A Connectionist Approach to Learn Association between Sentences and Behavioral Patterns of a Robot

Yuuya Sugita
BSI, RIKEN
Hirosawa 2-1, Wako-shi
Saitama 3510198 JAPAN
sugita@bdc.brain.riken.go.jp

Jun Tani
BSI, RIKEN
Hirosawa 2-1, Wako-shi
Saitama 3510198 JAPAN
tani@bdc.brain.riken.go.jp

Abstract

We present a novel connectionist model for acquiring the semantics of a simple language through the behavioral experiences of a real robot. We focus on the "compositionality" of semantics, a fundamental characteristic of human language, which is the ability to understand the meaning of a sentence as a combination of the meanings of words. We also pay much attention to the "embodiment" of a robot, which means that the robot should acquire semantics which matches its body, or sensory-motor system. The essential claim is that an embodied compositional semantic representation can be self-organized from generalized correspondences between sentences and behavioral patterns. This claim is examined and confirmed through simple experiments in which a robot generates corresponding behaviors from unlearned sentences by analogy with the correspondences between learned sentences and behaviors.

1. Introduction

Implementing language acquisition systems is one of the most difficult problems, since not only the complexity of the syntactical structure, but also the diversity in the domain of meaning make this problem complicated and intractable. In particular, how linguistic meaning can be represented in the system is crucial. This problem has been investigated for many years.

In this paper, we introduce a connectionist model that acquires the semantics of a simple finite language with respect to correspondences between sentences and the behavioral patterns of a real robot. An essential question is how compositional semantics can be acquired in the proposed connectionist model without providing any representations of the meaning of a word or behavior routines *a priori*. By "compositionality", we refer to the fundamental human ability to understand a sentence from (1) the meanings of its constituents, and (2) the way in which they are put together. It is possible for a language acquisition system that acquires compositional semantics to derive the meaning of an unknown sentence from the meanings of known sentences. Consider the unknown sentence: "John likes birds." It could be understood by learning these three sentences: "John likes cats."; "Mary likes birds."; and "Mary likes cats." That is to say, generalization of meaning can be achieved through compositional semantics.

From the point of view of compositionality, the symbolic representation of word meaning is advantageous for processing the linguistic meaning of sentences. In general, AI-based models employ semantic symbols to represent word meanings and have much affinity with compositionality in terms of the meanings of sentences (e.g., (Thompson and Mooney, 1998)). Thus, this approach assumes that the meanings of words (i.e., lexicon) is independent of the usages of words in sentences (i.e., syntax).

According to this observation, various learning models have been proposed to acquire the embodied semantics of language. For example, some models learn semantics in the form of correspondences between sentences and non-linguistic objects, i.e., visual images (Roy, 2002), sequences of video images (Siskind, 2001), or the sensory-motor patterns of a robot (Iwahashi, 2003, Steels, 2000). In these models, the meanings of words are labeled by the words themselves (a.k.a., semantic symbols). Consequently, a significant part of semantic learning can be reduced to the learning of the syntactic structure of the sentence. This means that the acquisition of meanings of sentences can be translated into two relatively separate steps, the acquisition of word meanings and the acquisition of syntax.

Although this separated learning approach seems to be plausible from the requirements of compositionality, it causes inevitable difficulties in representing the grounded meaning of a sentence. A priori separation of lexicon and syntax requires a pre-defined manner of combining word meanings to compose the meaning of a sentence. As a result, these approaches require relatively heavy manual pre-programming to realize compositional semantic representations. Not all of these models aim for self-organization of grounded compositional semantic representations. We nevertheless point out possible problems with their approaches within the framework of acquiring grounded compositional semantics.

In Iwahashi's model, the class of a word is assumed to be given prior to learning its meaning because different acquisition algorithms are required for nouns and verbs (c.f., (Siskind, 2001)). Roy's model does not require a priori knowledge of word classes, but requires the strong assumption, that the meaning of a word can be assigned to some pre-defined attributes of non-linguistic objects. This assumption is not realistic in more complex cases, such as when the meaning of a word needs to be extracted from non-linguistic spatio-temporal patterns, as in case of learning verbs.

Recently, several connectionist models for acquiring embodied language have been proposed (Billard, 2002, Sugita and Tani, 2002). These models don't require separate treatment of words and syntax. However, they can demonstrate only few compositional characteristics. Also, it cannot be said that embodied semantics is self-organized fully from scratch, since the models assume behavior primitives *a priori.*

In this paper, we discuss an essential mechanism for self-organizing embodied compositional semantic representations. Our model implements compositional semantics by utilizing the generalization capability of a recurrent neural network (RNN), where the meaning of each word cannot exist independently, but emerges from the relations with others (c.f., reverse compositionality, (Fodor, 1999)). In this situation, a sort of generalization can be expected, such that the meanings of novel sentences can be inferred by analogy with learned ones.

The experiments were conducted using a real mobile robot with an arm and with various sensors, including a vision system. A finite set of two-word sentences consisting of a verb followed by a noun was considered. We assume that a sentence is represented as a sequence of pre-defined words, however, employ neither semantic symbols nor composition rules *a priori.* Our analysis will clarify what sorts of internal neural structures should be self-organized for achieving compositional semantics grounded to a robot's behavioral experiences. Although our experimental design is limited, the current study will suggest an essential mechanism for acquiring grounded compositional semantics, with the minimal combinatorial structure of this finite language (Evans, 1981).

2. Task Design

The aim of our experimental task is to understand an essential mechanism for self-organizing compositional semantics grounded to the behavior of a robot. In the training phase, our neural network model learns the relationships between sentences and the corresponding behavioral sensory-motor sequences of a robot in a supervised manner. It is then tested to generate behavioral sequences from a given sentence. We regard compositional semantics as being acquired if appropriate behavioral sequences can be generated from unlearned sentences by analogy with learned data.

