Multiple Categorization by iCub: Learning Relationships between Multiple Modalities and Words*

Akira Taniguchi[1], Tadahiro Taniguchi[1] and Angelo Cangelosi[2]

***Abstract*— Human infants can acquire word meanings by estimating the relationships among multiple situations and words. In this paper, we propose a Bayesian probabilistic model that can learn multiple categorizations and words related to any of four modalities (action, object, position, and color). This paper focuses on a cross-situational learning using the co-occurrence of sentences and situations. We conducted a learning experiment using the humanoid iCub robot. In this experiment, the human tutor describes a sentence about an object of visual attention to the robot and an action of the robot. The experimental results show that the proposed method was able to estimate the multiple categorizations and to accurately learn the relationships between multiple modalities and words.**

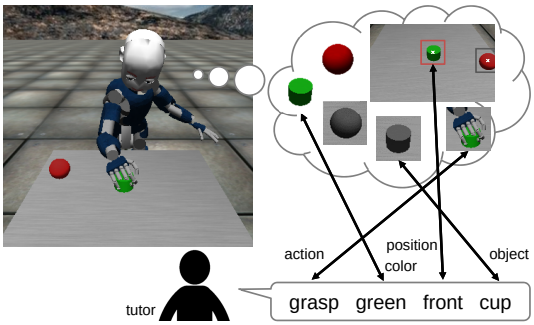


Fig. 1. Overview of the proposed method

I. INTRODUCTION

Human infants can acquire word meanings by estimating the relationships between multiple situations and words. For example, if an infant grasps a green cup at hand, the parent may describe the actions of the infant to the infant using a sentence such as "grasp green front cup". In this case, the infant does not know the relationship between words and situations because the infant has not acquired the word meanings. In other words, the infant cannot determine whether the word "green" indicates an action, an object, or a color. However, we consider that the infant can learn that the word "green" represents the green color by observing the co-occurrence of the word "green" with objects of green color in various situations, e.g., "touch green box" and "green right ball". This is called cross-situational learning, which has been both studied in children [1] and modeled in simulated agents and robots [2].

In this paper, the goal is to develop a novel method for learning the relationships between words and the four modalities (action, object, color, and position) from the robot's experience of hearing sentences describing object manipulation scenes. Many studies of language acquisition focus on the estimation of phonemes and words from speech signals [3]–[5]. In this paper, we assume that the robot can recognize spoken words without errors, as this work focuses specifically on 1) the categorization for each modality, 2) the learning of the relationships between words and modalities, and 3) the grounding of words in multiple categories. Peniak et al. [6] performed action learning by a multiple time-scale recurrent neural network. Attamimi et al. [7] performed an estimation of the relationships among words and multiple concepts by weighting the learned words according to their mutual information as post processing. In this paper, we perform cross-situational learning, including action learning, by a Bayesian probabilistic model. The proposed method can estimate multiple categories and the relationships of words and modalities simultaneously.

*This work was partially supported by JST, CREST.

[1]Akira Taniguchi and Tadahiro Taniguchi are with Ritsumeikan University, 1-1-1 Noji Higashi, Kusatsu, Shiga 525-8577, Japan {a.taniguchi, taniguchi} @em.ci.ritsumei.ac.jp

[2]Angelo Cangelosi is with the Centre for Robotics and Neural Systems, Plymouth University, Plymouth, PL4 8AA, United Kingdom A.Cangelosi@plymouth.ac.uk

II. OVERVIEW OF THE TASK

Fig. 1 shows an overview of the task. The training procedure consists of the following steps:

1) The robot is in front of the table with objects on it.
2) The robot performs an action on an object.
3) The human tutor utters a sentence about the object and the action of the robot.
4) The robot processes the sentence to discover the meanings of the words.

This process (steps 1–4) is carried out many times in different situations. The robot learns word meanings and multiple categories by using visual, tactile, and proprioceptive information, as well as words.

