka and Iwahashi [13] and Sugiura et al. [14]. With the visual features of objects and the motion classes, the conceptual structures in the videos are

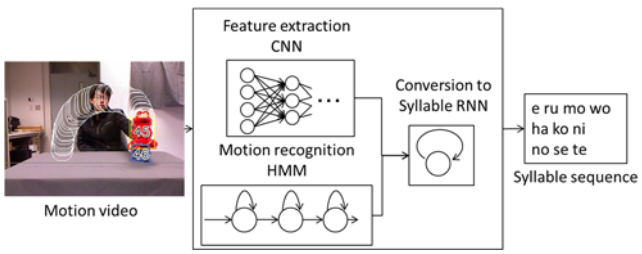

Fig. 1. Overview of the motion to syllable learning process

expressed. Forth, speech describing object manipulation is then converted to syllable sequences using the Japanese speech recognition engine Julius [15]. Finally, a conceptual structure to syllable sequence learning process is conducted using RNN with the parallel data of visual feature vectors (i.e., motion and object image) and syllable sequences.

2.2 Motion recognition

In the motion recognition step, we obtain motion class, trajector (TRJ), and landmark (LND). Motion class shows the action that people do, trajector shows the object that is the target of the action, and landmark shows the object that is the reference point of the action. They are obtained by learning with the L-Core motion recognition program using RPD-HMM.

Fig. 2 shows two motion recognition example results. Fig. 2(1) shows the recognition result for the motion 'move onto', with correctly recognised trajector TRJ:45 and landmark LND:46. Fig. 2(2) shows an example for null landmark situation such as the motion 'lift'. In this case, the motion 'lift' relates only to one object, TRJ:93, and thus the result for landmark is LND:null.

2.3 Conceptual structure to syllable sequence learning with RNN

Our proposed method deals with conversion from conceptual structure, which are expressed with visual feature vectors, to syllable sequence. In conventional image description research, image information is described with word sequences. In contrast, syllable units were used for the description in the proposed method without any information on words.

A conventional encoder-decoder model and its input vector sequence are shown in Figs. 3, 4, and 5, is difficult to use, in this case. Although a baseline encoder-decoder is able to convert conceptual structures expressed with three conceptual attributes (TRJ, LND, and motion), an additional object clustering step is needed to detect TRJ and LND objects. Each vector of the conceptual structure contains a motion part and an object part. A vector representing motions is filled with a respective motion vector value. In general, the output of the motion recognition step has a defined motion ID binary value indicating on and off (i.e., 1 or 0). Similarly, for the TRJ and LND vectors, we made their representative binary values on and off based on recognised outcomes of the object recognition process.

The encoder in the proposed method is shown in Fig. 6. The motion vector part of the proposed method has the same format as the conventional method (see Figs. 5 and 7). In contrast, the object part of the input format to the RNN differs for both methods because the proposed method does not have explicit process for the object recognition. More specifically, the LND and TRJ vectors of the proposed method represent the quantity of the object feature extracted with CNN. The input format is changed to enable training directly from the object image to the syllable sequences.

The complete network architecture of our proposed RNN encoder-decoder can be viewed by combining Fig. 6 (encoder) and Fig. 4 (decoder). In the encoder of the proposed system, the input layer accepts visual feature

(1) (2)

Fig. 2. Example results of motion recognition: (1) motion 'move onto'; (2) motion 'lift'