Our mobile robot has three actuators, with two wheels and a joint on the arm; a colored vision sensor; and two torque sensors, on the wheel and the arm (Figure 1a). The robot operates in an environment where three colored objects (red, blue, and green) are placed on the floor (Figure 1b). The positions of these objects can be varied so long as the robot sees the red object (R) on the left side of its field of view, the blue object in the middle (B), and the green object (G) on the right at the start of every trial of behavioral sequences.

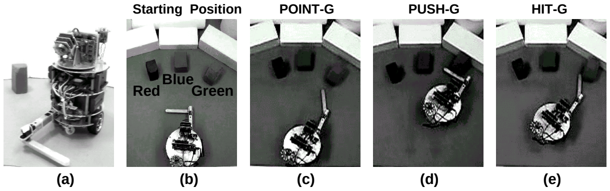

Figure 1: The mobile robot (a) starts from a fixed position in the environment and (b) ends each behavior by pointing at (c), pushing (d), or hitting (e) either the red, blue, or green object.

The robot learns nine categories of behavioral patterns, consisting of pointing at, pushing, and hitting each of the three objects, in a supervised manner. These categories are denoted as POINT-R, POINT-B, POINT-G, PUSH-R, PUSH-B, PUSH-G, HIT-R, HIT-B, and HIT-G (Figure 1c-e). It should be noted that the robot learns these behaviors as sensory-motor time sequences, in which there are no obvious relationships among behavioral categories. For example, one can not generate a sensory-motor sequence belonging to POINT-R as a combination of POINT-B and PUSH-R.

The robot also learns sentences which consist of one of three verbs (`point`, `push`, `hit`) followed by one of six nouns (`red`, `left`, `blue`, `center`, `green`, `right`). The meanings of these 18 possible sentences are given in terms of fixed correspondences with the 9 behavioral categories (Figure 2). For example, "`point red`" and "`point left`" correspond to POINT-R, "`point blue`" and "`point center`" to POINT-B, and so on.

In these correspondences, "`left`," "`center`," and "`right`" have exactly the same meaning as "`red`," "`blue`," and "`green`" respectively. These synonyms are introduced to observe how the behavioral similarity affects the acquired linguistic semantic structure.

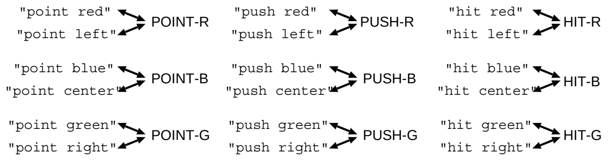

Figure 2: The correspondences between sentences and behavioral categories. For each behavioral category, there are two corresponding sentences.

3. Proposed Model

Our model employs two RNNs with parametric bias nodes (RNNPBs) (Tani, 2003, Tani and Ito, 2003, Ito and Tani, 2003) in order to implement a linguistic module and a behavioral module (Figure 3). The RNNPB, like the conventional Jordan-type RNN (Jordan and Rumelhart, 1992), is a connectionist model for learning time sequences. The linguistic module learns the above sentences represented as time sequences of words (Elman, 1990), while the behavioral module learns the behavioral sensory-motor time sequences of the robot. To acquire the correspondences between the sentences and behavioral sequences, these two modules are connected to each other by using the parametric bias binding method. Before discussing this binding method in detail, we introduce the overall architecture of RNNPB and both modules.

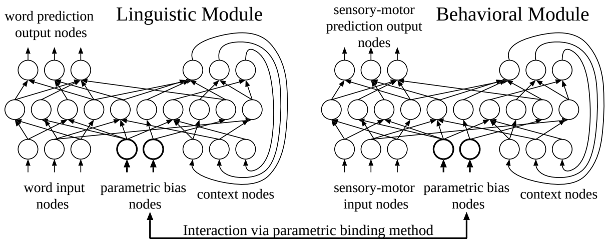
word prediction output nodes
Linguistic Module
sensory-motor prediction output nodes
Behavioral Module
word input nodes
parametric bias nodes
context nodes
sensory-motor input nodes
parametric bias nodes
context nodes
Interaction via parametric binding method

Figure 3: Our model is composed of two RNNs with parametric bias nodes (RNNPBs), one for a linguistic module and the other for a behavioral module. Both modules interact with each other during the learning process via the parametric bias method introduced in the text.

3.1 RNNPB

The RNNPB has the same neural architecture as the Jordan-type RNN except for the PB nodes in the input layer (c.f., each module of Figure 3). Unlike the other input nodes, these PB nodes take a specific constant vector throughout each time sequence, and are employed to implement a mapping between fixed-length vectors and time sequences.

Like the conventional Jordan-type RNN, the RNNPB learns time sequences in a supervised manner. The difference is that in the RNNPB, the vectors that encode the time sequences are self-organized in PB nodes during the learning process. The common structural properties of all the training time sequences are acquired as connection weight values by using the back-propagation through time (BPTT) algorithm, as used also in the conventional RNN (Jordan and Rumelhart, 1992, Rumelhart et al., 1986). Meanwhile, the specific properties of each individual time sequence are simultaneously encoded as PB vectors (c.f., (Miikkulainen, 1993)). As a result, the RNNPB self-organizes a mapping between the PB vectors and the time sequences.

