ヒューマンロボットマルチモーダル言語インタラクションのニューラルネットワークモデル

Neural Network Model for Human-Robot Multimodal Linguistic Interaction

守屋綾祐 *　高渕健太　岩橋直人
Ryosuke Moriya　Kenta Takabuchi　Naoto Iwahashi

岡山県立大学
Okayama Prefectual Univercity

Abstract: This paper proposes a novel neural network model for human-robot multimodal linguistic interaction, which learns the associations between human acts, comprising a linguistic expression and a physical action, and the appropriate robot response acts to them. The model uses encoder and decoder recurrent neural netowks (RNNs), convolutional neural network (CNN) image feature extractors, and a multilayer perceptron (MLP). The input and output of the model are a human action and the robot ' s response to it, respectively. Its originalities are: 1) three input modalities (language, image, and motion), 2) two output modalities (language and motion), 3) support for two types of human requests (movement directive and visual question), and 4) learning with no prior knowledge of words, grammar, or object concepts. Experimental evaluations demonstrated that the model shows a promising performance, with the robot's response to a human instruction or question being accurate approximately 74% of the time.

1 INTRODUCTION

For multimodal linguistic interaction between humans and robots, it is necessary for both participants to share the association between language and sensory-motor perceptive information in a given situation. In this regard, Winograd [1] pointed out the difficulties in the realization of structural coupling and description of shared beliefs in systems. Harnad [2] reframed the challenge and addressed it as a symbol grounding problem.

Recent studies that address these problems primarily focus on language acquisition by robots, particularly word and grammar acquisition. A comprehensive overview of such studies in the context of symbol emergence problem has been provided by Taniguchi et al. [3]. Takabuchi et al. [4] proposed a neural network model that learns the association between robots' actions and the syllable sequences that describe them. In their study, the input to the model is a human action that is represented by the human motion and an image of the environment in which this motion is performed, and the output is a sentence (a syllable sequence) describing the input action. Their methodology enables association between actions and sentences without any prior knowledge of words, grammar, and object concepts. Similarly, most language acquisition studies have focused on the association between sensory-motor perceptive information and the corresponding linguistic expressions. However, we believe that learning models should progress beyond such association models to realize multimodal linguistic interaction between human and robots.

Certain studies have explored such multimodal linguistic interaction models. For instance, Iwahashi et al. [5] developed a multimodal linguistic interaction model that enables robots to learn how to understand human acts including speech and motion, and to respond to them in unfamiliar situations. Their model is implemented using statistical graphical models. Taguchi et al. [6] demonstrated a learning model that enables robots to learn how to respond via speech

*連絡先：Okayama Prefectural University Graduate School of Information Science and Engineering Department of Systems Engineering
〒 719-1197 Okayama prefecture Soja city Kuboki 111 Faculty of Information Engineering 2617
E-mail: cd29042r@c.oka-pu.ac.jp

and motion to simple human movement directives and visual questions in specific situations. An example of a movement directive is "Place the big stuffed toy on the box", and examples of visual questions are "What is this?" and "Which of these items is a box?" Their model was the extension of the model by Iwahashi et al. [5]. However, this model has limited ability in word acquisition, and it needs to be given some prior linguistic knowledge manually.

Next, we examine some neural network models from the perspective of multimodal linguistic interaction. The neural network model for visual question-answering (e.g. [7]) is a type of multimodal linguistic interaction model. In this model, the input to the model is a visual question expression (a word sequence) and the image regarding which the question is asked, and the output is an answer expression (a word sequence). Another type of multimodal interaction model is the neural network model proposed by Saha et al. [8]. It enacts the role of a store clerk in Internet shopping. Its input is a human linguistic request (a word sequence) and the images of commercial goods that the human linguistic request is made about, and its output is a linguistic response (a word sequence) and images for additional information, such as images of recommended commercial goods.