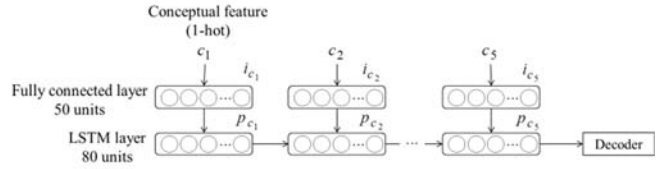

Fig. 3. Baseline encoder

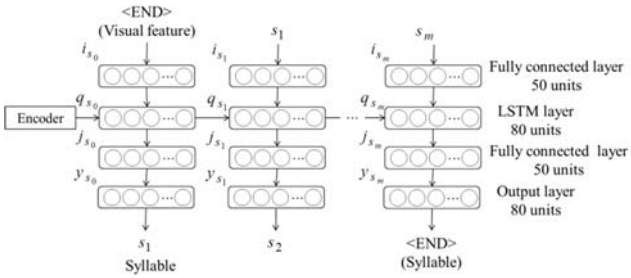

Fig. 4. Decoder utilised in the baseline and proposed system

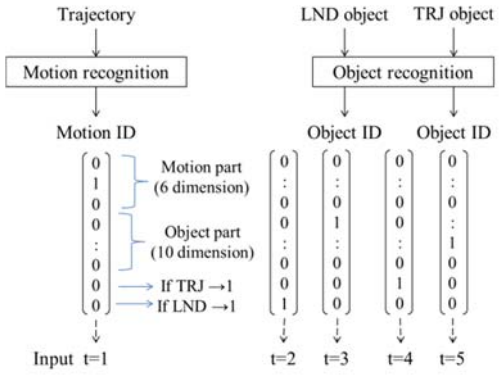

Fig. 5. Conceptual feature vector sequence, which expresses the conceptual structure, of the baseline method

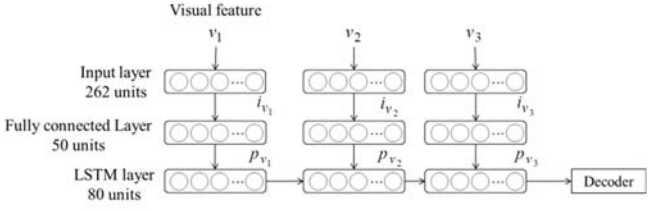

Fig. 6. Encoder utilised in the proposed system

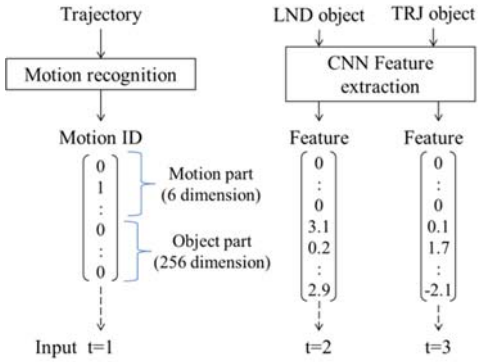

Fig. 7. Visual feature vector sequence, which expresses the conceptual structure, of the proposed method

elements as inputs. The fully connected layer acts as a link and passes on 80 units (Eq. 1). The LSTM [16] layer has two inputs: (1) output of the fully connected layer and (2) output of the LSTM layer in the previous phase (Eq. 2).

$$\mathbf{i}_{v_n} = \tanh(W_{v_n i_n} \cdot v_n) \quad (1)$$

$$\mathbf{p}_{v_n} = \mathrm{LSTM}(W_{i_{v_n} p_{v_n}} \cdot \mathbf{i}_{v_n} + W_{p_{v_n} p_{v_{n-1}}} \cdot \mathbf{p}_{v_{n-1}}) \quad (2)$$

3 EXPERIMENTAL SETUP

3.1 Data preparation

We prepared five hundred parallel motion images, manually constructed a conceptual structure, and manually segmented syllables from Japanese sentences for reference. Ten objects (see Fig. 8) and six defined motions (see Fig. 9) for taking motion videos with Kinect v1 were used. Motion images were taken with a maximum of two objects contained in one image. Four hundred and one hundred parallel data were used for training and test, respectively.

3.2 Preprocessing

A conceptual structure and segmented syllables from Japanese sentences were prepared for conceptual structure to syllable sequence learning with RNN using the following procedure:

1. Motion recognition of all motion images from videos with our in-house L-Core.
2. Image segmentation to obtain the images of objects from the motion images.
3. Extraction of object features from images of objects.
4. Dimension reduction of features from 4096 to 256 with PCA.
5. Building of the conceptual structure expressed with visual feature sequence by combining the motion ID and features of objects (see Fig. 7).
6. The Japanese syllable sequences were prepared using the syllable recogniser Julius.

