Implementation of Biases Observed in Children's Language Development into Agents

Ryo Taguchi, Masashi Kimura, Shuji Shinohara, Kouichi Katsurada, and Tsuneo Nitta

Graduate School of Engineering, Toyohashi University of Technology
1-1 Hibariga-oka, Tempaku-cho, Toyohashi-city, 441-8580 Japan
{taguchi, kimura, shinohara, katurada, nitta}
@vox.tutkie.tut.ac.jp
http://www.vox.tutkie.tut.ac.jp/

Abstract. This paper describes efficient word meaning acquisition for infant agents (IAs) based on learning biases that are observed in children's language development. An IA acquires word meanings through learning the relations among visual features of objects and acoustic features of human speech. In this task, the IA has to find out which visual features are indicated by the speech. Previous works introduced stochastic approaches to do this, however, such approaches need many examples to achieve high accuracy. In this paper, firstly, we propose a word meaning acquisition method for the IA based on an Online-EM algorithm without learning biases. Then, we implement two types of biases into it to accelerate the word meaning acquisition. Experimental results show that the proposed method with biases can efficiently acquire word meanings.

1 Introduction

The demand for language-mediated natural communication with PDAs, navigation systems, and robots is increasing in line with the development and spread of IT technologies. In the study of communication, an important problem has been how to handle the meanings of symbols such as words and gestures and transfer them without misunderstanding. In classical AI such as the Semantic Net and Physical Symbol System, the meaning of each symbol is defined by another symbol, so some external systems are needed to connect the meaning with real objects in such schemes. This is called the "Symbol Grounding Problem" [1, 2]. One of the solutions to this problem is to give a computer, or an agent, the capability to acquire symbols representing the relations among visual features of objects and acoustic features of human speech through interactions with the real world. Moreover, in language-mediated natural communication between a human and an agent, the two parties need to share the symbols held by each other in order to correctly understand what the other party wants to say.

Recently, studies on word meaning acquisition, in which a human teaches words to an agent through human–agent interaction, have begun. Akaho et al. [3], Roy et al. [4] and Iwahashi et al. [5] respectively proposed mechanisms to acquire the word meanings that represent relations among visual features of objects and acoustic features of

P. Vogt et al. (Eds.): EELC 2006, LNAI 4211, pp. 89 – 99, 2006.

human speech using a machine learning method. By applying these mechanisms, agents can learn and understand word meanings in the real world.

Such studies are divided into two types: (1) an object has a name [4], and (2) an object has some features which have corresponding words and the words are taught [3, 5]. For example, suppose that a human shows an agent a picture of a rabbit: in type (1) the human speaks "rabbit", while in type (2) the human speaks "rabbit" together with "white" or "big". Type (2) is a more difficult task than type (1) because the agent has to find out which visual features are represented by a word. In studies [3] and [5], the features are identified by using stochastic methods, however, these methods need a lot of examples. To overcome this problem, we propose a word meaning acquisition mechanism with two types of learning biases, the mutual exclusivity bias [8] and the shape bias [9], which are observed in children's language development. When the agent with learning biases watches an object and listens to an unknown word at the same time, the agent can guess its word meaning based on the meanings of other known words. Therefore, the biases are expected to make the word meaning acquisition more efficient than a stochastic only approach.

In section 2, we propose a basic word meaning acquisition mechanism using an Online-EM algorithm without biases. In section 3, we discuss formulations and implementation of the biases. In section 4, we conduct experiments to test the effectiveness of the bias. Lastly, in section 5, we describe the conclusions of this paper.

2 Word Meaning Acquisition Mechanism

2.1 Infant Agent

A human infant learns language mainly based on the triadic relationship among him/herself, his/her parent, and an object. This relationship is also important in natural communication between a human and an agent because the agent can directly sense the object's features, share them with the human, and acquire word meanings on the basis of them. For this reason, we have developed Infant Agents (IAs) that are modeled after the language acquisition process of human infants. In the learning process of an IA, a human, who is a teacher, shows an object to the IA and speaks a word that represents certain features of the object. The IA perceives both human speech and the object's features through its audio-visual sensors, and acquires the relationship between visual information and auditory information. The IA regards this relationship as a meaning of the word.

