特集論文

Information Theoretic Competitive Learning and Linguistic Rule Acquisition

Ryotaro Kamimura — Information Science Laboratory, Tokai University
ryo@cc.u-tokai.ac.jp

Taeko Kamimura — Department of English, Senshu University
taekok@isc.senshu-u.ac.jp

Thomas R. Shultz — Department of Psychology, McGill University
shultz@psych.mcgill.ca

keywords: information maximization, competitive learning, language acquisition

Summary

In this paper, we propose a new information theoretic method for competitive learning, and demonstrate that it can discover some linguistic rules in unsupervised ways more explicitly than the traditional competitive method. The new method can directly control competitive unit activation patterns to which input-competitive connections are adjusted. This direct control of the activation patterns permits considerable flexibility for connections, and shows the ability to detect salient features not captured by the traditional competitive method. We applied the new method to a linguistic rule acquisition problem. In this problem, unsupervised methods are needed because children learn rules without any explicit instruction. Our results confirmed that the new method can give similar results as those by the traditional competitive method when input data are appropriately coded. However, we could see that when unnecessary information is given to a network, the new method can filter it out, while the performance of the traditional method is degraded by unnecessary information. Because data in actual cognitive and engineering problems usually contain redundant and unnecessary information, the new method has good potential for discovering regularity in actual problems.

1. Introduction

In this paper, we propose a new information theoretic method for competitive learning and apply it to a problem of linguistic rule acquisition. The new method can contribute to neural computing from three perspectives: (1) it is a competitive learning method that can directly control competitive unit activation patterns; (2) competition is realized by maximizing information in competitive units; and (3) the method can provide a tool and guiding principles to simulate a process of language acquisition. Let us discuss these points in more detail.

First, our new method for competition is considerably different from the traditional competitive method. A number of unsupervised learning methods have been based upon competitive learning [Grossberg 87, von der Malsburg 73, Fukushima 75, Kohonen 95]. In addition, a number of heuristic methods have been proposed to solve problems inherent in traditional competitive learning such as dead units: for example, leaky learning [Rumelhart and Zipser 86, Grossberg 87] a conscience method [Diesieno 88], frequency-sensitive learning [Ahalt et al. 90], rival penalized competitive learning [Xu 93] and lotto-type competitive learning [Luk and Lien 00]. However, because these traditional competitive methods must realize the appropriate competitive activation patterns indirectly by approximating input patterns over connections, it is clear that the methods need to be sensitive to detailed parts of input patterns [Rumelhart and Zipser 86, Hagan and Demuth 96]. On the other hand, by directly controlling competitive unit activation, the new information theoretic method can produce the same kind of connections as the traditional competitive method does, but without requiring representations of the input patterns. This permits considerable flexibility for connections, enabling networks to detect salient features not captured by the traditional competitive method.

Second, competition in the new method is realized by maximizing mutual information between in-

put patterns and competitive units. Information theoretic methods applied to neural computing have given promising results in various aspects of neural computing and pattern recognition [Linsker 88, Linsker 89, Linsker 92, Atick and Redlich 90, Becker 96, Becker and Hinton 93]. However, the limitations of these approaches have severely constrained their range of application. Contrary to those traditional information theoretic methods, the new method is easy to implement in any neural network architecture, and a process of information maximization can be intuitively interpreted. In brief, the information can be defined by using competitive unit activation. By maximizing the information, only one unit wins the competition, and all the other competitive units are off. In addition, the mutual information maximization forces all competitive units available to be equally used. This prevents only one competitive unit from always winning, and thus it can be expected to eliminate the dead neuron problem.

Third, in order to demonstrate the characteristics and better performance of the new information theoretic method, we applied it to an unsupervised language rule acquisition problem. Natural language acquisition processes have been so far simulated with supervised neural networks [Plunkett and Marchman 91, Marchman 93, Rumelhart and McClelland 86]. However, it has been argued that children acquire their first language without being given explicit target forms [Brown and Hanlon 70, Demetras et al. 86]. This is why unsupervised learning methods are needed to build more realistic models of language acquisition. Moreover, in past research, no guiding principle to govern overall acquisition has been postulated except for some technical learning rules. In this paper, we propose a principle of information maximization that realizes competition by which networks can eventually acquire linguistic rules.

2. Information in Systems

In this paper, we focus upon information stored in a system [Gatlin 72, Kamimura and Nakanishi 95]. Information is defined as the decrease in uncertainty from an initial state to another state after receiving an input signal. Initial uncertainty H_1 is described by first order entropy:

$$H_1 = -\sum_j p(j)\log p(j), \tag{1}$$

where $p(j)$ denotes the probability of the jth unit being turned on. First order entropy H_1 may be further decreased to second order uncertainty H_2 after receiving an input signal:

$$H_2 = -\sum_s \sum_j p(s)p(j \mid s)\log p(j \mid s), \tag{2}$$

where $p(s)$ represents the probability of the input signal s, and $p(j \mid s)$ denotes the conditional probability of the jth unit after receiving the s input signal. This uncertainty decrease or information is defined by

$$\begin{aligned} I &= H_1 - H_2 \\ &= -\sum_j p(j)\log p(j) \\ &\quad + \sum_s \sum_j p(s)p(j \mid s)\log p(j \mid s). \end{aligned} \tag{3}$$

3. Competition by Information Maximization

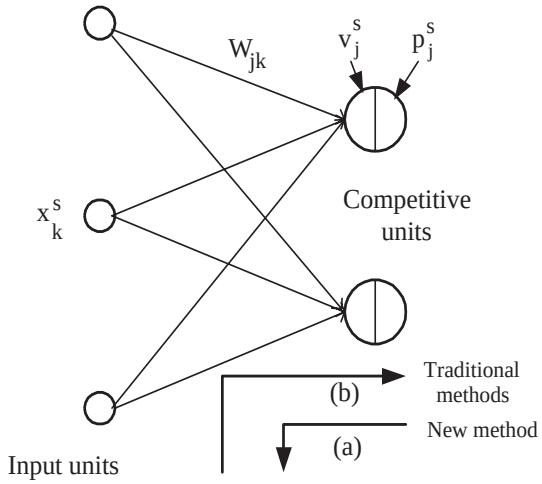


Fig. 1 A network architecture for defining information. The information theoretic method can directly control competitive unit activation patterns (a), while the traditional methods approximate input patterns over connections and indirectly control activation patterns (b).