The learning algorithm for the PB vectors is a variant of the BPTT algorithm. For each time sequence, the back-propagated errors with respect to the PB nodes are accumulated for all time steps to update the PB vectors. Formally, the update rule for the PB vector p_{x_i} encoding the i-th time sequence $\boldsymbol{x}_i$ is given as follows:

$$\begin{aligned} \delta^2 p_{x_i} &= \frac{1}{l_i}\sum_{t=0}^{l_i-1} error_{p_{x_i}}(t) && (1)\\ \delta p_{x_i} &= \epsilon \cdot \delta^2 p_{x_i} + \eta \cdot \delta p_{x_i}^{old} && (2)\\ p_{x_i} &= p_{x_i}^{old} + \delta p_{x_i}. && (3) \end{aligned}$$

In equation (1), the update of PB vector $\delta^2 p_{x_i}$ is obtained from the average back-propagated error with respect to a PB node $error_{p_{x_i}}(t)$ through all time steps from $t = 0$ to $l_i - 1$, where l_i is the length of $\boldsymbol{x}_i$. In equation (2), this update is low-pass filtered to inhibit frequent rapid changes in the PB vectors.

Here, we introduce an abstracted operational notation for the RNNPB to facilitate a later explanation of our proposed method of binding language and behavior. After successfully learning the time sequences $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_N$, the RNNPB can generate a time sequence $\boldsymbol{x}_i$ from its corresponding parametric bias p_{x_i}. By applying an operator $RNNPB$, the generation of $\boldsymbol{x}_i$ is described as follows:

$$RNNPB(p_{x_i}) \quad \rightarrow \quad \boldsymbol{x}_i, \quad i = 1, \cdots, N. \qquad (4)$$

We note here that the actual generation process of a time sequence is implemented by iteratively utilizing the RNNPB with input vectors for each time step. Both the environmental sensory information and the internal prediction of the RNNPB are employed as input vectors depending on the required functionality of the module.

Furthermore, the RNNPB can be used not only for sequence generation processes but also for recognition processes. For a given sequence $\boldsymbol{x}_i$, the corresponding PB vector p_{x_i} can be obtained by using the update rules for the PB vectors (equations (1) to (3)), without updating the connection weight values. This inverse operation for generation is regarded as recognition, and is hence denoted as follows:

$$RNNPB^{-1}(\boldsymbol{x}_i) \quad \rightarrow \quad p_{x_i}, \quad i = 1, \cdots, N. \qquad (5)$$

The other important characteristic nature of the RNNPB is that the relational structure among the training time sequences can be acquired in the PB space through the learning process. This generalization capability of RNNPB can be employed to generate and recognize unseen time sequences without any additional learning. For instance, by learning several cyclic time sequences of different frequency, novel time sequences of intermediate frequency can be generated (Ito and Tani, 2003).

3.2 Linguistic Module

The linguistic module learns the sentences shown in section 2to acquire the underlying syntactic structure. The sentences are represented as a time sequence of words, which starts with a fixed starting symbol. Each word is locally represented, such that, each input node of the module corresponds to a specific word and only one input node takes 1.0 while the others take 0.0. The module has 10 input nodes for each of 9 words (`point`, `push`, `hit`, `red`, `left`, `blue`, `center`, `green`, and `right`) and a starting symbol. The module also has 6 parametric bias nodes, 4 context nodes, 50 hidden nodes, and 10 prediction output nodes. This representation is almost the same as in Elman's previous work (Elman, 1990). However, the acquired dynamics shows a different characteristic nature.

Elman's model predicts a probabilistic distribution of the next words. Therefore, the prediction output can be a probabilistic superposition of word vectors. In contrast, our linguistic module can deterministically predict the next word by utilizing a given PB vector. Recall that a PB vector encodes a specific sentence by means of equation (4). Thus, RNNPB acquires properties of specific sentences in addition to common syntactic properties of sentences. We'll see the role of this characteristic nature in treating the meaning of a sentence in a later section.

3.3 Behavioral Module

The behavioral module learns the time sequences of sensory-motor vectors involving the robot's behaviors presented in section 2.After successfully learning them, the module generates motor commands and a prediction of the sensory image at the next time step from the current sensory-motor vector (Tani, 1996).

The robot can generate behavior by iteratively utilizing the module with changing sensory inputs to generate motor commands every one-third of a second. A behavioral sequence is thus created by sampling three sensory-motor vectors per second during a trial of the robot's behavior. Typical behavioral sequences are about 5 to 25 seconds long, and therefore have about 15 to 75 sensory-motor vectors.

A sensory-motor vector is a real-numbered 26-dimensional vector consisting of 3 motor values (for angular velocities of 2 wheels and an angle of the arm joint), 2 values from torque sensors (of the wheels and the arm), and 21 values encoding the visual image. The visual field is divided vertically into 7 regions, and each region is represented by (1) the fraction of the region covered by the object, (2) dominant hue of the object in the region and (3) the bottom border of the object in the region, which is proportional to the distance of the object from the camera[1]. We note here that the visual information is not the most important for the acquisition of semantics. It occupies 21 of 26 dimensions in the sensory-motor vector only due to the characteristic nature of visual information. All the sensory-motor information is complementary and interdependent.

The module has 26 input nodes for the sensory-motor vector, 6 parametric nodes, 4 context nodes, 70 hidden nodes, 6 prediction output nodes which correspond to 3 motor values, 2 required torque values, and a hue value for the center region of the visual field. In the actual behavior generation process, the predicted motor values are used as the actual motor values at the next time step. The rest of the values of the sensory vector are not predicted to reduce the learning time. The module can enable the robot to generate behavior appropriately without predicting the entire sensory-motor vector.

It should be emphasized that the behavioral sequences are not separable. Hence they can not be decomposed into plausible primitives, unlike the sentences which can be broken down into words. This implies that no direct correspondences between some parts of behavioral sequences and certain words can be established. As discussed in later sections, the correspondences are acquired in a non-trivial way.

3.4 Parametric Bias Binding Method

In the proposed model, corresponding sentences and behavioral sequences are constrained to have the same PB vectors in both modules. Under this condition, corresponding behavioral sequences can be generated naturally from sentences.