2.2 IA's Sensory Information

(1) Visual Information (see Fig. 1)
When a human shows an object, an IA receives it as a bitmap image and extracts the visual features from it. These features are divided into three types of attributes (shape, hue and lightness) by the difference of extraction procedures. Hue and lightness features are obtained by converting RGB signals of the image into HSV colors that contain hue, saturation and value (lightness). The shape feature is

obtained by the following process. First, three monochrome images with different resolutions (100%, 50%, 25%) are generated from the original image. Then, contour exaction is applied to each image. Lastly, 25-dimensional Higher-order Local Auto-Correlation (HLAC) features are calculated for each image [6], and the total of 75 (25×3) dimensional features is used as the shape feature.

(2) Auditory Information

In this paper, words from keyboard input are used as auditory information to avoid experimental complexity caused by recognition errors.

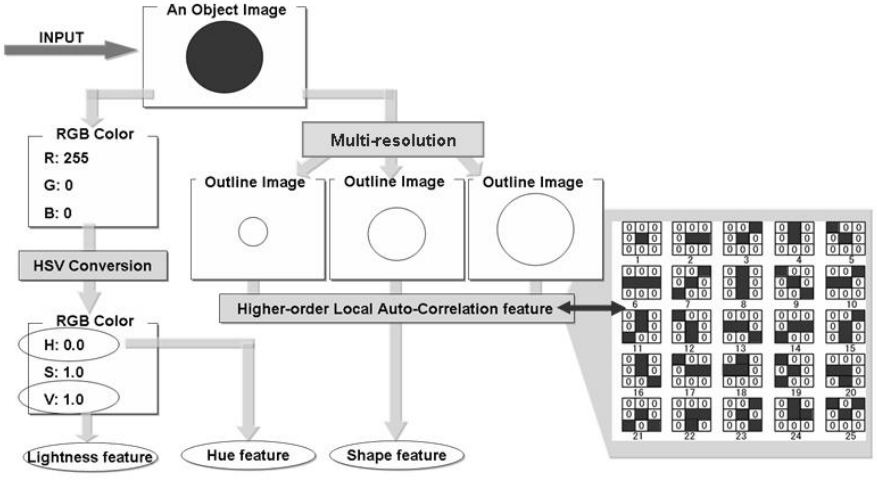

Fig. 1. Visual Information of an IA

2.3 Word Meaning Acquisition

In our approach, a human teaches a word related to the IA's sensory information such as "circle", "red", "dark", etc. For example, the word "circle" represents a set of specific shapes or a specific range of shape features. Note that it does not represent other attributes such as hue and lightness. However, at this point, the IA is not taught which features of objects are represented by each word. Therefore, the IA has not only to learn the range, but also to identify the target attributes represented by each word.

Learning the range is easier than identifying the target attributes because an IA can calculate the range by counting co-occurrence frequency between a word and visual features. In this paper, we express this co-occurrence frequency by probability distributions of the frequency with respect to each attribute. We apply the Online-EM algorithm [7] for calculating probability distributions in which probability distributions are generated, modified, and sometimes deleted through the E-steps and M-steps of the EM algorithm.

Identifying the target attributes needs more complex calculation. If an attribute is a target of a word, its probability distribution will be different from the ones obtained from other words because the distribution is calculated from the specific objects that are to be distinguished from others by the attributes. For example, when an IA learns the word "circle", the probability distribution of the shape attribute will be learned from shape features of only circle objects. On the other hand, the probability distribution of the hue attribute will not show any specific difference from the one obtained from "square" or "dark", because they are not the targets of each word and will be similarly learned from various hue features. Therefore, in this paper, we use the difference between the probability distribution of an attribute and that obtained from all objects (we call this distribution the Basis Distribution) to identify the target attribute (see Fig. 2). The Basis Distribution is calculated by the Online-EM algorithm before the word meaning acquisition.