Next, we apply this concept of information to a neural network architecture. As shown in Figure 1, the network architecture is composed of input units x_k^s and competitive units v_j^s. Let us define information for competitive units and try to control competitive unit activation. The jth competitive unit receives a net input from input units, and an output from the jth competitive unit can be computed by

$$v_j^s = f(\sum_k w_{jk}x_k^s), \tag{4}$$

where w_{jk} denotes a connection from the kth input unit to the jth competitive unit, and a function $f(t)$ is defined by $1/(1+\exp(-t))$. In modeling competition among units, one of the easiest ways is to normalize the outputs from the competitive units as follows:

$$p_j^s = \frac{v_j^s}{\sum_m v_m^s}. \tag{5}$$

The conditional probability is approximated by this normalized competitive unit output, that is,

$$p(j \mid s) \approx p_j^s. \tag{6}$$

Because input patterns are supposed to be uniformly given to networks, the probability of the jth competitive unit is approximated by

$$\begin{aligned} p(j) &= \sum_s p(s)p(j \mid s) \\ &\approx \frac{1}{S}\sum_s p_j^s \\ &= p_j. \end{aligned} \tag{7}$$

Using these probabilities, first order entropy is approximated by

$$\begin{aligned} H_1 &= -\sum_j p(j)\log p(j) \\ &\approx -\sum_j p_j \log p_j. \end{aligned} \tag{8}$$

Second order entropy H_2 is approximated by

$$\begin{aligned} H_2 &= -\sum_s \sum_j p(s)p(j \mid s)\log p(j \mid s) \\ &\approx -\frac{1}{S}\sum_s \sum_j p_j^s \log p_j^s. \end{aligned} \tag{9}$$

As second order entropy becomes smaller, the specific pairs of input patterns and competitive units become strongly correlated. By using these two types of entropy, information is approximated by

$$I \approx -\sum_j p_j \log p_j + \sum_s \frac{1}{S}\sum_j p_j^s \log p_j^s. \tag{10}$$

Differentiating the information I with respect to input-competitive connections w_{jk}, we have the final update rules:

$$\begin{aligned} \Delta w_{jk} = &-\beta \sum_s \left(\log p_j - \sum_m p_m^s \log p_m \right) Q_{jk} \\ &+\beta \sum_s \left(\log p_j^s - \sum_m p_m^s \log p_m^s \right) \\ &\times Q_{jk}, \end{aligned} \tag{11}$$

where β is the learning parameter and

$$Q_{jk} = \frac{1}{S} p_j^s (1 - v_j^s) x_k^s. \tag{12}$$

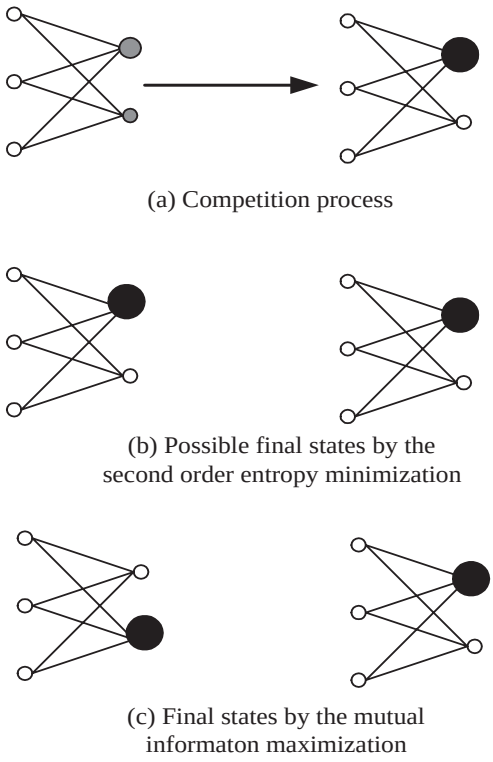


Fig. 2 Information maximization process (a), possible final activation patterns by second order entropy minimization (b) and mutual information maximization (c).

.

By using these update rules, mutual information is increased as much as possible.

Let us consider how information maximization applied to neural computing can be used to model competition. In the conventional competitive learning methods [Rumelhart and Zipser 86, Grossberg 87], distortion between the input vector and the weights is computed, and the unit with the smallest distortion is regarded as a winner. Then, the weight vector is updated toward the input vector. On the other hand, the information theoretic method simulates competition not by approximating input patterns but by directly controlling competitive unit activation patterns. At the initial state of learning, no information is given to a network, which corresponds to a minimum information state, as shown in the left figure of Figure 2(a). By update rules (11), information is increased, and reaches a final maximum information state in which only one competitive unit is turned on, while all the other units are close to off (the right figure of Figure 2(a)). This shows a fundamental process of information maximization to simulate competition among neurons. However, we can imagine a situation in which only one unit always wins, while all the other units are off, as shown in Figure 2(b). This corresponds to a case of dead neurons in conventional com-

petitive learning, and causes uncertainty in predicting final states by entropy minimization (second order entropy minimization) in neural networks [Kamimura and Nakanishi 95]. In our new method, this problem is clearly solved. In the information theoretic method, mutual information should be increased maximally. For this purpose, first order entropy $-\sum_j p_j \log p_j$ should be maximized; and at the same time, second order entropy $-1/S \sum_s \sum_j p_j^s \log p_j^s$ should be minimized. When an input pattern responds strongly to a specific competitive unit, second order entropy is minimized, which does not necessarily guarantee the final states in Figure 2(c). However, in mutual information maximization, first order entropy should be increased as much as possible simultaneously. This maximization is possible when the competitive units are uniformly distributed, that is, when the probability p_j is uniformly distributed. This corresponds to a situation of Figure 2(c), in which each competitive unit responds to a different input pattern. As can be inferred from the equation of mutual information, each competitive unit tends to respond to the same number of input patterns. In sum, the information theoretic methods can realize the same competitive activation patterns as those by the conventional competitive learning method without approximating the input vector over connections. For this reason, there is great flexibility for controlling connections.

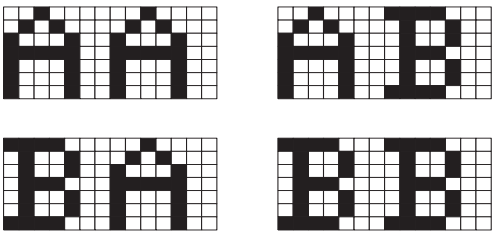


Fig. 3 Four words composed of two letters for the letter and word detection problem.