Fig. 8. The ten objects utilised for object recognition

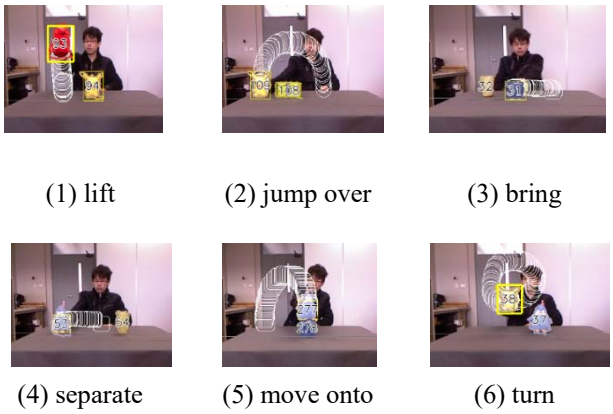

Fig. 9. The six motions utilised for motion recognition

3.3 Framework for RNN training

Chainer [17] is a framework for neural network

development that provides an easy and straightforward way to implement complex deep learning architectures. Some key features of Chainer include support as a Python library (PyPI: Chainer) and capability to run on both CUDA and multi-GPU computers. We used the Chainer Python module (version 1.15.0.1) in our conceptual structure to syllable sequence conversion experiments based on RNN trained for five hundred epochs.

3.4 Evaluation criteria

We used two automatic criteria BLEU and RIBES. Further, we used the score speech recognition system output (SCLITE) program from the NIST scoring toolkit (SCTK) version 2.4.10 [18] to calculate the word error rate (WER). In our case, WER is equal to the syllable error rate for segmented Japanese syllable sequences. The SCLITE scoring method calculates the erroneous words in WER as follows: First, an alignment is made of the G2P hypothesis (the output from the trained model) and the reference (human transcribed) word strings. Then, global minimisation of the Levenshtein distance function is performed. This process weights the cost of correct words via insertions (I), selections (D), and substitutions (S). The formula for WER is as follows:

$$WER = (I + D + S)\ 100/N$$

4 RESULTS

The results for the baseline and the proposed method are given in Tables 1 and 2. In the tables, 'Manual' signifies execution with manually prepared conceptual structures and syllable sequences, and 'Syllable Recognition' signifies execution with conceptual structures built from online motion and object recognition results and the syllable sequences output from Julius.

The results in Tables 1 and 2 clearly show that training with our proposed method is comparable with the baseline in terms of RIBES score, sentence accuracy percentage, and syllable accuracy percentage. However, the BLEU score of the proposed method is 2.3 points below that of the baseline. The sentence and syllable accuracy percentages are also comparable with the baseline results for both manual and syllable recognition (see Tables 1 and 2).

We also conducted recognition error analyses in which 16 images failed to be recognised, resulting in motion recognition of 96.8%. In syllable sequences recognition with Julius, syllable accuracy was 88.6% and sentence accuracy was 15.4%.

Table 1. Results of baseline method in terms of BLEU, RIBES, sentence accuracy, and syllable accuracy

	BLEU	RIBES	Sentence Accuracy (%)	Syllable Accuracy (%)
Manual	98.3	0.996	96	98.7
Syllable Recognition	74.6	0.966	15	86.4

Table 2. Results of proposed method in terms of BLEU, RIBES, sentence accuracy, and syllable accuracy

	BLEU	RIBES	Sentence Accuracy (%)	Syllable Accuracy (%)
Manual	96.0	0.985	91	95.8
Syllable Recognition	74.9	0.961	18	84.8

5 DISCUSSION

As outlined above, the conversion results from conceptual structure to Japanese syllable sequences by our proposed method are, in general, comparable with the baseline. In terms of comparison of the conversion results between manual and syllable recognition, Tables 1 and 2 clearly show that they were affected by the recognition results for motion and speech. Among them, as shown in Section 4, the speech recognition error rate is significantly higher than that for motion. Therefore, if the speech recognition error rate could be reduced, then better performance for conversion from image to syllable sequences could be obtained. Fig. 10 shows actual speech recognition results from Julius with related motion images attached. In the figure, sentences (1) and (2) were correctly recognised, whereas the other two sentences were not. In sentence (3), 'ko pu' should be 'koppu' and in sentence (4), 'tsu' should be 'chi'.