Here, we formalize our word meaning acquisition mechanism. When a human shows an object to an IA and gives a word w, the IA extracts visual features $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2, ..., \mathbf{x}^i, ..., \mathbf{x}^I)$ from the object. The index i represents one of I attributes (I=3) and each $\mathbf{x}^i$ is a J(i)-dimensional vector ($\mathbf{x}^i = (x_1^i, x_2^i, ..., x_j^i, ...x_{J(i)}^i)$). In this paper, $\mathbf{x}^1$ is a shape feature vector with seventy five dimensions, $\mathbf{x}^2$ is a hue feature vector with one dimension, and $\mathbf{x}^3$ is a lightness feature vector with one dimension. Then the IA calculates probability distribution $P(\mathbf{x}^i|w)$ for each attribute i and a word w by using the Online-EM algorithm. The confidence measure $Conf(i,w)$ ($0 \le Conf(i,w) \le 1$) that indicates whether a word w targets attribute i or not, is calculated. $Conf(i,w)$ is given by using the correlation $Corr(i,w)$ ($0 \le Corr(i,w) \le 1$) between Basis Distributions $P(\mathbf{x}^i)$ and $P(\mathbf{x}^i|w)$. The correlation $Corr(i,w)$ and the confidence measure $Conf(i,w)$ are calculated as follows.

$$Corr(i, w) = \left(\frac{1}{J(i)} \right) \sum_{j=1}^{J(i)} \int P(x_j^i) P(x_j^i | w) dx_j^i \qquad i = 1, 2, 3 \tag{1}$$

$$Conf(i, w) = 1 - Corr(i, w) \qquad i = 1, 2, 3 \tag{2}$$

If $Conf(i,w)$ is less than a threshold Th_i, the attribute i is determined as a non-target attribute of a word w. When an object is shown to an IA, the occurrence probability of a word w $P(w \mid \mathbf{X})$ is calculated by the next equation.

$$P(w \mid \mathbf{X}) = P(w) \prod_{i \in \arg[Conf(i,w) > Th_i]} \frac{P(\mathbf{x}^i \mid w)}{P(\mathbf{x}^i)} \tag{3}$$

where $P(w)$ is probability of a word w.

If a word w has a higher value of $P(w \mid \mathbf{X})$ than those of the other words that have the same set of target attributes as w, the IA considers the word w as a word representing the features of the object.