III. MULTIPLE CATEGORIZATIONS AND WORD MEANING LEARNING

Fig. 2 shows a graphical model of the proposed method. The n-th word in the d-th trial is denoted as w_{dn}. An index of the word distribution is denoted as l. The model associates the word distributions θ with categories z_{dm} on four modalities, namely, the action a_d, the position $p_{dm}$of the object on the table, object feature o_{dm}, and object color c_{dm}. An index of the object of attention selected by iCub from the multiple objects on the table is denoted as $A_d = m$. The sequence representing the respective modalities associated

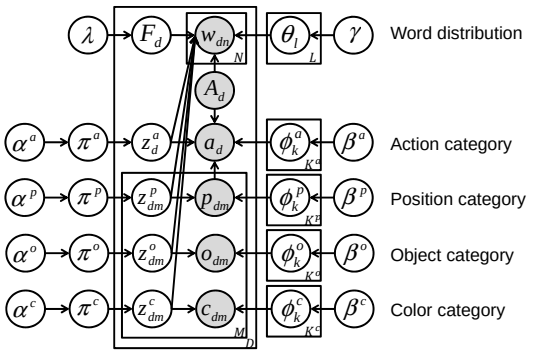


Fig. 2. The proposed graphical model; the action, position, color, and object categories are represented by a component in Gaussian mixture models (GMMs). A word distribution is related to a category on GMMs.

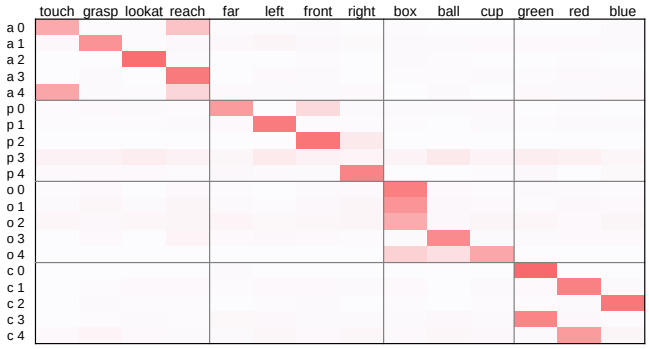


Fig. 3. Word probability distribution across the multiple categories

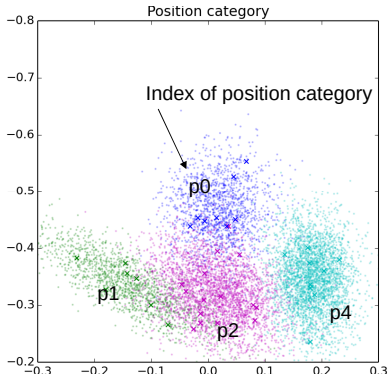


Fig. 4. Learning result of position category; for example, the index of position category p1 corresponds to the word "left". A point group of each color represents each Gaussian distribution of position category. The crosses of each color represent the object positions of the learning data. The each color represents the each position category.

with each word in the sentence is denoted as F_d, e.g., $F_d = (\mathrm{a}, \mathrm{p}, \mathrm{c}, \mathrm{o})$. The number of categories for each modality is K. The number of words in the sentence is N. The set of word distributions is denoted as $\theta = \{\theta_{l=(F_{dn}, z_{dm}^{F_{dn}})} \mid F_{dn} \in \{\mathrm{o}, \mathrm{c}, \mathrm{p}, \mathrm{a}\}, z_{dm}^{F_{dn}} \in \{1, 2, ..., K^{F_{dn}}\}\}$. The action category ϕ_k^a, the position category ϕ_k^p, the object category ϕ_k^o, and the color category ϕ_k^c are represented by a Gaussian distribution. The hyperparameter λ of the Uniform distribution has the constraint that each modality exists only once in the sentence. The hyperparameter of the mixture weights π is denoted as α. The hyperparameter of the Gaussian-inverse-Wishart distribution is denoted as β. The hyperparameter of the Dirichlet distribution is denoted as γ. Parameters of the proposed model are estimated by Gibbs sampling. The learning algorithm is obtained by repeatedly sampling from the conditional posterior distributions for each parameter.