4. Letter and Word Detection

We conducted an experiment to show that the information theoretic competitive method can give the same kind of representation as the traditional competitive method. For the experiment, we used the letter and word detection problem of Rumelhart and Zipser [Rumelhart and Zipser 86]. In this experiment, we considered four words: AA, AB, BA and BB, each composed of two letters, A and B, as shown in Figure 3. Networks were supposed to classify these four words into two groups by detecting a letter in the words. In addition, the four words were supposed to be classified into four groups. This means that networks were expected to detect words composed of two letters.

The numbers of input and competitive units were 98 and 2, respectively. The learning parameter β was 1, and we did not use any techniques for improving performance. Figure 4 shows information as a function of learning epochs. We can see that information almost reaches its maximum point in only 50 epochs.

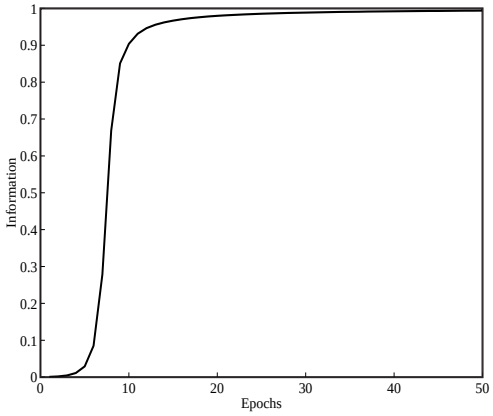
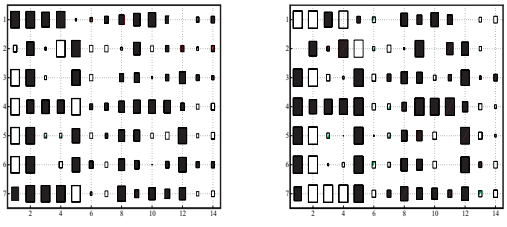


Fig. 4 Information as a function of learning epochs for the letter and word detection experiment.

These results confirmed that networks classify input patterns into two groups, independent of different initial conditions, as in the traditional competitive method [Rumelhart and Zipser 86]. Figure 5 shows connections into the first and the second competitive unit in the Hinton diagram. In the figure, black and white squares denote negative and positive connections, respectively. The size of the squares represents the absolute magnitude of connections. Strongly negative connections on the left hand side in (a) correspond to a letter B. On the other hand, strongly negative connections on the left hand side in (b) correspond to a letter A. These results show that networks can detect the first letter in the four words, and



(a) (b)

Fig. 5 Connections into the first (a) and second competitive unit (b).

thereby classifying the four words into two groups. This means that the letters are detected by networks.

When the number of competitive units was increased from two to four, each competitive unit responded to different words, that is, specialized competitive units could be detected. This means that each competitive unit became a word detector.

5. Inference of Linguistic Rules

5·1 Donatory Verb System

The following three experiments concern the Japanese donatary verb system. This system is a rather complicated one, and thus a brief explanation of how different donatary verbs are used in Japanese is necessary. Any action of transfer can be looked at from two different points of view. When viewed from the giver's perspective, an action of transfer is a "giving event"; on the other hand, the same action becomes a "receiving event" if it is viewed from the receiver's perspective.

In describing a giving event, English has only one verb *give*. On the other hand, Japanese has a set of different verbs, and a speaker of Japanese selects an appropriate verb, depending on contextual factors that characterize the relationship between the giver and receiver. The "in-group/out-group" relationship and the difference in social status between the giver and the receiver are the two contextual factors that determine the selection of appropriate donatory verbs in Japanese.

In this paper, we consider a giving event where the second contextual factor, i.e., social status, is kept constant. In this case, either *ageru* or *kureru* is used in Japanese, according to the notion of in-group/out-group relationship. A giver and a receiver in a giving event can be located on the "in-group/out-group continuum." The speaker is the core of the in-group, and the in-group consists of those to whom the speaker feels psychological proximity. As a basic rule, when the speaker or an in-group member is a giver and the receiver is from the out-group, *ageru* is chosen; on the other hand, when the giver belongs to the speaker's out-group and the receiver is from the in-group, *kureru* is used.

Next, consider a "receiving action." In describing a receiving event, English uses *receive*, while Japanese uses *morau*. Here again, the notion of in-group/out-group affects the use of the verb *morau*. As a basic principle, when using the verb *morau*, the grammatical subject of a sentence (which corresponds to the receiver in a receiving action) has to be the speaker or an in-group member. It is extremely rare to place an out-group member in the grammatical subject [Makino 96].

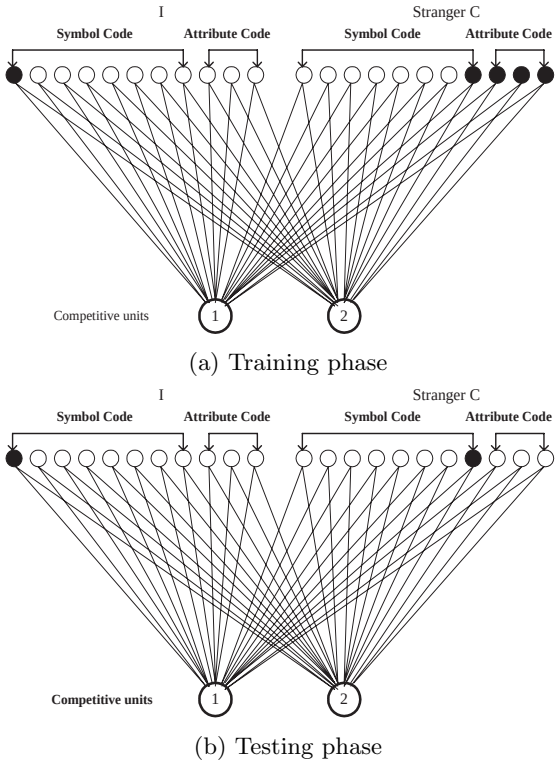


Fig. 7 A network architecture for inferring the two verbs for the training phase (a) and the testing phase (b).

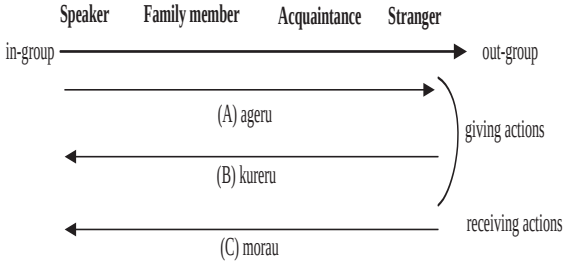


Fig. 6 Use of *ageru*, *kureru*, *morau*.