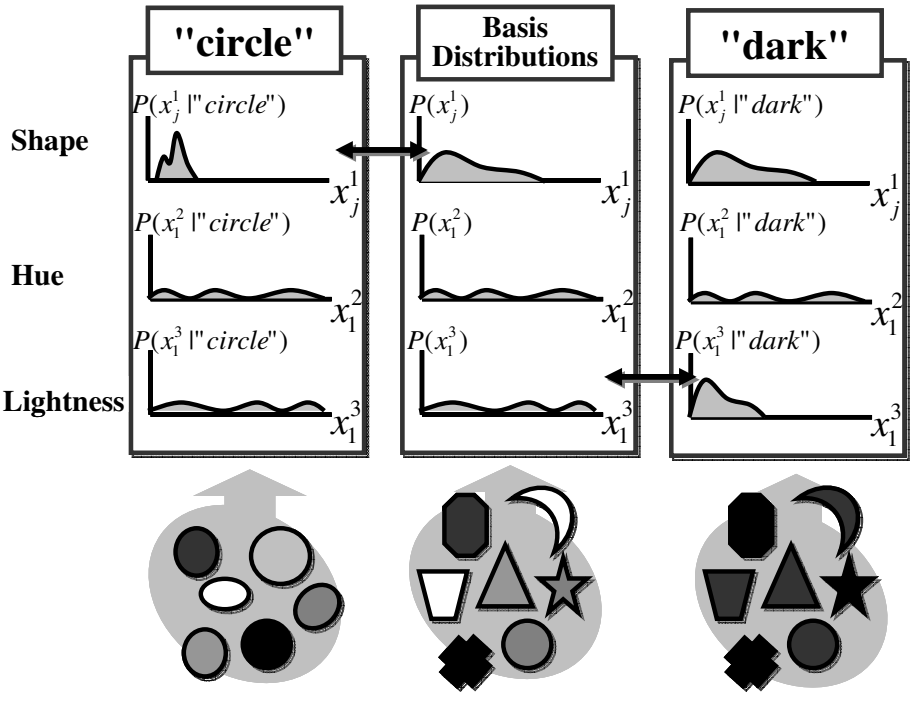

Fig. 2. Word meaning acquisition using stochastic methods

3 Implementation of Learning Bias

3.1 Learning Bias Observed in Children's Language Development

The philosopher Quine pointed out the following problem. If an unknown word is given to a child in some context, it is very difficult to determine the referent indicated by the word because there are huge amounts of candidate hypotheses to be a referent. However, when a child hears a word for the first time, he/she does not test these hypotheses completely but understands the referent of the word quickly with few errors. This is called fast mapping. Currently, many psychologists consider that fast mapping is executed based on some learning biases which children have inherently [8,9,10,11]. These biases may also be important for the agent that acquires word meanings in the real world. In this paper, we incorporate two biases, the mutual exclusivity bias and the shape bias, into an IA to acquire word meanings efficiently.

3.2 Formulations and Implementation

In our framework, the above hypotheses correspond to combinations of attributes. To test these hypotheses (combinations of attributes), our IAs use the confidence measure *Conf*(*i*,*w*). Therefore, biases should be some parameters that inhibit *Conf*(*i*,*w*). In this

paper, we introduce $B(i,w,t)$ ($0 \le B(i,w,t) \le 1$), which includes both the mutual exclusivity bias and the shape bias, to inhibit $Conf(i,w)$.

$$Conf(i,w) = B(i,w,t) \cdot \left(1.0 - Corr(i,w)\right) \quad (4)$$
$$where \quad B(i,w,0) = E(i,w) \cdot S(i,w)$$

In the above expression, $E(i,w)$ is the mutual exclusivity bias and $S(i,w)$ is the shape bias. $Conf(i,w)$ becomes a lower value than the value given by equation (2) if $B(i,w,t)$ is a low value. In this case, the attribute i will be identified as a non-target from equation (3). The following sections describe the details of $E(i,w)$, $S(i,w)$ and attenuation of $B(i,w,t)$.

(1) Mutual exclusivity bias

When a human infant learns a new word about an object, he/she is known to use a rule that the word meaning is not congruent with other word meanings [8]. That is, if an infant hears an unknown word, he/she seeks its meaning outside the meanings of known words. This rule is called the mutual exclusivity bias. We formulate this bias as follows.

If an IA has already known some words **W'** ($w' \in W'$) related an object and hears an unknown word w when looking at the object, the IA determines the target attribute of w not to become the target attributes of **W'**. However, target attributes of a known word w' are not always correct if w' has not been learned enough. Therefore this bias should be controlled according to the number of times of learning w'. The mutual exclusivity bias $E(i,w)$ is calculated by using the following equation.

$$E(i,w) = \begin{cases} 1.0 & (\Theta_i \leqq Th_i) \\ 0.5 & (\Theta_i > Th_i) \end{cases} \quad (5)$$
$$where \quad \Theta_i = \max_{w' \in W} \left[\frac{Conf(i,w)}{1.0 + \exp\{-\alpha(t^{w'} - \beta)\}} \right]$$

Here, α and β are the parameters of the sigmoid function, and $t^{w'}$ represents the number of times learning w'.