However, in order to realize true human-robot multimodal linguistic interaction, we need the model to learn the association between human acts, which include linguistic expressions and motions under specific situations, and the robots' acts, which include linguistic expressions and motions, as responses to them. Thus far, no model has been proposed to address this critical issue.

In this paper, we propose a neural network model for human-robot multimodal linguistic interaction that enables robots to learn the association between human acts and robots' acts as responses to them. The inputs to the model comprise human acts, which are composed of three modalities: 1) a human linguistic expression, 2) a human action, and 3) an image that represents the situation in which the human performs the act. The outputs comprise the robot's response act to the human act, and are composed of two modalities: 1) the robot's linguistic expression and 2) the robot's motion. This neural network model is implemented using recurrent neural networks (RNNs) and convolutional neural networks (CNNs). Its originalities are as follows:

1. **Three input modalities:** A human linguistic expression, human motion, and an image that represents the situation in which the human performs the act comprise the input to the model.
2. **Two output modalities:** A robot's linguistic expression and robot's motion comprise the output of the model.
3. **Two types of human requests:** A robot can respond adaptively to two types of human requests, namely, 1) a movement directive and 2) a visual question about the objects in its view.
4. **Learning with no prior information:** Learning can be carried out from scratch without any prior knowledge of words, grammar, and object concepts. The linguistic expressions are represented by syllable sequences in the input as well as the output.

The rest of this paper is organized as follows. Section 2 describes the proposed neural network model. Section 3 presents the experimental evaluation of the proposed model. Section 4 discusses the experimental results. Section 5 concludes the paper.

2 PROPOSED MODEL

To clarify the differences between previous models and the proposed model, we contrast the inputs and outputs of the models. In Takabuchi's language acquisition study [4], the input is image and motion information, and the output is a syllable sequence, as explained in Section 1 (Fig. 1). In the visual question-answering studies (e.g. [7]), the input is a word sequence and an image, and the output is a word sequence (Fig. 2).

By contrast, in the proposed model, the inputs are a syllable sequence, an image, and motion information, and the outputs are a syllable sequence and motion information (Fig. 3).

2.1 Feature representations of inputs and outputs

Next, the feature representations of the inputs and outputs of the proposed model are described .

The syllable sequences in the input and output are represented by Japanese syllable symbol (1-hot vector) sequences. The input syllable sequence describes

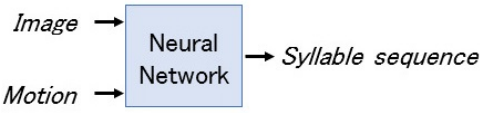


Figure. 1: Takabuchi's language acquisition model

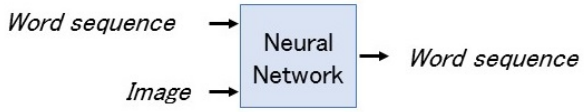


Figure. 2: Visual question-answering model

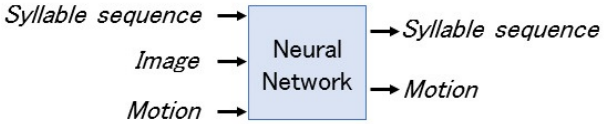


Figure. 3: The proposed human-robot multimodal linguistic interaction model

a human request that is either a movement directive, such as "Raise up the box," represented by the syllable sequence [ha ko o a ge te], or a visual question, such as "What is this?" represented by the syllable sequence [ko re wa na ni]. The output syllable sequence describes a robot answer to the human visual question, such as "It's a lunch-box" represented by syllable sequence [be N to o da yo].

Second, the input image is represented by RGB-D information (continuous values). The images used in this study include two foregrounding objects, as shown in Figure 4.