5·2 Coding System

When dealing with a symbolic system such as language, neural networks cannot immediately detect the semantic relations between discrete symbols. Symbols themselves are arbitrary in nature [Saussure 69]. Therefore, in the process of language acquisition, children need to establish a connection between a given symbol (e.g., a word "dog") and its meaning ("a four-

legged animal that barks") by utilizing the context in which the symbol appears. This contextual learning was proposed by Rumelhart and Zipser [Rumelhart and Zipser 86], and discussed fully by Ritter and Kohonen [Ritter and Kohonen 89].

Following Ritter and Kohonen [Ritter and Kohonen 89], we conducted a series of experiments to associate discrete symbols with their meanings by adopting two codes: symbol code and attribute code. The former corresponds to a label arbitrarily given in a symbolic system. The latter provides contextual information in which the former is used. In our experiment, we first trained networks by employing both the symbol and attribute codes. Then, by using the symbol code alone, we examined whether the networks had acquired the linguistic rules, because language is a collection of arbitrary symbols that correspond to our symbol codes.

5·3 Experiment 1

The objective of Experiment 1 is to show to what extent networks with an information maximization component can infer the correct use of the two verbs *ageru* and *kureru*. In this experiment, we considered four groups of participants who are of the same social status: the speaker, his family members, his acquaintances, and strangers. We further considered several instances of participants who fall into each group. The speaker is always unique; within the category of family members, the speaker's wife is the only possible person who has the same social status as he does. Thus, the final hypothetical participants in this experiment were eight as follows: the speaker ("I"), his family member ("my wife"), his acquaintances ("Taro-san," "Jiro-san," and "Hanako-san"), and strangers ("a," "b," and "c"). As Figure 6 shows, those four groups of participants can be located on the in-group/out-group continuum. The giving actions conducted between an in-group giver and an out-group receiver can be depicted by the arrow (A), while those conducted between an out-group giver and an in-group receiver can be represented by the arrow (B) in Figure 6.

Figure 7 shows a network architecture for this experiment. In this experiment, each participant was represented by a symbol code and an attribute code. The symbol code is a label corresponding to a person's name. Because there were eight participants in this experiment, eight bits were used to locally represent each participant. For instance, the speaker ("I") was encoded as 10000000, while a stranger "c" was encoded as 00000001.

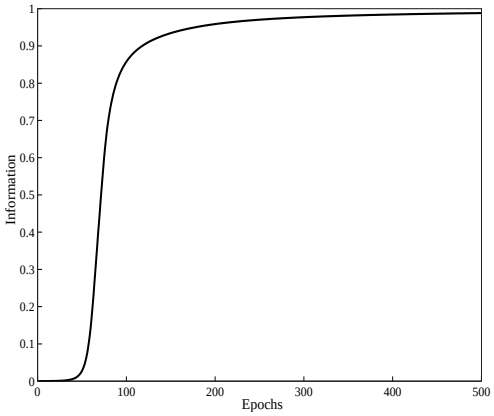


Fig. 8 Information as a function of the learning epochs for the two verbs.

The attribute code provides contextual information about relational distances among participants. For the attribute code, we represent the relational distance that each participant has on the in/out-group continuum as shown in Figure 6. For this purpose, the attribute codes for the speaker, his family member, his acquaintances, and strangers were encoded as $000, 001, 011, 111$, respectively. Because there were both a giver and a receiver, the number of input units was 22. For instance, consider a sentence in which the speaker ("I") gives something to a stranger "c." This sentence was encoded as 10000000000 for the speaker and 00000001111 for the receiver, as shown in Figure 7(a). As for the participants who belong to the same category on the in-group/out-group continuum (see Figure 6), it is difficult to determine which one is a member of the speaker's in-group; therefore, we did not consider the events of transfer that could be conducted by pairs of participants in the same category on the continuum, for example, "Taro-san" and "Jiro-san," who belong to the same group named "acquaintances." Consequently, the total number of pairs considered for encoding was 44. An experiment was conducted to examine whether only using the symbol code networks can infer the correct use of the two donatory verbs. First, we trained the networks to learn all the input patterns, as shown in Figure 7(a), including both the attribute and symbol codes. After checking whether networks had learned all the input patterns appropriately, we tested the performance by using only the symbol code, as shown in Figure 7(b).

Figure 8 illustrates the information I as a function of learning epochs. Information I was normalized for

its values to range between 0 and 1. As can be seen in the figure, information rapidly increases, and reaches its maximum at about 500 epochs.

The networks could accurately classify the input data into *ageru* and *kureru*. Figure 9 shows connections into the first competitive unit (a) and the second competitive unit (b). In the figure, we can clearly see that the networks respond strongly to the attribute code, containing information about the distances between the participants. At the bottom in Figure 9(a), two input sentences are shown. The first corresponds to a sentence "The speaker ('I') gives something to a stranger 'c'," while the second corresponds to a sentence "A stranger 'c' gives something to the speaker ('me')." For the former case, the first competitive unit is on by virtue of the positive connections for the second attribute code. In addition, we can see that by using only the symbol code, the networks can appropriately infer the rules. In the first sentence, the first bit in the first symbol code is on, which eventually turns on the first competitive unit. For the second sentence, the situation is reversed. Because the three bits in the first attribute code are all on for the stranger, the first competitive unit is off by the strongly negative connections. When we use only the symbol code for testing, the last bit in the first symbol code and the first bit in the second symbol code are on, which turns off the first competitive unit by the corresponding negative connections. Figure 9(b) shows connections to the second competitive unit. As shown in the figure, we can see that completely reversed patterns of connections are obtained.

Figure 10 shows connections into the first competitive units by the traditional competitive method *1. We can interpret these connections in the same way as in the information theoretic method. For example, consider a case in which the speaker ("I") gives something to a stranger "c" (the first input in Figure 10). The connections for the second attribute code are strongly positive, and this leads the first competitive unit to win the competition.