(2) Shape bias

Human infants use another rule. When they hear a new word about an object, they tend to interpret the word as indicating the shape of the object. This is called the shape bias [9]. This bias can be formulated by inhibiting each $Conf(i,w)$ for non-shape attributes of the word as shown in equation (6). However, it is not used if the shape attribute has already been inhibited by the mutual exclusivity bias.

$$S(i,w) = \begin{cases} 1.0 & (i = 1 or\ E(i,w) < 1.0) \\ 0.5 & (i \neq 1) \end{cases} \quad (6)$$

(3) Attenuation of bias

The above two biases may decrease the efficiency of word meaning acquisition depending on the order of teaching words. For example, if a human teaches a word representing color (hue or lightness) of an object first, the IA will assume that the word

represents the shape of the object due to the shape bias. To avoid this problem, we attenuate $B_i^w(t)$ according to the number of learning t as follows.

$$B(i,w,t) = B(i,w,t-1) + \gamma\left[1.0 - B(i,w,t-1)\right] \tag{7}$$

where γ is an attenuation rate ($0 < \gamma < 1$).

4 Experiments

4.1 Experimental Setup

In the experiments, we prepared 1,080,000 objects with different features (Fig. 3 shows a part of them). Each object has one of 108 shape features, one of 100 gradations of hue features, and one of 100 levels of lightness features. We assume that each attribute is represented by 7 words, and so an IA is taught a total of 21 words such as "circle", "square", "red", and so on. Note that each word targets only an attribute and does not have duplicated meanings. We taught the words according to the following three types of teaching sequence.

Table 1. Parameters used in the experiments

Th_1	0.2
Th_2	0.4
Th_3	0.4
α	0.5
β	20
γ	0.1

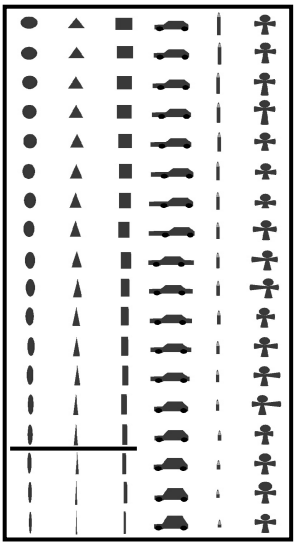

Fig. 3. Objects used in the experiments

TS1: Taught words are chosen randomly.
TS2: Seven shape words are taught first. After that, seven hue words are taught and then seven lightness words.
TS3: Seven hue words are taught first. After that, seven lightness words are taught and then seven shape words.

We evaluate the word meanings acquired by the IA each time we teach one word. In the evaluation, we show the IA 200 objects chosen randomly, and the IA speaks the words that represent the features of those objects. When the spoken words correctly represent the features of the object and their target attributes are correct, we consider that the IA has acquired correct word meanings. Table 1 shows parameters used in word meaning acquisition.

4.2 Evaluation of Word Meaning Acquisition Mechanism Without the Biases

We calculate correct rates and confusion rates to evaluate our basic word meaning acquisition mechanism without the biases. The confusion rate was calculated from the frequency that an IA had correctly identified the target attribute of a word but the IA used the word as having a different meaning. Figure 4 shows the correct rates and confusion rates of the word meanings that were acquired by the IA after 2,000 iterations of teaching according to the teaching sequence TS1.

The average of the correct rates was more than 90%, showing that the IA correctly acquired word meanings. However, the confusion rates of shape words were higher than those of hue and lightness concepts. Figure 5 shows the difficulty of correctly acquiring shape words (the horizontal axis represents the number of teaching while the vertical axis represents the correct rate). Shape words are more complex than hue and lightness because they are represented by 75-dimensional features. Moreover, the feature ranges of shape words are also narrower than others, and some parts of them overlap, causing confusion. To resolve this problem, we are now considering reducing the number of dimensions by using principal component analysis.