Third, the input motion (human behavior) information is represented by 1-hot vectors, and the output motion (robot's response behavior) information is represented by a sequence of 1-hot vectors. The input vector comprises information regarding human pointing and gazing behaviors. The pointing and gazing information are encoded as 1-hot vectors. Examples of such behaviors are shown in Figure 5. The motion output vector sequence includes two types of information: 1) information on which of two objects in front of the robot it should grasp, and 2) the category of the action that the robot should perform to execute

Figure. 4: Example input images

(a) Pointing (b) Gazing

(c) Pointing and Gazing

Figure. 5: Example images with human pointing and gazing behaviors

this. An example of the motion output sequence is [OBJ1, AGERU] for the robot motion "Raise up the left object," where OBJ1 and AGERU represent the object that the robot should grasp and the action that the robot should perform to execute it, respectively.

2.2 Network architecture

Figure 6 shows the architecture of the proposed neural network model. This neural network model was implemented using Chainer [14]. The model is composed of encoder and decoder RNNs [9], both of which include long short-time memories (LSTM) [10], an image segmentation (IS) module, CNN image feature extractors (CNN-FEs), and MLP. The encoder RNN has 200 nodes, while the decoder RNN has 400 nodes.

The input syllable sequences are fed into the encoder RNN. In IS, each object image among two foregrounding objects in an RGB-D image is segmented

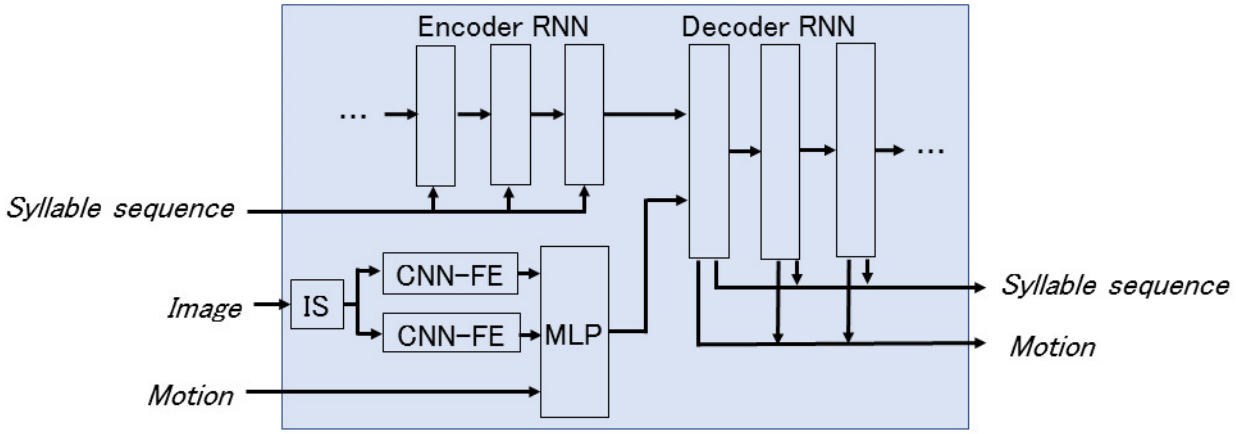


Figure. 6: Architecture of proposed neural network model for human-robot multimodal linguistic interaction

using depth information based on the assumption that the object that the robot is supposed to grasp is located in an area that the robot can reach. These object image segments represented by RGB are fed separately into CNN-FEs implemented with [11, 12, 13]. Each CNN-FE outputs 4096 dimensional features. The motion information with these image features is fed into the MLP. The outputs from the MLP with the output from the encoder RNN are fed into the decoder RNN. Finally, the decoder RNN outputs syllable sequences as well as sequences representing the motion information.