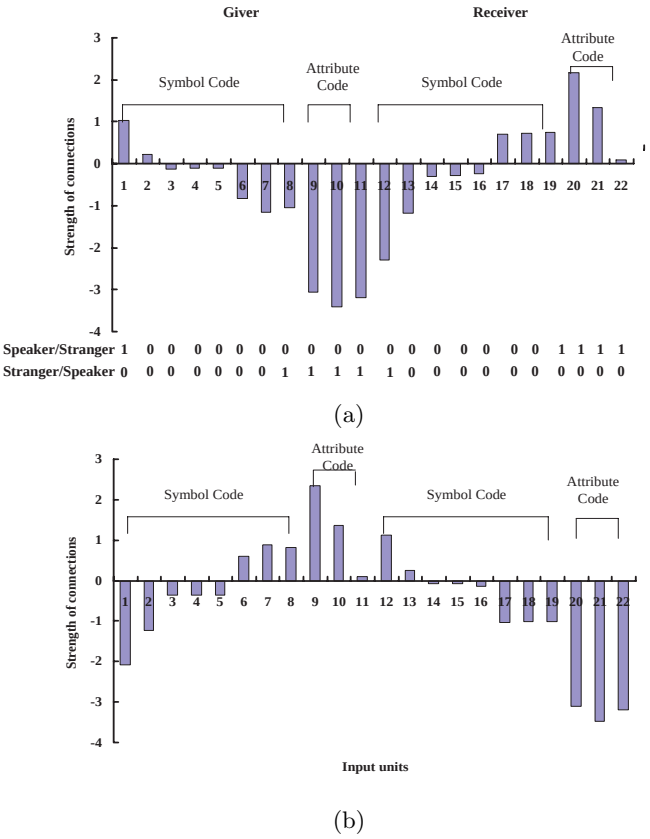


Fig. 9 Connections into the first competitive unit (a) and the second competitive unit (b) by the information theoretic method.

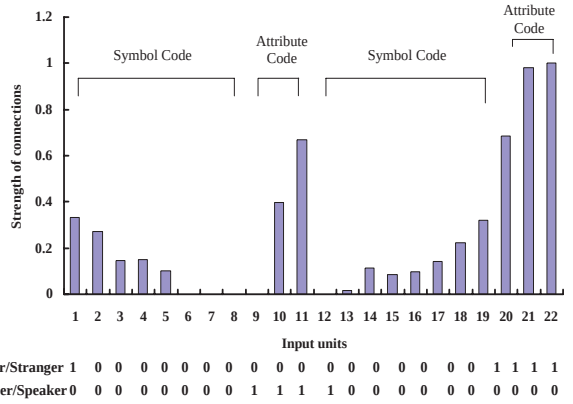


Fig. 10 Connections into the first competitive unit by the traditional method.

5·4 *Experiment 2*

In Experiment 2, the third verb *morau* was introduced (see Arrow (C) in Figure 6). As mentioned before, the verb *morau* is used when an action of transfer is looked at from the receiver's point of view, whereas *ageru* and *kureru* are used to describe an action of transfer viewed from the giver's perspective. In other words, when expressing a certain action of transfer in a sentence, if a speaker puts a receiver in the position of the grammatical subject, that sentence describes the action as a receiving event and thus the verb to be chosen is *morau*. In contrast, if a sentence has a giver as the grammatical subject, this sentence expresses a giving action and either *ageru* or *kureru* is used. Therefore, in our encoding system for Experiment 2, we needed to add an additional attribute code (attribute code No.2) to specify the two participants

*1 The competitive learning used in this paper was the conscience method [Diesieno 88]

in a certain action either as a giver or a receiver. By doing so, for example, we attempted to differentiate the following two sentences with the same grammatical subject: "I give something to a stranger," where the speaker ("I") is a giver and a stranger is a receiver versus "I receive something from a stranger," in which the speaker ("I") is a receiver while a stranger is a giver.

Figure 11 represents a network architecture for the receiving situation where a stranger is the giver and the speaker ("I") is the receiver. The giver and the receiver are represented by turning on and off the attribute code No. 2, respectively. We used the redundant two bits, because neither the information maximization nor the competitive method could classify appropriately input patterns into three groups with a single bit for the giver and receiver information.

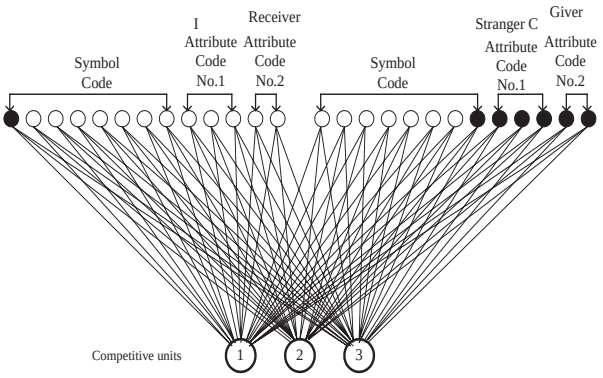


Fig. 11 A network architecture for inferring the three verbs.

As shown in Figure 12, information approaches the maximum in 1,000 epochs. This time, we decreased the learning parameter β from 1 to 0.5, because some instability was frequently observed when the parameter β was 1.

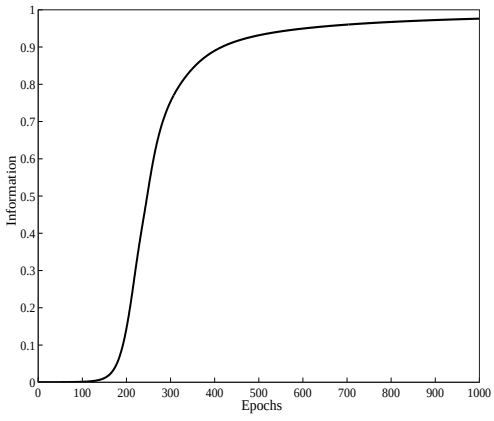
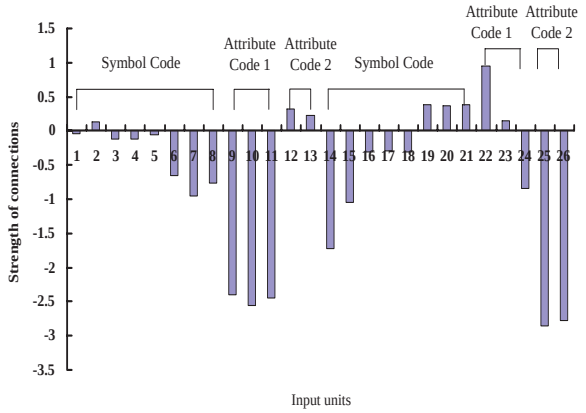
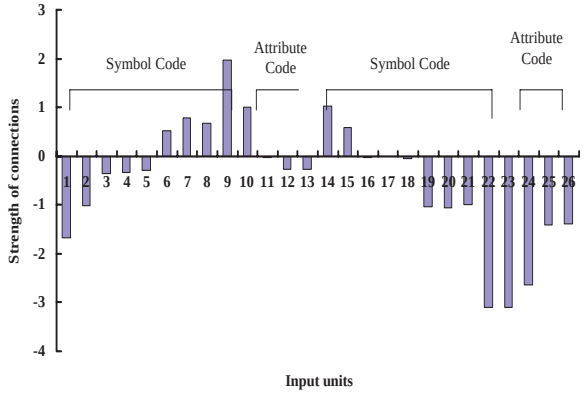


Fig. 12 Information as a function of the learning epochs for the three verbs.