When a sentence $\boldsymbol{s}_i$ and its corresponding behavioral sequence $\boldsymbol{b}_i$ have the same PB vector, we can obtain $\boldsymbol{b}_i$ from $\boldsymbol{s}_i$ as follows:

$$RNNPB_B(RNNPB_L^{-1}(\boldsymbol{s}_i)) \quad \rightarrow \quad \boldsymbol{b}_i \qquad (6)$$

where $RNNPB_L$ and $RNNPB_B$ are abstracted operators for the linguistic module and the behavioral module, respectively.

The PB vector p_{s_i} is obtained by recognizing the sentence $\boldsymbol{s}_i$. Because of the constraint that corresponding sentences and behavioral sequences must have the same

[1] For the region in which there is no colored area, the hue takes a pre-defined constant value 1.0, and the bottom border position takes 0.0, which designates very far.

PB vectors, p_{b_i} is equal to p_{s_i}. Therefore, we can obtain the corresponding behavioral sequence $\boldsymbol{b}_i$ by utilizing the behavioral module with p_{b_i}. In the same way we can also obtain the $\boldsymbol{s}_i$ from $\boldsymbol{b}_i$. Thus, sentences and behavioral sequences are connected bi-directionally as observed in the mirror system (Rizzolatti et al., 1996).

The constraint is implemented by introducing an interaction term into part of the update rule for the PB vectors (equation (3)).

$$p_{s_i} = p_{s_i}^{old} + \delta p_{s_i} + \gamma_L \cdot (p_{b_i}^{old} - p_{s_i}^{old}) \quad (7)$$

$$p_{b_i} = p_{b_i}^{old} + \delta p_{b_i} + \gamma_B \cdot (p_{s_i}^{old} - p_{b_i}^{old}) \quad (8)$$

where γ_L and γ_B are positive coefficients that determine the strength of the binding. Equations (7) and (8) are the constrained update rules for the linguistic module and the behavior module, respectively. Under these rules, the PB vectors of a corresponding sentence $\boldsymbol{s}_i$ and behavioral sequence $\boldsymbol{b}_i$ attract each other.

Actually, the corresponding PB vectors p_{s_i} and p_{b_i} need not be completely equalized to acquire a correspondence. The epsilon errors of the PB vectors can be neglected because of the continuity of PB spaces.

3.5 Generalization of Correspondences

The compositional semantics of a simple language self-organized in our model is now explained. Originally, "compositionality" referred to the characteristic nature of semantics in which the meaning of a sentence can be represented as a combination of the meanings of the words. However, for compositional semantics, a substantial requirement is the meaning of an unlearned sentence can be derived from the meanings of known sentences.

Our model implements compositional semantics without introducing any explicit representations of the meanings of words. We instead regard the model as acquiring compositional semantics when it can generate appropriate behavioral sequences from all sentences without learning all correspondences.

To achieve this, an unlearned sentence and its corresponding behavioral sequences must have the same PB vector. Nevertheless, the PB binding method only equalizes the PB vectors for given corresponding sentences and behavioral sequences (e.g., equation (7) and (8)).

Implicit binding, or in other words, inter-module generalization of correspondences, is achieved by dynamic coordination between the PB binding method and the intra-module generalization of each module. The local effect of the PB binding method spreads over the whole PB space, because each individual PB vector depends on the others in order to self-organize PB structures reflecting the relationships among training data. Consider the situation in which the linguistic PB vector of "`point red`" is perturbed via the PB binding method. To keep the relationship among sentences, the perturbation of "`point red`" is propagated to "`point *`" and "`* red`", and then spreads over the whole linguistic PB space. A similar process occurs in the behavioral module.

Thus, the PB structures of both modules closely interact via the PB binding methods. Finally, both PB structures converge into a common PB structure, in which the structures of both sentences and behavioral sequences are unified. Under the condition that both modules share a common PB structure, all corresponding sentences and behavioral sequences then share the same PB vectors automatically.

4. Experiments

To observe self-organization of the compositional semantics of language based on the behavioral experiences of a robot, we designed three experiments. Here, we briefly explain each experiment.

In experiment I, only the linguistic module was employed to investigate the acquisition of the pure syntactic structure of language. The module acquired the complete syntax by learning 14 of the 18 possible sentences.

In experiment II, only the behavioral module was employed to investigate the acquisition of the pure embodied structure of the behaviors. The module learned sensory-motor sequences of all nine behavioral categories in a supervised manner.

To generate behaviors robustly, the behavioral module needs to acquire not the behavioral trajectories themselves but the functionality to generate motor commands coupled with the environment. Thus, 10 different sensory-motor sequences were given for each behavioral category, and a total of 90 training sequences were generated through human guidance. To differentiate these sequences, the positions of the objects in the environment were slightly varied for each generated sequence. The variation was at most 20 percent of the distance between the starting position of the robot and the original position of each object in every direction.

Finally, in experiment III, both modules were employed to investigate the acquisition of the compositional semantics. As in the previous experiment, the linguistic module learned 14 of the 18 sentences and the behavioral module learned the behavioral sequences of the nine categories. For each sentence, five different pairs of sentences and the corresponding behavioral sequences were learned. Thus, the linguistic module learned 70 sentences, and the behavioral module learned 70 behavioral sequences with binding and 20 without binding. In addition, the behavioral module learned the same 90 behavioral sequences without binding.

5. Results and Analysis

The results of the last experiment showed that the compositional semantics of a simple language could be acquired by generalizing the correspondences between sentences and behavioral sequences in the proposed model. By employing the acquired semantics, unlearned sentences could be recognized and understood in order to generate appropriate behaviors.