3 EXPERIMENTAL EVALUATION

3.1 Setup

3.1.1 Data preparation

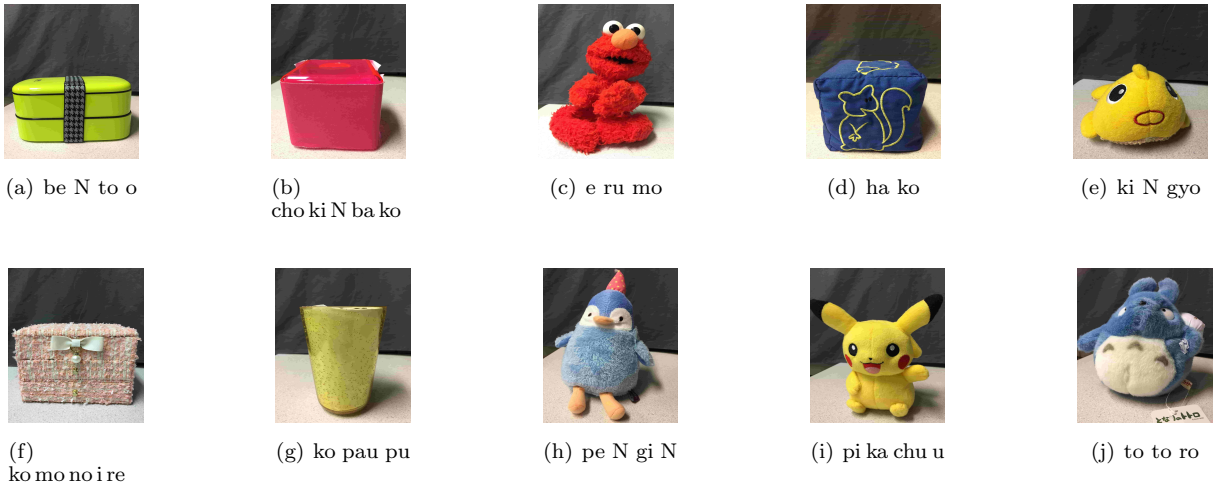
The prepared experimental data set D comprised 840 input-output pair samples, where the inputs represented human acts and the outputs represented robot's response acts as described so far. Each input (a human act) data consists of a syllable sequence (a human linguistic act (request)), an image (situation under which the human request is made), and motion information (a human behavioral act). Each output (a robot's response act) data consists of a syllable sequence (a robot's linguistic act) or motion information (a robot's behavioral act).

Examples of input-output pair samples are shown in Table 1. There were 340 movement directive and 480 visual question samples. The average lengths of the input and output syllable sequences were 4.9 and 3.5 syllables, respectively.

Each of these input-output pairs included an image taken with Kinect V1. Ten objects were used for the data preparation; these are shown in Figure 7. We note that although the number of objects were lim-

Table. 1: Examples of input and output (human act and robot's response act) pair samples

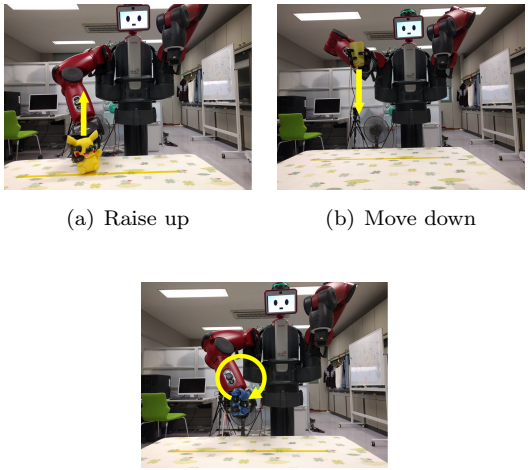
	Input (Human act)	Output (Robot's response act)
Syllable sequence	[ha ko o a ge te] (Raise up the box)	-
Motion	-	Raise up the box
Syllable sequence	[ko re na ni] (What is this?)	[e ru mo da yo] ("It's ERUMO.")
Motion	Pointing to erumo	-
Syllable sequence	[pe n gi N o ma wa shi te] (Move-circle the penguin)	-
Motion	-	Move-circle the penguin
Syllable sequence	[pi ka chu u sa ge te] (Move down the pikachu)	-
Motion	-	Move down the pikachu
Syllable sequence	[a ge te] (Raise up)	-
Motion	Gazing at box	Raise up the box
Syllable sequence	[ma wa shi te] (Move-circle)	-
Motion	Pointing to totoro	Move-circle the totoro
Sylable sequence	[sa ge te] (Move down)	-
Motion	Pointing and gazing at kingyo	Move down the goldfish