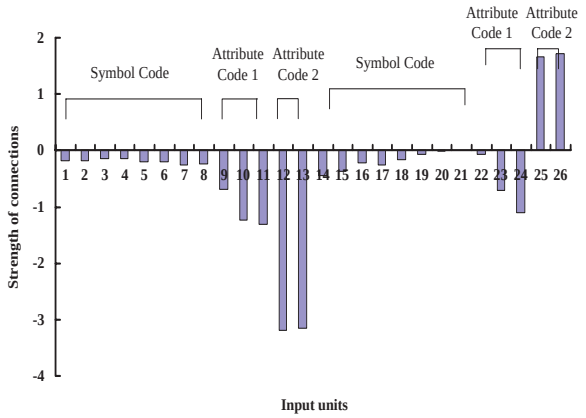
Figure 13 shows connections into the first (a), the second (b) and the third competitive unit (c). Connections into the first (a) and the second unit (b) correspond to connections (a) and (b) in Figure 9 in the previous section except connections from the attribute code No.2 for differentiating the giver and the receiver. Connections into the third competitive unit (c) respond strongly to the attribute code No.2. The connections into the third unit are used to differentiate the giver and receiver.



(a)



(b)



(c)

Fig. 13 Connections into the first (a), the second (b) and the third (c) competitive unit.

When we used the traditional competitive method, we could see that the networks can produce the same connections as those discussed in Experiment 2 except for the second attribute code No.2 for differentiating the giver and the receiver. However, we can find a striking difference between the information theoretic method and the traditional competitive method in terms of connections into the third competitive unit. Figure 14 shows connections into the third competitive unit by the traditional competitive method. We can see that networks respond almost equally to the attribute code No.1 and No.2. On the other hand, by using the information theoretic method, connections into the third competitive unit respond strongly to the attribute code No.2, as shown in Figure 13(c). This makes a great difference in performance between the two methods, which will be discussed in the next section.

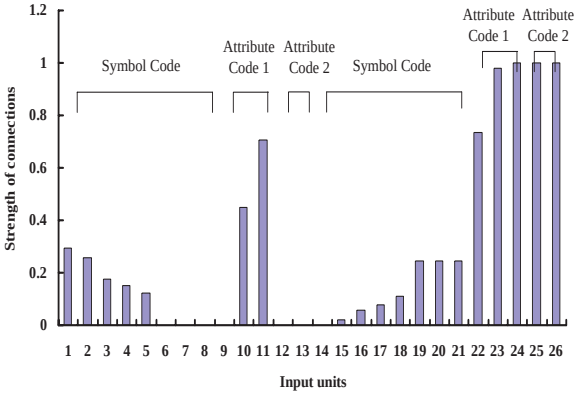


Fig. 14 Connections into the third competitive unit by the traditional method.

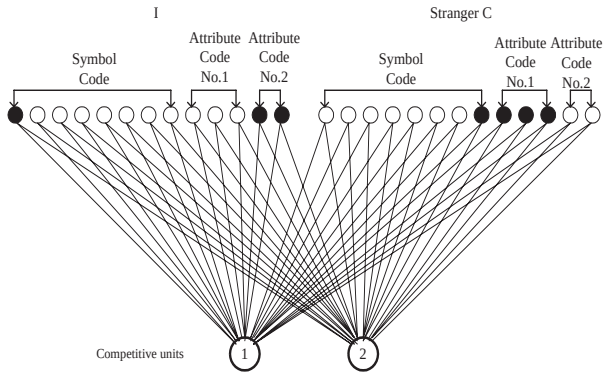


Fig. 15 Network architecture for the experiment 3.

learning method *2. The success rates are averaged over ten different runs. The information theoretic method always gives the correct answers for the ten different runs. On the other hand, the success rate by the traditional competitive method is about 0.87. The reason why the traditional method gives such a low success rate can be explained as follows.

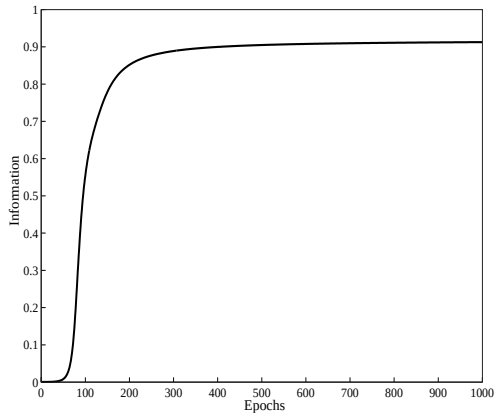


Fig. 16 Information as a function of the number of epochs for Experiment 3.

5·5 Experiment 3

Experiment 3 was conducted to clarify the differences between the information theoretic competitive method and the traditional method. In this experiment, we tried to classify the input patterns used in Experiment 2 into two groups. In this case, we naturally expect the networks to classify input patterns into the giving event and the receiving event.

Figure 15 shows a network architecture for the experiment. The number of competitive units was restricted to two. As can be seen in Figure 16, information reaches its maximum values in 500 epochs.

Figure 17 shows the success rates by the information theoretic method and the traditional competitive

Figure 18 shows connections into the first (a) and the second (b) competitive unit. The networks respond only to the attribute code No.2 for differentiating the two events of transfer. On the other hand, Figure 19 depicts connections into the first competitive unit by the traditional competitive method. As can be seen in the figure, the networks respond to both the attribute code No.1 and No.2. The traditional competitive method must produce appropriate activation patterns by imitating input patterns. In this experiment, the attribute code No.1 for repre-

*2 We conducted the same experiments with normalized input data. With the normalized data, the learning was slightly stabilized, but the success rates were almost the same as those presented here.

senting relational distances among participants is of no use in classification. The networks must ignore the unnecessary part of input data as much as possible. However, the traditional competitive method treats these two attribute codes equally and tries to imitate them over connections. This property of the traditional method tends to cause difficulty in classifying input patterns appropriately. On the other hand, the new information theoretic method does not imitate input patterns over connections. It directly controls the competitive activation patterns by which appropriate connections are generated. This means that if some information in input patterns are of no use for producing the competitive activation patterns, it can be ignored.