This generalization was achieved by sharing a common PB structure between modules. In the following analysis, we show that this common structure possesses (1) the combinatorial properties of the pure linguistic structure and (2) a metric based on the similarities of behavioral sequences from the pure behavioral structure.

5.1 Linguistic Module: Syntactic Structure

In this section we analyze the results of experiment I, in which only the linguistic module learned. The final averaged output error of each node was 0.0060 after 50000-step learning. Analysis of the linguistic module reveals two important properties: (1) acquisition of syntax through the process of generalizing sentences, and (2) the structure of PB space, which reflects the combinatorial structure of the sentences.

In this experiment, 14 of the 18 two-word sentences were employed as training sentences. The four remaining sentences, "`point green`", "`point right`", "`push red`", and "`push left`" were used to evaluate the intra-module generalization.

The linguistic module acquired the underlying syntax from the given sentences. This was confirmed by the fact that the module could correctly generate only grammatical sentences. In other words, the minimized regeneration error of grammatical sentences (verb-noun) is much smaller than of ungrammatical 2-word sequences (verb-verb, noun-verb, noun-noun). For a given 2-word sequence $\boldsymbol{s}$, we define the minimized regeneration error as the total RMS error between $\boldsymbol{s}$ and the regenerated sequence $\hat{\boldsymbol{s}}$, where

$$\hat{\boldsymbol{s}} \quad \leftarrow \quad RNNPB_L(RNNPB_L^{-1}(\boldsymbol{s})). \qquad (9)$$

Here, it should be noted that the most likely PB vector encoding $\boldsymbol{s}$ can be obtained by the recognition process. Therefore, $\hat{\boldsymbol{s}}$ is the most likely estimation of $\boldsymbol{s}$ in the generable sequences of the module.

Next, we show the acquired PB space representing the combinatorial structure of the sentences. Figure 4a is a plot of the PB vectors obtained by recognizing all the sentences including the unlearned ones. The PB space is 6-dimensional, so only two selected parametric nodes are plotted here[2]. We can find three congruent sub-structures for each verb (Figure 4b), and six congruent sub-structures for each noun (Figure 4c)[3].

The congruency of the sub-structures for verbs and nouns represents the combinatorial structure of the sentences. This means that the words included in a sentence determine the PB vector of the sentence. For example, the PB vector of "`push green`" must be on the cross-over point between sub-structures for `push` and `green`. In addition, the PB vectors of the unlearned sequences take part in these sub-structures, and this supports intra-module generalization.

5.2 Behavioral Module: Behavioral Similarity

In this section, we analyze the results of experiment II, in which only the behavioral module learned. The final averaged output error of each output node was 0.012 after 50000-step learning.

After successful learning, the robot can generate the specified behavior regardless of the arrangement of the objects so long as the red, blue, and green objects are placed from the left to the right (Figure 5). The behavioral module generalizes behavioral sequences as functions which generate motor commands depending on the current sensory images, and therefore the robot can generate appropriate behavioral trajectories even in a noisy environment.

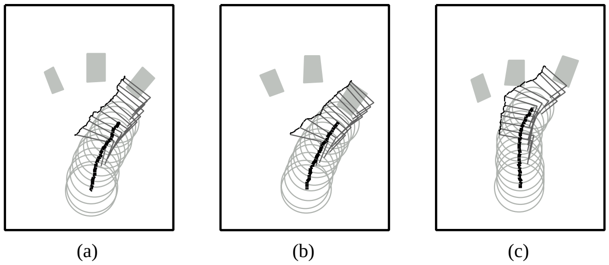

Figure 5: These three trajectories (a, b, and c) show that the robot could regenerate PUSH-G behavior regardless of the position of the objects. The robot could regenerate behavioral sequences appropriately so long as the red, blue, and green objects were placed in order from the left to the right, and the robot could see the target object at the starting position. This means that the behavioral module learned the given behavioral patterns by acquiring appropriate sensory-motor functions. The thick black lines represent trajectories of the center of the robot, and the thin black lines represent trajectories of the tip of the arm.

Although no combinatorial structure among the behavioral categories could be seen in the acquired structure of the behavioral module, we did find a metric reflecting the similarity of behavioral sequences (Figure 6). In this figure, the 6-dimensional PB vectors acquired at

[2] Unlike in figure 6 and figure 7, we did not employ the conventional principal component analysis (PCA) method to determine the axes. The plot projected on the surface obtained by PCA is not suitable for seeing congruent sub-structures because each sub-structure is plotted on a line.

[3] You can find similar diagrams in "Cours de linguistique générale (de Saussure, 1966)."