6 CONCLUSION

The main contribution of this paper is a proposed method that uses the features of image objects directly in a conceptual structure to learn syllable sequences for their descriptions. The experimental results obtained show that the proposed method can achieve results that are on par with those of the current baseline method. Moreover, our experiments prove that language acquisition research without language-specific morphological analysis and image object clustering is possible. On the other hand, the training time of our proposed method is considerably longer than that of the baseline. We

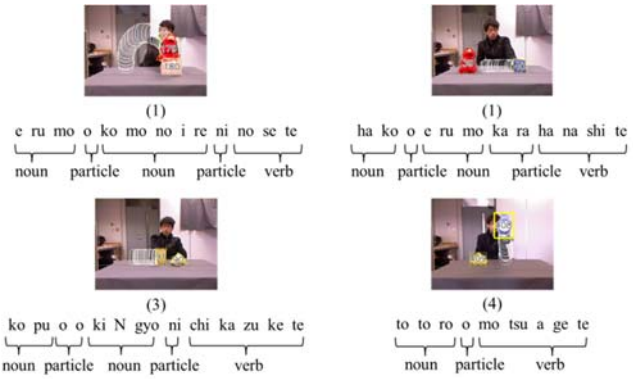

Fig. 10. Example results of speech recognition with Julius

plan to extend our Encoder-Decoder study to attention approaches.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI (grant number 15K00244) and JST CREST ('Symbol Emergence in Robotics for Future Human-Machine Collaboration').

REFERENCES

[1] Iwahashi N (2007), Robots That Learn Language: Developmental Approach to Human-Machine Conversations. Human Robot Interaction I-Tech Education and Publishing, pp. 95–118

[2] Iwahashi N, Sugiura K, Taguchi R, et al. (2010), Robots That Learn to Communicate: A Developmental Approach to Personally and Physically Situated Human-Robot Conversations. In Proceedings of the AAAI Fall Symposium on Dialog with Robots, pp.38–43

[3] Takabuchi K, Iwahashi N, Kunishima T (2015), A Language Acquisition Method Based on Statistical Machine Translation for Application to Robots. The Sixth Joint IEEE International Conference on Developmental Learning and Epigenetic Robotics, Poster Session I.

[4] Cangelosi A and Schlesinger M (2015), Developmental Robotics: From Babies to Robots. The MIT Press.

[5] Taniguchi T, Nagai T, Nakamura T, Iwahashi N, Ogata T, and Asoh H (2016), Symbol emergence in robotics: a survey," Advanced Robotics. 30(11-12), pp.706–728

[6] Sutskever I, Vinyals O, Le QV (2014), Sequence to Sequence Learning with Neural Networks. Advances in Neural Information Processing Systems 27 (NIPS)

[7] Cho K, Merrienboerr Bv, Gulcehre C, et al. (2014), Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.1724–1734

[8] Papineni K, Roukos S, Ward T, et al. (2001), BLEU: A Method for Automatic Evaluation of Machine Translation. IBM Research Report rc22176 (w0109022), Thomas J. Watson Research Center.

[9] Isozaki H, Hirao T, Duh K, et al. (2010), Automatic Evaluation of Translation Quality For Distant Language Pairs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 944–952.

[10] Donahue J, Jia Y, Vinyals O, et al. (2014), DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In Proceedings of the International Conference on Machine Learning (ICML), pp.647–655

[11] Jia Y, Shelhamer E, Donahue J, et al. (2014), Caffe: Convolutional Architecture for Fast Feature Embedding. In Proceedings of the 22nd ACM International Conference on Multimedia, pp. 675–678.

[12] Deng J, Dong W, Socher R, et al. (2009), ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 248-255.

[13] Haoka T, Iwahashi N (2000), Learning of the reference-point-dependent concepts on movement for language acquisition (in Japanese). Technical Report of IEICE PRMU 100(442), pp.39–46

[14] Sugiura K, Iwahashi N, Kashioka H, et al. (2011), Learning, Generation and Recognition of Motions by Reference-Point-Dependent Probabilistic Models. Advanced Robotics 25(6-7): 825–848

[15] Kawahara T, Lee A, Kobayashi T, et al. (2000), Free Software Toolkit for Japanese Large Vocabulary Continuous Speech Recognition. In Proc. Int'l Conf. on Spoken Language Processing (ICSLP), Vol. 4, pp. 476–479

[16] Hochreiter S, Schmidhuber J (1997), Long Short-Term Memory. Neural Computation, 9(8), 1735–1780.

[17] https://github.com/pfnet/chainer

[18] http://www1.icsi.berkeley.edu/Speech/docs/sctk-1.2/sclite.htm