IV. Experiment

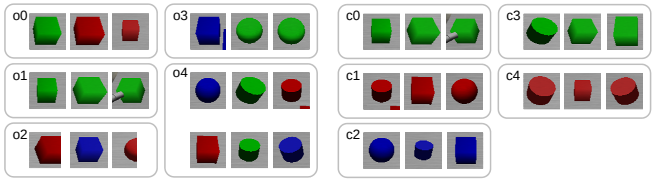
We conducted an experiment on the learning of categories for each modality and words associated with each category. The number of action trials was $D = 20$ for learning. The number of objects M on the table for each trial was a number from one to three. The number of words in the sentence was $N = 4$. We assume that a word related to each modality is spoken only once in each sentence. The word order for each category was $F_d = (\mathrm{a}, \mathrm{p}, \mathrm{c}, \mathrm{o})$ in all of the sentences. This experiment used 14 words. The number of categories for each modality was $K = 5$.

Fig. 3 shows the word probability distributions θ. Higher probability values are represented by darker shades. Fig. 4 shows the learning result for the position category ϕ^p. Fig. 5 shows part of the examples of categorization results for the object and color categories. F_d was correctly estimated in the all of the data. The results demonstrate that the proposed method was able to accurately associate each word with its respective modality.

V. Conclusion

In this paper, we have proposed a learning method that can estimate multiple categories and the relationships between words and multiple modalities. The experimental results showed that it is possible for a robot to learn the combination of an element in a complex situation and a word from their co-occurrence. In the experiment, there were four words within an uttered sentence, however, we consider that the proposed method can learn word meanings from more uncertain sentence, i.e., the sentence could have four words or less with a changing order. The proposed method did not take into account a grammatical information and the sentence of five words or more, which we will try to address in our future studies.

In ongoing work, our focus will be on new experiments with action generation tasks and action description tasks. In addition, we plan to consider integrating the proposed method into Bayesian probabilistic models of object con-



(a) Object (b) Color

Fig. 5. Part of the examples of categorization results

cepts [8] and spatial concepts [9].

REFERENCES

[1] K. Smith, A. D. Smith, and R. A. Blythe, "Cross-situational learning: An experimental study of word-learning mechanisms," *Cognitive Science*, vol. 35, no. 3, pp. 480–498, 2011.

[2] J. F. Fontanari, V. Tikhanoff, A. Cangelosi, R. Ilin, and L. I. Perlovsky, "Cross-situational learning of object–word mapping using neural modeling fields," *Neural Networks*, vol. 22, no. 5, pp. 579–585, 2009.

[3] D. Roy and A. Pentland, "Learning words from sights and sounds: A computational model," *Cognitive Science*, vol. 26, no. 1, pp. 113–146, 2002.

[4] S. Goldwater, T. L. Griffiths, and M. Johnson, "A Bayesian framework for word segmentation: Exploring the effects of context," *Cognition*, vol. 112, no. 1, pp. 21–54, 2009.

[5] T. Taniguchi, R. Nakashima, H. Liu, and S. Nagasaka, "Double articulation analyzer with deep sparse autoencoder for unsupervised word discovery from speech signals," *Advanced Robotics*, vol. 30, no. 11-12, pp. 770–783, 2016.

[6] M. Peniak, D. Marocco, J. Tani, Y. Yamashita, K. Fischer, and A. Cangelosi, "Multiple time scales recurrent neural network for complex action acquisition," in *ICDL/Epirob Joint Conference*, 2011.

[7] M. Attamimi, Y. Ando, T. Nakamura, T. Nagai, D. Mochihashi, I. Kobayashi, and H. Asoh, "Learning word meanings and grammar for verbalization of daily life activities using multilayered multimodal latent Dirichlet allocation and Bayesian hidden Markov models," *Advanced Robotics*, vol. 30, no. 11-12, pp. 806–824, 2016.

[8] T. Nakamura, T. Araki, T. Nagai, and N. Iwahashi, "Grounding of word meanings in latent Dirichlet allocation-based multimodal concepts," *Advanced Robotics*, vol. 25, no. 17, pp. 2189–2206, 2011.

[9] A. Taniguchi, T. Taniguchi, and T. Inamura, "Spatial concept acquisition for a mobile robot that integrates self-localization and unsupervised word discovery from spoken sentences," *IEEE Transactions on Cognitive and Developmental Systems*, 2016.