Available online at www.sciencedirect.com

ScienceDirect

IFAC-PapersOnLine 49-19 (2016) 150–153

Learning of Relative Spatial Concepts from ambiguous instructions

Zhixiang Gu*, Ryo Taguchi*, Koosuke Hattori, Masahiro Hoguro**, Taizo Umezaki***

** Nagoya Institute of Technology, Japan*
(taguchi.ryo@nitech.ac.jp).
***Chubu University, Japan*

Abstract: Nowadays, the research that robots learn the meaning of a word by relating speech instructions with signals of the sensorimotor system has started. Lots of pre-research on learning 'concept' that means the meaning of a word by robot have used the co-occurrence information of words and sensor signals directly. However, relative concepts, such as color, size or location, require comparing with other objects or concepts. The objects or concepts used for comparison are called "reference object." In human-computer interaction, reference point may not be explicitly communicated, because human utterance is ambiguous. Therefore, estimating a true reference object from the multi-candidate is necessary to learn relative concepts. In this paper, we propose a learning method based on EM algorithm to estimate the reference point and learn relative concepts. In our experiments, a robot learned concepts expressing relative location, such as 'in front of' or 'behind.' The robot was given text labels expressing its locations from a human teacher in several situations. Nevertheless, the instructions didn't include the true reference points and the coordinate system of each concept. Under these conditions, the robot could learn the concepts based on coordinate systems have the orientations of the reference objects or the teacher.

Keywords: symbol-grounding, relative concept, spatial concept

1. INTRODUCTION

Spoken dialog systems need to have linguistic knowledge. In particular, communication robots, household robots and nursing-care robots have to understand the various linguistic representations that indicate objects and events in the real world in order to work cooperatively with users. However, it is difficult for the developers to describe all of the knowledge in advance, because such systems can be used in situations other than that the developers assume. Therefore, it is desirable that systems automatically learn such knowledge through interactions with users.

Nowadays, the research that robots learn the meaning of a word by relating speech instructions with signals of the sensorimotor system has started. Roy and Pentland (2002) have conducted experiments to extract semantically useful phoneme sequences from natural utterances on the basis of relations between visual features of objects and acoustic features of utterances. Iwahashi (2006) has proposed a belief system, which is represented by a dynamic graphical model, to learn grammar and pragmatic capability in addition to word meanings. Nakamura et al. (2014) have proposed a method by which a robot can mutually learn a language model and object concepts based on multimodal latent Dirichlet allocation and the nested Pitman-Yor language model.

In this paper, the meaning of a word is called "concept". A lot of pre-research including the above have directly used co-occurrence information of words and sensor signals. However, concepts of color, size, or location are relative, and cannot be learned from just co-occurrence information. Fig. 1 shows an example of learning the concept of "big". We assume that this

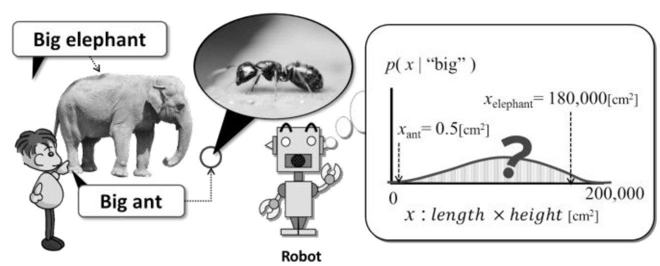

Fig. 1. An example of learning a relative concept.

robot can get the length and height of an object by using a 3D measurement system. The feature indicated by "big" is size of an object. The robot approximates the feature by $length \times height$. A user said "This is big" seeing an elephant which is 600 cm in length and 300 cm in height. After that, the user said "This is big" seeing an ant elephant which is 1 cm in length and 0.5 cm in height. The robot knew the meaning of "This is xxx." The robot tried to estimate the concept of "big" by using those features but could not acquire the suitable distribution. In practice, the utterances mean "This elephant (ant) is bigger than a general elephant (ant)." Therefore, in order to correctly learn relative concepts, the robot has to estimate the object by comparison according to each concept and each scene, and extract relative features between the target object and the estimated object. An object to be compared is called a reference object (Landau and Jackendoff, 1993).

In this paper, we treat relative spatial concept as an example of relative concepts, and propose a method by which a robot learns relative spatial concepts from ambiguous instructions. The proposed method is based on EM algorithm. It estimates both the reference object used in each instruction and the

10.1016/j.ifacol.2016.10.477

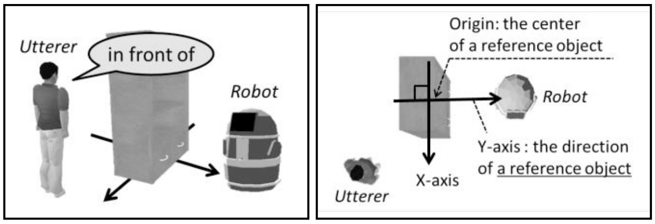

Fig. 2. Reference frame based on the orientation of a reference object.

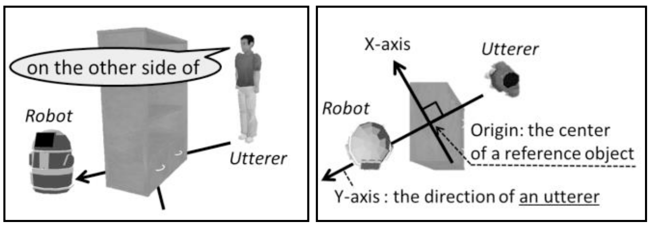

Fig. 3. Reference frame based on the orientation of an utterer.

parameters of statistical distributions representing each concept at the same time. Iwahashi (2006) has proposed a method based on EM algorithm for learning the concepts of motions depending on a reference object. However relative spatial concepts have not been discussed.