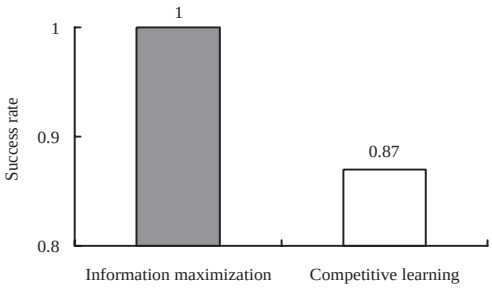


Fig. 17 Success rates by the information method (left) and the competitive method (right). The success rates were averaged over ten runs.

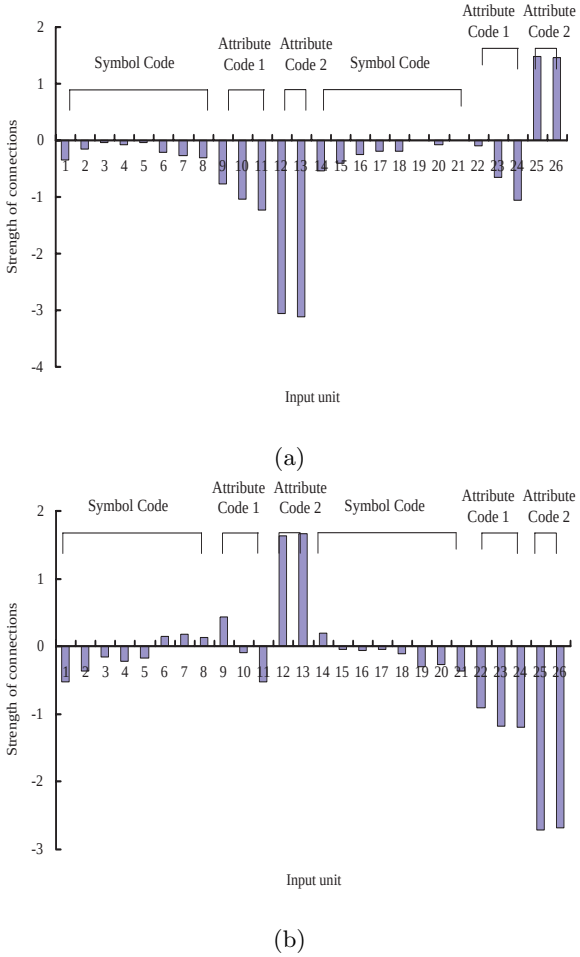


Fig. 18 Connections into the first (a) and the second (b) competitive unit.

6. Discussion

We discuss here the implications and the problems of the new method from psychological, computational and linguistic perspectives: (1) the new method is related to *attention* in human information processing; (2) the method can solve some shortcomings of the traditional methods; and (3) the coding system should be elaborated for more realistic simulation of language acquisition.

First, the new method is closely related to the psychological construct of *attention* [Anderson 80]. Human beings and neural networks are both highly parallel processing systems. However, one of the basic differences between the two is that human beings can focus on a particular piece of information while processing enormous amounts of data in parallel. As discussed in Section 5.5, the information theoretic method can classify input patterns by responding only to the attribute code No.2, ignoring the attribute code

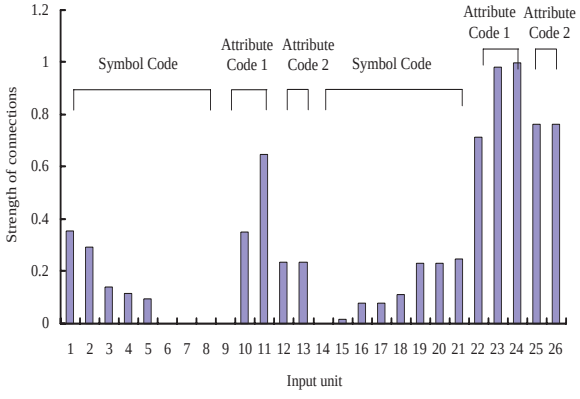


Fig. 19 Connections into the first competitive unit by the traditional method.

No.1. This means that the network can focus upon some specific features to achieve global activation patterns enforced by mutual information maximization. This suggests that the new method can simulate a process of attention in human information processing.

Second, as already suggested in Introduction, the conventional competitive method suffered from the so-called *dead* neurons. If data are complex or initial conditions are far from the final states, some neurons never win in competition, which degrades classification processes [Rumelhart and Zipser 86]. For solving this problem, many heuristic techniques have been proposed [Diesieno 88, Ahalt et al. 90, Xu 93]. These heuristic techniques and our methods are closely related. For example, Ahalt [Ahalt et al. 90] used entropy of competitive units for measuring the uniform utilization of different competitive neurons. He argues that if the entropy is higher, the number of dead neurons is smaller. This concept of entropy has been incorporated in our formulation of mutual information. In maximizing mutual information, first order entropy should be as large as possible. This means the equal use of all different competitive units. Though a detailed mathematical and experimental comparison will be necessary, in principle, the problem of dead neurons is solved in the information theoretic method.

Third, coding in linguistic problems is critical. Linguistic or symbolic elements have so far been coded arbitrarily in neural network approaches to linguistic problems [Rumelhart and Zipser 86, Marchman 93, Plunkett and Marchman 91]. Our coding was based on Ritter and Kohonen [Ritter and Kohonen 89]. The coding system in our linguistic experiments as well as those in other studies [Rumelhart and Zipser 86, Marchman 93, Plunkett and Marchman 91] are not necessarily systematic. For example, we added both the attribute code No.1 and the attribute code No.2 for classifying the donatory verbs. The properties of the first and the second code are different from each other, which might affect the performance of the information theoretic method. We need a more unified approach to coding for more realistic models of language acquisition. One of the possible solutions to this problem is that we could enumerate all possible semantic features of the donatory verbs. Then, each verb could be represented by combinations of these fundamental features.

7. Conclusion

We have proposed a new information theoretic method for competitive learning that can directly produce appropriate competitive activation patterns. Compared with the traditional methods that indirectly produce appropriate patterns by imitating input patterns, the new method has revealed better performance when unnecessary information exists in input patterns.