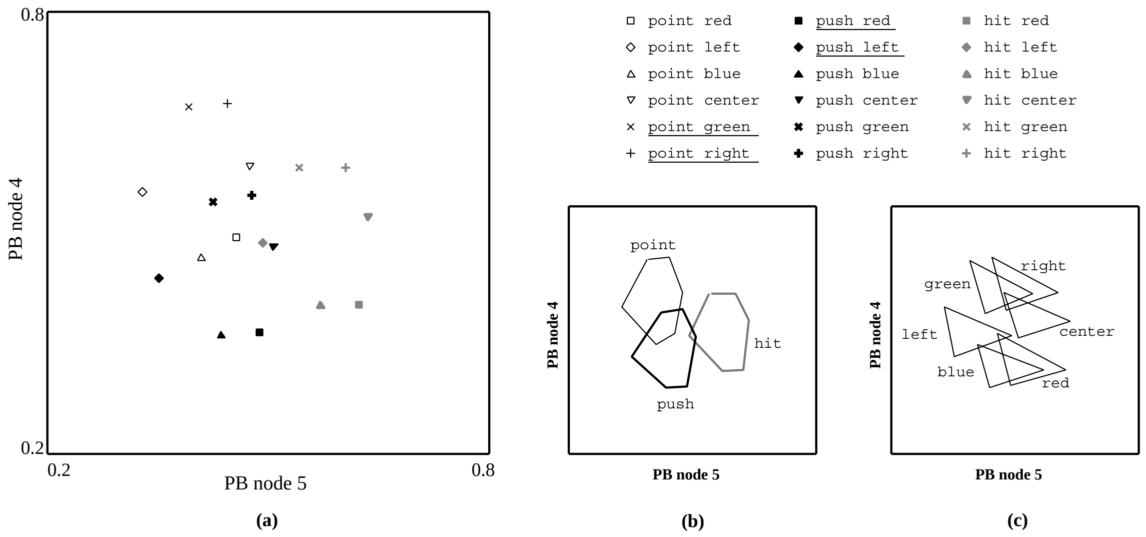

Figure 4: A plot of the linguistic PB space acquired without binding. The underlined sentences were unlearned. Two of six nodes are selected for ease of observing the combinatorial structure (a). The plot has three congruent sub-structures for each verb (b), and six congruent sub-structures for each noun (c).

the learning phase was projected onto a surface which maximizes the deviation of the plots using the conventional principle component analysis (PCA) method.

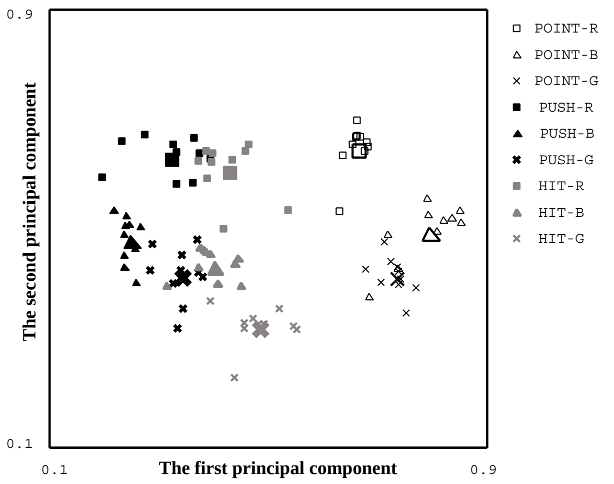

Figure 6: A plot of the behavioral sequences in the PB space acquired without binding. The large plots show the averaged PB vectors for each behavioral category. The 6-dimensional PB space was projected onto a surface that maximizes the deviation of the plots by using the PCA method. The accumulated contribution ratio of the two axes is 70%.

As nine clusters were observed, corresponding to the nine behavioral categories, we can conclude a metric based on the behavioral categories is internally represented in the PB space. This was confirmed by an additional experiment, in which it was shown that the robot could robustly regenerate the behavioral trajectory of each behavioral category using a PB vector value from the corresponding cluster.

We could not find a combinatorial arrangement of the behavioral clusters differing from the linguistic PB space shown in section 5.1. In the plot of the averaged PB vector over each behavioral category, this non-combinatorial arrangement is clearly observed (see large plots in figure 6). For example, it is not possible to estimate the PB vector of PUSH-G from the relationship among the PB vectors of PUSH-B, HIT-G and HIT-B. This confirms that the PB structure reflecting the behavioral sensory-motor sequences has no combinatorial property.

5.3 Unified Structure

In this section, we analyze the results of experiment III, in which we combined the linguistic and behavioral modules by using the PB binding method to learn the sentences and behavioral sequences simultaneously. The final averaged output error of each node was 0.0091 for the linguistic module, and 0.025 for the behavioral module after 50000-step learning. The analysis reveals that the PB binding method could fill an essential role in self-organizing the compositional semantics of language through the behavioral experiences of the robot.

As mentioned in section 4,the training data for this experiment did not include all the correspondences. Four sentences (“`point green`”, “`point right`”, “`push red`”, and “`push left`”) were not included in the training data for the linguistic module. As a result, although the behavioral module was trained with the behavioral sequences of all behavioral categories, those in two of the categories, whose corresponding sentences were not in the linguistic training set (POINT-G and PUSH-R), could not be bound (c.f., Figure 2).

The most important result was that these dangling behavioral sequences could be bound with appropriate sentences. That is to say, the resulting semantics could recognize all four unlearned sentences and properly generate the corresponding behaviors. This means that both modules acquired a common PB structure by generalizing the given correspondences.

Comparing the PB spaces of both modules shows that they indeed shared a common structure as a result of binding. The linguistic PB vectors are computed by recognizing all the possible 18 sentences including 4 unseen ones (Figure 7a), and the behavioral PB vectors are computed at the learning phase for all the corresponding 90 behavioral sequences in the training data (Figure 7b). The acquired correspondences between sentences and behavioral sequences can be examined according to equation (6). In particular, the coincidence of the four unlearned correspondences (“`point green`”↔POINT-G, “`point right`”↔POINT-G, “`push red`”↔PUSH-R, and “`push left`”↔PUSH-R) demonstrates acquisition of the underlying semantics, or the generalized correspondences.

The acquired common structure has two striking characteristics: (1) the combinatorial structure originated from the linguistic module, and (2) the metric based on the similarity of behavioral sequences originated from the behavioral module. The interaction between modules enabled the linguistic PB space (Figure 7a) to simultaneously self-organize not only the combinatorial structure but also the embodied structure. We can see this embodied structure introduced into the linguistic PB space as the similarity of the PB vectors of sentences that correspond to the same behavioral category. For example, the two sentences corresponding to POINT-R (“`point red`” and “`point left`”) are encoded in similar PB vectors. Recall here that such a metric nature was not observed in experiment I (Figure 4a). All nouns were plotted symmetrically in the PB space by means of the syntactical constraints.