(a) be N to o (b) cho ki N ba ko (c) e ru mo (d) ha ko (e) ki N gyo

(f) ko mo no i re (g) ko pau pu (h) pe N gi N (i) pi ka chu u (j) to to ro

Figure. 7: The objects used for data preparation



(a) Raise up (b) Move down

(c) Move-circle

Figure. 8: The categories of robot's response movements in the experimental data

ited, all images were different. The categories of the robot response act were "Raise up", "Move down", and "Move-circle", and these are shown in Figure 8.

3.1.2 Evaluation metric

The experimental data D (840 input-output pair samples) were divided into 700 pair samples for training and 140 pair samples for test. This division was performed six times and six-fold cross validation was performed. Learning the parameters of the neural network model was executed based on a softmax cross entropy criterion. The number of samples whose outputs from the neural network model were completely correct was counted, and their accuracies were calculated.

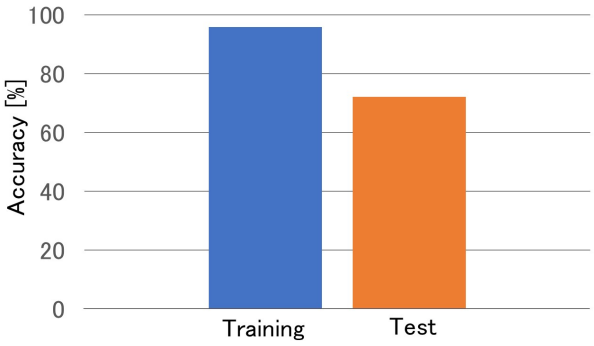


Figure. 9: Training and test data accuracies in whole samples

3.2 Results

First, the overall performance is presented as the output accuracies in Figure 9. The training and test data accuracies, obtained as the average among 6-fold cross-validation experiments, were 98.6% and 73.5%, respectively.

Second, the training and test identification accuracies for two types of human requests were 100% and 100% in movement directive samples, respectively, and 100% and 99.7% in visual question samples, respectively (Fig. 10). From these results, we observe that the proposed model can correctly distinguish between the two types of human requests.

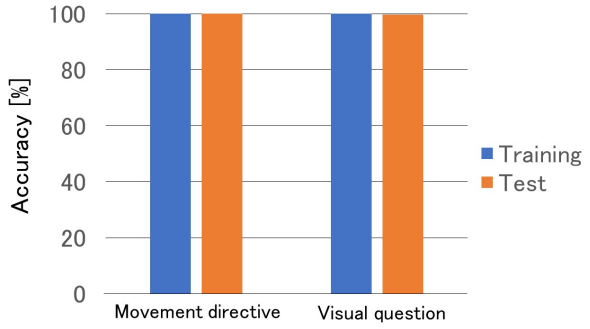


Figure. 10: Training and test data accuracies in movement directive and visual question samples

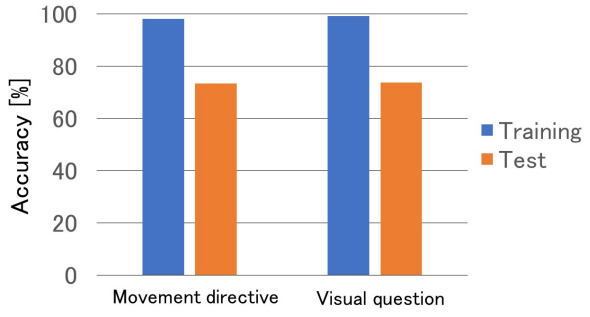


Figure. 11: Training and test data output accuracies in movement directive and visual question samples

Third, the training and test accuracies for two types of human requests were 98.1% and 73.3% in movement directive samples, respectively, and 99.2% and 73.4% in visual question samples, respectively (Fig. 11). From these results, we observe that the proposed model can produce the appropriate robot responses for various types of human requests.