In addition, we have shown that information maximization used in the information theoretic method alone can allow unsupervised neural networks to acquire the complex grammatical rules underlying Japanese donatory verbs. In natural language acquisition, children are not always exposed to correct input data, but instead inputs surrounding children contain abundant unnecessary information. They must infer useful information for themselves [Anderson 80]. The new method, therefore, may be well suited to describe natural language acquisition processes.

The present study has been concerned only with competition. The next step is to explore spatial constraints of competitive units as in Kohonen's self-organization map [Kohonen 87]. However, we assume that spatial constraints can be incorporated in our framework [Kamimura et al. 01], because we do not imitate input patterns over connections. The spatial constraint is only one of those to be introduced in our method.

Finally, although further studies with more complex linguistic, psychological and engineering data are needed, the present results underscore the potential of the new information theoretic method in a wider range of research fields.

Acknowledgments

The authors are very grateful to anonymous reviewers for their valuable comments and suggestions.

♢ References ♢

[Ahalt et al. 90] Ahalt, S. C., andd P. Chen, A. K. K., and Melton, D. E. (1990). Competitive learning algorithms for vector quantization. *Neural Networks*, 3:277–290.

[Anderson 80] Anderson, J. R. (1980). *Cognitive Psychology and its Implication*. Worth Publishers, New York.

[Atick and Redlich 90] Atick, J. J. and Redlich, A. N. (1990). Toward a theory of early visual processing. *Neural Computation*, 2:308–320.

[Becker 96] Becker, S. (1996). Mutual information maximization: models of cortical self-organization. *Network: Computation in Neural Systems*, 7:7–31.

[Becker and Hinton 93] Becker, S. and Hinton, G. E. (1993). Learning mixture models of spatial coherence. *Neural Computation*, 5:267–277.

[Brown and Hanlon 70] Brown, R. and Hanlon, C. (1970). Derivational complexity and order of acquisition. In Hayes, R., editor, *Cognition and Development of Language*. Wiley, New York.

[Demetras et al. 86] Demetras, M. J., Post, K. N., and Snow, C. E. (1986). Feedback to first language learners: The role of repetitions and clarification questions. *Journal of Child Language*, 13:275–292.

[Diesieno 88] Diesieno, D. (1988). Adding a conscience to competitive learning. In *Proceedings of IEEE International Conference on Neural Networks*, pages 117–124. IEEE.

[Fukushima 75] Fukushima, K. (1975). Cognitron: a self-organizing multi-layered neural network. *Biological Cybernetics*, 20:121–136.

[Gatlin 72] Gatlin, L. L. (1972). *Information Theory and Living Systems*. Columbia University Press.

[Grossberg 87] Grossberg, S. (1987). Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science*, 11:23–63.

[Hagan and Demuth 96] Hagan, M. T. and Demuth, H. B. (1996). *Neural Network Design*. PWS Publishing Company, Boston, MA.

[Kamimura et al. 01] Kamimura, R., Kamimura, T., and Shultz, T. R. (01). Self-organization by information control(in press). In *Proceedings of IASTED International Conference on Applied Informatics*.

[Kamimura and Nakanishi 95] Kamimura, R. and Nakanishi, S. (1995). Hidden information maximization for feature detection and rule discovery. *Network*, 6:577–622.

[Kohonen 87] Kohonen, T. (1987). *Self-Organization and Associative Memory*. Springer-Verlag.

[Kohonen 95] Kohonen, T. (1995). *Self-Organization Maps*. Springer-Verlag.

[Linsker 88] Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21:105–117.

[Linsker 89] Linsker, R. (1989). How to generate ordered maps by maximizing the mutual information between input and output. *Neural Computation*, 1:402–411.

[Linsker 92] Linsker, R. (1992). Local synaptic rules suffice to maximize mutual information in a linear network. *Neural Computation*, 4:691–702.

[Luk and Lien 00] Luk, A. and Lien, S. (00). Properties of the generalized lotto-type competitive learning. In *Proceedings of International conference on neural information processing*, pages 1180–1185.

[Makino 96] Makino, S. (1996). *Uchi to Soto no Gengogaku (A Linguistic-cultural Study of Uchi and Soto*. ALC, Tokyo.

[Marchman 93] Marchman, V. A. (1993). Constraints on plasticity in a connectionist model of the english past tense. *Journal of Cognitive Neuroscience*, 5(2):215–234.

[Plunkett and Marchman 91] Plunkett, K. and Marchman, V. (1991). U-shaped learning and frequency effects in a multilayered perceptron: Implication for child language acquisition. *Cognition*, 38:43–102.

[Ritter and Kohonen 89] Ritter, H. and Kohonen, T. (1989). Self-organizing semantic map. *Biological Cybernetics*, 61:241–254.

[Rumelhart and McClelland 86] Rumelhart, D. E. and McClelland, J. L. (1986). On learning the past tenses of english verbs. In Rumelhart, D. E., Hinton, G. E., and Williams, R. J., editors, *Parallel Distributed Processing*, volume 2, pages 216–271. MIT Press, Cambrige.

[Rumelhart and Zipser 86] Rumelhart, D. E. and Zipser, D. (1986). Feature discovery by competitive learning. In Rumelhart, D. E., Hinton, G. E., and Williams, R. J., editors, *Parallel Distributed Processing*, volume 1, pages 151–193. MIT Press, Cambrige.

[Saussure 69] Saussure (1969). *Cours de Linguitique Generale*. Payot, Paris.

[von der Malsburg 73] von der Malsburg, C. (1973). Self-organization of orientation sensitive cells in the striata cortex. *Kybernetik*, 14:85–100.

[Xu 93] Xu, L. (1993). Rival penalized competitive learning for clustering analysis, rbf net, and curve detection. *IEEE Transaction on Neural Networks*, 4(4):636–649.

〔担当委員：新田克己〕

Received October 31, 2000.

Author's Profile

Ryotaro Kamimura (Member)

He is a professor in Information Science Laboratory in Tokai University. He is currently working in the areas of neural networks and their application to problems in linguistics, cognitive science and management information systems.

Taeko Kamimura

She is a professor in the Department of English in Senshu University. Her research interests include interrelations between first and second language writing and procedural facilitation to promote second language writing development.

Thomas R. Shultz

He is a professor in the Department of Psychology in McGill University. His research interests include cognitive development, cognitive science and the relation between knowledge and learning.