At the same time, the combinatorial structure was introduced into the behavioral module (Figure 7b). In contrast to the behavioral PB space acquired in experiment II (Figure 6), we can find geometric regularity in the relationships among behavioral categories.

The PB vectors of the unlearned sentences and the corresponding behavioral sequences successfully coincided without binding because of the common structure shared by both modules. As explained in section 3.5, the structural unification of both PB spaces is realized by the local interaction of PB vectors. When a corresponding sentence and behavioral pattern pair is bound according to equations (7) and (8), the original PB structural regularity of both modules recedes. For example, the displacement of the linguistic PB vector of “`point red`” breaks the congruency reflecting the syntactic relationships among sentences (c.f., Figure 4). In the subsequent learning process of the linguistic module, all linguistic PB vectors are updated to recover regularity among sentences according to equations (1) to (3). Thus a local perturbation introduced by the PB binding spreads over the whole PB space. We especially note that the learning process affects the PB vectors for the unlearned sentences owing to the intra-module generalization. Similar processes are also observed in the behavioral module. As a result of the iterative interaction between the modules, they share a common PB structure.

The above observation thus confirms that the embodied compositional semantics was self-organized through the unification of both modules, which was implemented by the local PB binding method. We also made experiments with different test sentences, and confirmed that similar results could be obtained.

6. Discussion and Summary

Our simple experiments showed that the minimal grounded compositional semantics of our language can be acquired by generalizing the correspondences between sentences and the behavioral sensory-motor sequences of a robot. Our experiments could not examine strong systematicity (Hadley, 1994), but could address the combinatorial characteristic nature of sentences. That is to say, the robot could understand relatively simple sentences in a systematic way, and could understand novel sentences. Our 2-word language is very similar to the language $\mathcal{L}_0$ in (Evans, 1981), and the argument about the compositional semantics of $\mathcal{L}_0$ holds true for our language as well. Therefore, our results can elucidate some important issues about the compositional semantic representation.

In our experiments, compositionality was implemented in a non-trivial way. The embodied meaning of a word was implicitly realized in terms of the relationships among the meanings of sentences based on behavioral experiences. That is to say, the robot could understand a sentence by means of a generated behavior as if the meaning of the sentence were composed of the meanings of the words included in it. Our model does not require any pre-programming of syntactic information, such as symbolic representation of word meaning, a predefined combinatorial structure in the semantic domain, or behavior routines. Instead, the essential structures accounting for both compositionality and generalization are fully self-organized in the iterative dynamics of the RNN, through the structural interactions between language and behavior using the PB binding method.

We claim that the acquisition of word meaning and syntax can not be separated from the standpoint of the

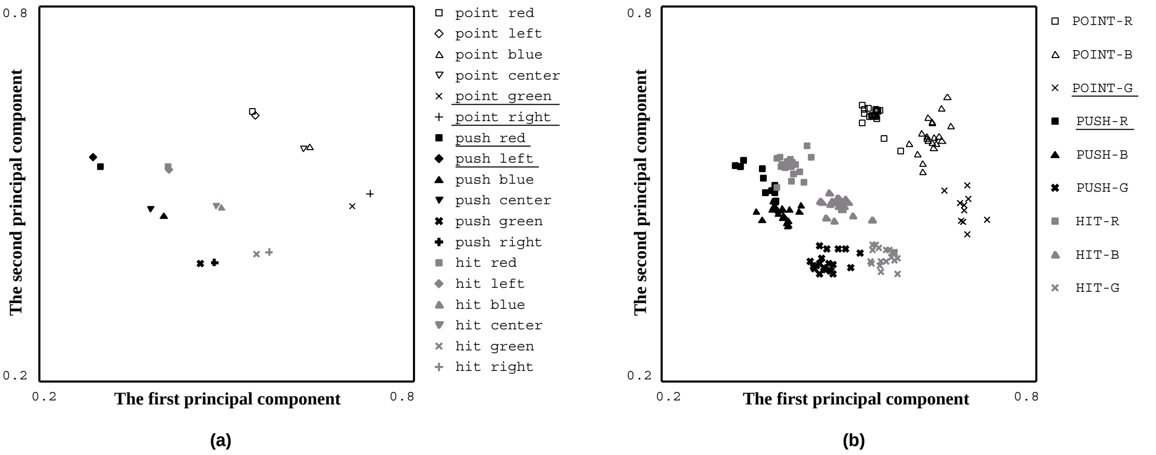

Figure 7: PB plots of the bound linguistic module (a) and the bound behavioral module (b). Both plots are projections of the PB spaces onto the same surface determined by the PCA method. Here, the accumulated contribution rate is about 73%. Unlearned sentences and their corresponding behavioral categories are underlined.

symbol grounding problem (Harnad, 1990). The meanings of words depend on each other to compose the meanings of sentences (Winograd, 1972). Consider the meaning of the word "`red`." The meaning of "`red`" must be something which combines with the meaning of "`point`", "`push`" or "`hit`" to form the grounded meanings of sentences. Therefore, a priori definition of the meaning of "`red`" substantially affects the organization of the other parts of the system, and often results in further preprogramming. This means that it is inevitably difficult to explicitly extract the meaning of a word from the meaning of a sentence.

Our model avoids this difficulty by implementing the grounded meaning of a word implicitly, in terms of the relationships among the meanings of sentences based on behavioral experiences. Thus, the robot can understand "`red`" through its behavioral interactions in the designed tasks in a bottom-up way (Tani, 1996). A similar argument holds true for verbs. For example, the robot understands "`point`" through pointing at red, blue, and green objects. Moreover, it should be noted that the meanings of nouns and verbs also depend on each other. One can not understand that a verb takes a noun as its object prior to the acquisition of semantics.