Fourth, Figures 12 and 13 show the effects of two types of input human motion, pointing and gazing. Figure 12 shows the accuracies in movement directive samples with pointing, gazing, both, and neither. Figure 13 shows the accuracies in visual question samples with pointing, gazing, and both. From these results, we can see that the proposed model could use the pointing and gazing information to produce appropriate responses by the robot, and that the objects were selected appropriately with human motion information, human linguistic information, or both.

Finally, the learning curves obtained by averaging the six-fold cross validation results are shown in Figure 14. We see that learning was done normally. However, the difference between the training and test data accuracies was large. This might be owing to the small size of the training data. Increasing training data might reduce the difference in accuracies and improve the learning performance.

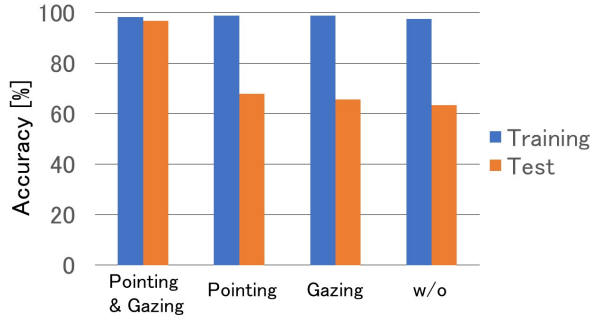


Figure. 12: The effect of pointing and gazing information in training and test data accuracies in movement directive samples

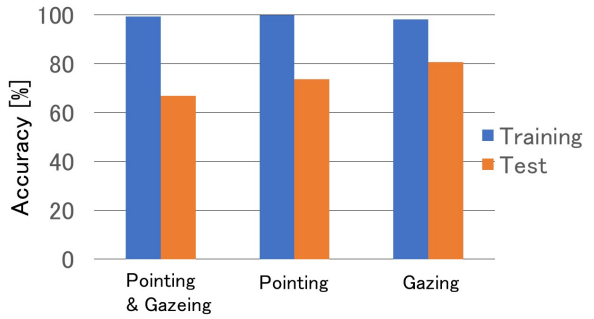


Figure. 13: The effect of pointing and gazing information in training and test data accuracies in visual question samples

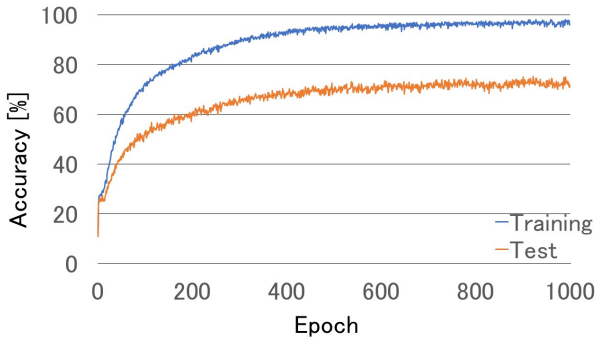


Figure. 14: Learning curves

4 DISCUSSION

From the experimental results, we can confirm that the proposed neural network model successfully achieved the following four original features as mentioned in Section 1.

1. Three input modalities (language, image, and motion)
2. Two output modalities (language and motion)
3. Two types of human requests (movement directive and visual question)
4. Learning with no prior knowledge of words, grammar, and object concepts

However, there are many limitations that should be addressed in future work. These are as follows:

The number of objects: The number of foregrounding objects in each image was fixed to two by considering the simplicity of the network architecture. However, in future, to deal with a variable number of objects, an RNN might be used. Alternatively, the input image could be processed without an object image segmentation process.

Human pointing and gazing behaviors: The information regarding pointing and gazing behaviors was manually provided in the data preparation process in this study. In future, we plan to use raw data of human behavior without any categorization processes.

Speech recognition: Input syllable sequences were manually transcribed in the data preparation process in this study. In future, we plan to use an automatic syllable recognizer. We note that it was reported in Takabuchi's language acquisition study [4] that the automatic syllable recognition process did not introduce a significant difference in the performance of their neural network model.