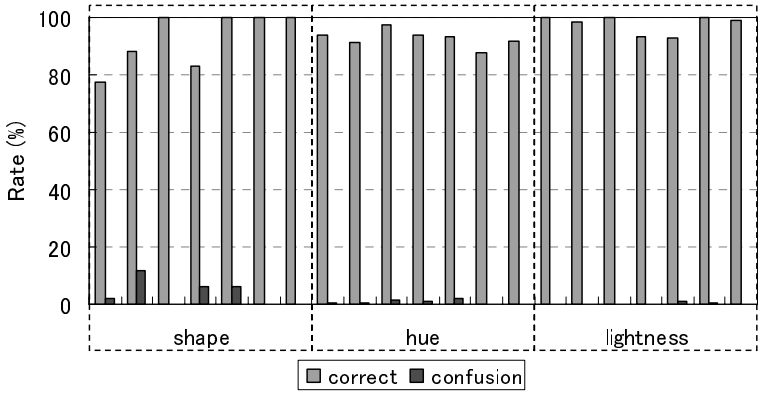

Fig. 4. Correct rate and confusion rate of each acquired word

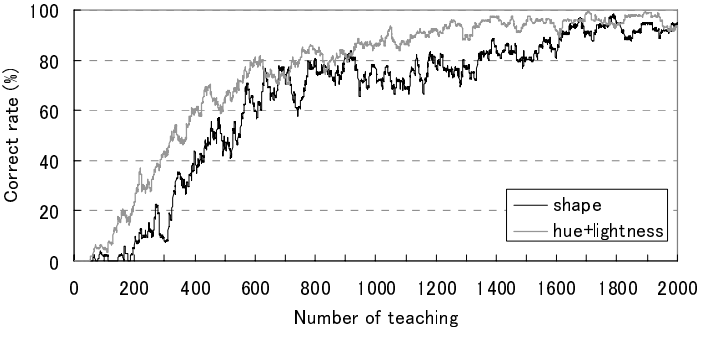

Fig. 5. Correct word rate of shape words and other words in the learning stage

4.3 Evaluation of the Biases

Figures 6 to 8 show the results of comparing between the presence and absence of the biases in the above condition TS1 to TS3. The horizontal axis represents the number of teaching while the vertical axis represents the correct word rate (%). These graphs show that the IA with the biases is able to learn word meanings more efficiently than without biases. When shape words are taught first (TS2), improvement of the correct word rate is quicker than the other conditions, because the shape bias is applied to the initial shape words and the mutual exclusivity bias is applied to subsequent words (see Fig. 7). On the other hand, when words are chosen randomly (TS1) or shape words are taught at the end (TS3), the shape bias is incorrectly applied to the hue and lightness words. However, in actuality, this was not found to have adverse influences and acquisition of shape words became faster because the IA was able to correctly determine target attributes by attenuating incorrect biases according to the number of learning. However, the most efficient word meaning acquisition is achieved by teaching shape words first.