Next, we focus on generalization in the behavioral module and its effects on the common semantic structure. We assume that the meaning of a sentence is represented as a corresponding category of behavioral patterns. Although clarifying to what extent this assumption is appropriate is an important future goal, this assumption plays an essential role in the self-organization of a common semantic structure. For example, behavioral generalization (or categorization) enables the linguistic module to learn the semantic similarity of "`point red`" and "`point left`", which can not be learned from syntactic properties.

In a separated learning approach, the meaning of a sentence is represented as a combination of the meanings of words, which are labeled by the words themselves. This assumption allows the acquisition of sentence meanings to be translated into two relatively separate steps, the acquisition of word meanings and the acquisition of syntax (Kirby and Hurford, 2001). However, it is difficult to find the correspondence of a word meaning in the spatio-temporal behavioral patterns. This difficulty could be avoided by introducing an adequate set of behavioral routines. However it is often a non-trivial task and tends to result in a significant amount of preprogramming.

In contrast, our model acquires not just a mere mapping between sentences and behavioral patterns. The PB spaces of both modules (a.k.a., domains of the acquired mapping) acquire the internal structures suitable for realizing a structural mapping through mutual interaction between both modules. As mentioned in section 5.2, the behavioral module learned multiple behavioral patterns in a generalized manner, and categorized them with respect to the relationships between sensory images and motor commands. Therefore, the semantic domain is self-organized as some sort of a functional space, in which each function outputs motor commands from various input sensory information to generate behavioral patterns in a specific category. This behavioral generalization is important not only to the self-organization of the common semantic space, but also to realizing robust generation of behavioral patterns regardless of the perturbations in the environment.

To the summary, the current study has shown the importance of generalization of the correspondences between sentences and behavioral patterns in the acquisition of an embodied language. In future studies, we plan to apply our model to larger language sets. In the current experiment, the training set consists of a large fraction of the legal input space, when compared with related works. Such a large training set is needed because our model has no a priori knowledge of syntax and composition rules. However, we think that our model requires relatively fewer fraction of sentences to learn a larger language set, for a given degree of syntactic complexity.

References

Billard, A. (2002). Imitation: a means to enhance learning of a synthetic proto-language in an autonomous robot. In Dautenhahn, K. and Nehaniv, C. L., (Eds.), *Imitation in Animals and Artifacts*. MIT Press.

de Saussure, F. (1966). *Course in General Linguistics*. McGraw Hill.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Evans, G. (1981). Semantic Theory and Tacit Knowledge. In Holzman, S. and Leich, C., (Eds.), *Wittgenstein: To Follow a Rule*. London: Routledge and Kegan Paul.

Fodor, J. (1999). Why Compositionality Won't Go Away: Reflections on Horwich's 'Deflationary' Theory. Technical Report 46, Rutgers University.

Hadley, R. F. (1994). Systematicity revisited: reply to Christiansen and Chater and Niklasson and van Gelder. *Mind and Language*, 9:431–444.

Harnad, S. (1990). The symbol grounding problem. *Physica D*, 42:335–346.

Ito, M. and Tani, J. (2003). Generalization and Diversity in Dynamic Pattern Learning and Generation by Distributed Representation Architecture . Technical Report 2003-3, Lab. for BDC, Brain Science Institute, RIKEN.

Iwahashi, N. (2003). Language acquisition by robots – Towards New Paradigm of Language Processing –. *Journal of Japanese Society for Artificial Intelligence*, 18(1):49–58.

Jordan, M. and Rumelhart, D. (1992). Forward models: supervised learning with a distal teacher. *Cognitive Science*, 16:307–354.

Kirby, S. and Hurford, J. R. (2001). The Emergence of Linguistic Structure: an Overview of the Iterated Learning Model. In Parisi, D. and Cangelosi, A., (Eds.), *Computational Approaches to the Evolution of Language and Communication*. Springer Verlag.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. MIT Press.

Rizzolatti, G., Fadiga, L., Galles, B., and Fogassi, L. (1996). Promotor cortex and the recognition of motor actions. *Cognitive Brain Research*, 3:131–141.

Roy, D. K. (2002). Learning visually grounded words and syntax for a scene description task. *Computer Speech and Language*, 16.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning internal representations by error propagation. In Rumelhart, D. E. and Mclelland, J. L., (Eds.), *Parallel Distributed Processing*. Cambridge, MA: MIT Press.

Siskind, J. M. (2001). Grounding the Lexical Semantics of Verbs in Visual Perception using Force Dynamics and Event Logic. *Artificial Intelligence Research*, 15:31–90.

Steels, L. (2000). The Emergence of Grammar in Communicating Autonomous Robotic Agents. In Horn, W., (Ed.), *Proceedings of European Conference of Artificial Intelligence*, pages 764–769. IOS Press.

Sugita, Y. and Tani, J. (2002). A Connectionist Model which Unifies the Behavioral and the Linguistic Processes: Results from Robot Learning Experiments. In Stamenov, M. I. and Gallese, V., (Eds.), *Mirror Neurons and the Evolution of Brain and Language*. John Benjamins.

Tani, J. (1996). Model-Based Learning for Mobile Robot Navigation from the Dynamical Systems Perspective. *IEEE Trans. on SMC (B)*, 26(3):421–436.

Tani, J. (2003). Learning to generate articulated behavior through the bottom-up and the top-down interaction process. *Neural Networks*, 16:11–23.

Tani, J. and Ito, M. (2003). Self-Organization of Behavioral Primitives as Multiple Attractor Dynamics: A Robot Experiment. *IEEE Trans. SMC-A*, 33(4):481–488.

Thompson, C. A. and Mooney, R. J. (1998). Semantic lexicon acquisition for learning natural language interfaces. Technical Report AI98-273, The University of Texas at Austin.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, 3(1):1–191.