Further improvements and extensions can also be considered, as follows.

Learning data: Based on the experimental results, it was found that the performance might be affected by the small size of the training dataset. It can be considered that accuracy can be further improved by increasing the number of data. We plan to record the state of interaction between two human beings and introduce it as learning data.

Evaluation: In this study, we experimented with and evaluated neural networks. As we plan to incorporate the proposed neural network model into a robot system, we would like to continue to evaluate the model from the viewpoint of human-robot interaction by actually operating a robot in the future.

5 CONCLUSION

A novel neural network model for human-robot multimodal linguistic interaction was proposed. The input to the model comprised information about a human act, and the output from the model was information on a robot's response act. The experimental evaluations showed that the model exhibited the following capabilities: 1) three input modalities, 2) two output modalities, 3) two types of human requests, and 4) learning with no prior linguistic and object knowledge.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI (Grant number 15K00244) and JST CREST (Grant number JPMJCR15E3, 'Symbol Emergence in Robotics for Future Human-Machine Collaboration').

REFERENCES

[1] Terry Winograd and Fernando Flores: *Understanding Computers and Cognition: A New Foundation for Design*, Addison-Wesley Professional (1987)

[2] Stevan Harnad: The Symbol Grounding Problem, *Pysica D*, Vol. 42, pp. 335-346 (1990)

[3] Tadahiro Taniguchi, Takayuki Nagai, Tomoaki Nakamura, Naoto Iwahashi, Tetsuya Ogata, and Hideki Asoh: Symbol emergence in robotics: a survey, *Adanced Robotics*, Vol. 30, pp. 706–728 (2016)

[4] Kenta Takabuchi, Naoto Iwahashi, Ye Kyaw Thu, and Takeo Kunishima.: DNN-based object manipulation to syllable sequence for language acquisition, *In Proceedings in International Symposium on Artificial Lifes and Robotics*, pp. 638–642 (2017)

[5] Naoto Iwahashi: Robots That Learn Language: Development Approach to Human-Machine Conversation, in Human-Robot Interaction (N. Snaker, Ed.), *I-tech Education and Publishing*, pp. 95–118 (2007)

[6] Ryo Taguchi, Naoto Iwahashi, and Tsuneo Nitta: Learning Communicative Meanings of Utterances by Robots, *New Frontiers in Artificial Intelligence*, LNCS/LNAI 5447, Springer, pp. 62-72 (2009)

[7] Aishwarya Agrawal and Jiasen Lu: VQA: Visual Question Answering, *In Proceedings on International Conference on Computer Vision*, pp. 2425–2433 (2015)

[8] Amrita Saha, Mitesh Khapra and Karthik Sankaranarayanan.: Multimodal Dialogs (MMD): A large-scale dataset for studying multimodal domain-aware conversations, *https://scirate.com/arxiv/1704.00200* (2017)

[9] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio.: Learning Phrase Representations using RNN EncoderDecoder for Statistical Machine Translation, *In Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp. 1724–1745 (2014)

[10] Sepp Hochreiter and Jrgen Schmidhuber: Long Short-Term Memory, *Neural Computation* vol. 9(8), pp. 1735–780 (1997)

[11] Yann LeCun, Patrick Haffner, Lon Bottou, and Yoshua Bengio: Gradient-based learning applied to document recognition., *Proceedings of the IEEE*, pp. 2278–2324 (1998)

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell: Caffe: Convolutional Architecture for Fast Feature Embedding, *In Proceedings of ACM International Conference on Multimedia*, pp. 675–678 (2014)

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei: ImageNet: A Large-Scale Hierarchical Image Database, *In Proceedings of International Conference on Computer Vision and Pattern Recognition* pp. 2–9 (2009)

[14] Chainer: A Powerful, Flexible, and Intuitive Framework for Neural Network, *http://chainer.org/*