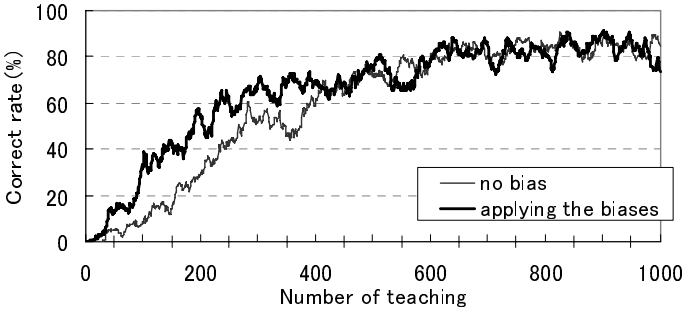

Fig. 6. Effectiveness of the biases in TS1

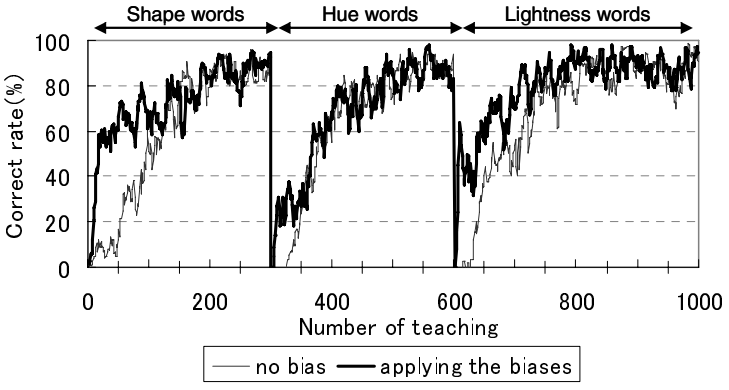

Fig. 7. Effectiveness of the biases in TS2

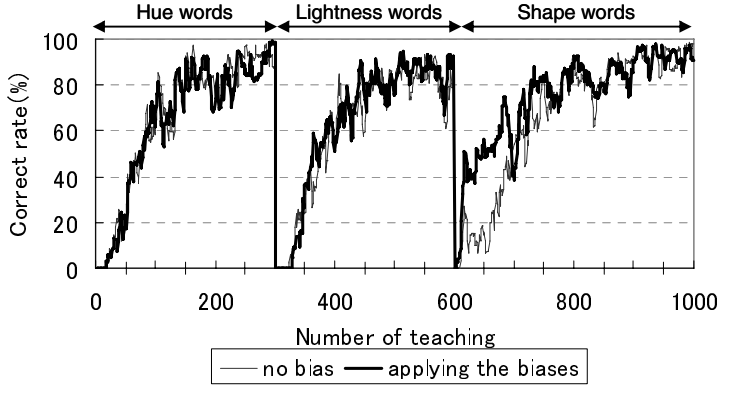

Fig. 8. Effectiveness of the biases in TS3

5 Conclusion

This paper described the efficient acquisition of word meanings based on learning biases. In order to acquire word meanings, an agent has to learn the feature range represented by each word and to identify the target attributes indicated by it. Our stochastic method for word meaning acquisition learns the feature ranges as probability distributions by using an Online-EM algorithm, and the target attributes are identified by comparing the correlation between these probability distributions and the Basis Distributions. The experimental results showed that the agent applying our stochastic method could acquire word meanings correctly. However, this method needs many examples.

In order to resolve this problem, we formulated two biases which are observed in children's language development, and implemented them into the agent. Although the effects of these biases depend on the teaching sequence, the results of comparative

experiments showed that this has few adverse influences and the most efficient learning is achieved by teaching shape words first.

In a future work, we will implement the principle of contrast [10]; this is a widely-known bias which is expected to make the acquisition of hierarchical meanings more efficient.

Acknowledgments

This research was supported by a 21st Century COE Program Grant for "Intelligent Human Sensing".

References

[1] S. Harnad, The symbol grounding problem, Physica D 42, 1990, 335-346.
[2] R. Pfeifer and C. Scheier, Understanding Intelligence. The MIT Press, Cambridge, MA, 1999.
[3] S. Akaho et al., Concept acquisition from multiple information sources by the EM algorithm, IEICE Trans, J80-A(9), 1997, 1546-1553.
[4] D. Roy, Integration of speech and vision using mutual information, in Proc. ICASSP, Vol. 4, 2000, 2369-2372.
[5] N. Iwahashi, Language acquisition through a human-robot interface by combining speech, visual, and behavioral information, Information Sciences, Vol. 156, 2003, 109-121.
[6] N. Otsu and T. Kurita, A new scheme for practical, flexible and intelligent vision systems, Proc. IAPR Workshop on Computer Vision, 1988, 431-435.
[7] M. Sato and S. Ishii, On-line EM algorithm for the normalized Gaussian network, Neural Computation, 12(2), 2000, 407-432.
[8] E.M. Markman, Categorization and naming in children, MIT Press, 1989.
[9] B. Landau et al., The importance of shape in early lexical learning, Cognitive Development, 3, 1988, 299-321.
[10] E. Clark, The principle of contrast: A constraint on language acquisition. In B. MacWhinney (Ed.), Mechanisms of language acquisition, Hillsdale, NJ: Lawrence Erlbaum Assoc, 1987, 1-33.
[11] M. Imai and E. Haryu, The nature of word learning biases: From a cross linguistic perspective, In D.G. Hall & S. Waxman, (Eds.), Weaving a lexicon. MIT Press, 2004, 411-444.