2. RELATIVE SPATIAL CONCEPTS

Spatial concepts are dependent on coordinate systems. Coordinate system for spatial concepts is called reference frame (Logan and Sadler, 1996). Reference frame is used to indicate the location of a target object. In this paper, we treat two types of reference frame as shown in Fig. 2 and 3.

First reference frame is based on the orientation of a reference object. The origin is the center of a reference object. The direction of Y-axis accords with the orientation of the reference object. The right direction perpendicular to the Y-axis is the positive X-axis. By using this frame, the utterer can represent the location of the robot by "The robot is *in front of* the cabinet."

Second reference frame is based on the orientation of an utterer. The origin is the center of a reference object same as the first reference frame. The Y-axis is the line drown to the origin from the utterer. The right direction perpendicular to the Y-axis is the positive X-axis. By using this frame, the utterer can represent the location of the robot by "The robot is on the other side of the cabinet."

In the case that a robot learns concepts through natural interaction with a user, each reference object may not be clearly indicated. In addition, the robot may not be able to understand the name of the reference object because it is in the middle of learning. Also the reference frame of each concept is not indicated. Therefore, the learning of spatial concepts needs to estimate not only the reference object, but also the reference frame.

3. LEARNING METHOD FOR RELATIVE SPATIAL CONCEPT

A human teaches a word representing the current location of a robot. By repeating the teaching while changing the location, the robot acquires the distribution of the locations that the word indicates. This distribution is a spatial concept. The relative location of the robot on a reference frame is represented as two-dimensional coordinate $\mathbf{x}_{n,c,k} = (x_{n,c,k}, y_{x,c,k})$, while n is the index of training data, c is a reference frame and k is a reference object. Although the candidates of reference objects and the candidates of reference frames are given, the true reference object and the true coordinate system, which are used for teaching, are not given.

A spatial concept is built by a normal distribution with mean $\boldsymbol{\mu}_c$ and variance $\boldsymbol{\sigma}_c{}^2$. The probability of that reference object k is the true reference is $\pi_{n,c,k}$. The number of candidates of reference object for scene n is represented as M_n, the probability of reference is $\boldsymbol{\pi}_{n,c} = (\pi_{n,c,1}, \cdots, \pi_{n,c,k}, \cdots, \pi_{n,c,M_n})$, the probability of reference object of all is $\boldsymbol{\pi}_c = (\boldsymbol{\pi}_{1,c}, \cdots, \boldsymbol{\pi}_{n,c}, \cdots, \boldsymbol{\pi}_{N,c})$. The parameters of model are represented as $\theta_c = (\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c{}^2, \boldsymbol{\pi}_c)$ .The likelihood of model $L(\theta_c)$ is shown as function (1). EM algorithm calculates the maximum parameter $\widehat{\theta_c}$ through iteration process (2).

$$L(\theta_c) = \sum_{n=1}^{N} \ln \sum_{k=1}^{M_n} \pi_{n,c,k} \, \mathrm{N}(x_{n,c,k}; \mu_{c,x}, \sigma_{c,x}^2) \, \mathrm{N}(y_{n,c,k}; \mu_{c,y}, \sigma_{c,y}^2) \tag{1}$$

$$\widehat{\theta_c} = \arg\max_{\theta_c} L(\theta_c) \tag{2}$$

The initial value of $\pi_{n,c,k}$ depends on $d_{n,c,k}$ denoting the Euclidean distance between the robot and each candidate of reference object as shown in functions (3) and (4). The shorter distance, the bigger initial value of $\pi_{n,c,k}$ will be distributed.

$$d_{n,c,k} = \sqrt{{x_{n,c,k}}^2 + {y_{n,c,k}}^2} \tag{3}$$

$$\pi_{n,c,k} = \frac{1/(d_{n,c,k})^{\alpha}}{\sum_k 1/(d_{n,c,k})^{\alpha}} \tag{4}$$

where α is a parameter to adjust the effect of the distance. We conducted a preliminary experiment to compare the effect of α on experimental results. In the experiment, we set the value of α to 0.0, 0.5, 1.0, 1.5 and 2.0, respectively. The results showed that the effect is best when α is 1.5. Therefore, in the later-mentioned experiment, we set the value of α to 1.5. The automatic adjustment of α is a future work.

Model parameters on each reference frame are calculated using the EM algorithm. After that, the maximum likelihood model is selected as the true reference frame $\hat{c}$.

$$\hat{c} = \arg\max_{c} L(\widehat{\theta_c}) \tag{5}$$

4. EXPERIMENT

4.1 The conditions of experiment

Experiments to learn relative spatial concepts were conducted on SIGVerse (Tan and Inamura, 2010), which is the simulator that combines dynamics, perception, and communication simulations for synthetic approaches to research into the genesis of social intelligence. Fig.4 shows the experimental environment. A user operates an avatar of human type by a controller. A robot follows the avatar. The user can teach the location name in the text at any point. Text is stored together with the locations of the robot and the avatar. There are six objects, which are a bookshelf, a television, a sofa, a plant,