Attractor Neural Networks for Simulating Dyslexic Patients' Behavior

Shin-ichi Asakawa
Center for Information Sciences, Tokyo Woman's Christian University,
Zempukuji 2, 6, 1, Suginami, 1678585 Tokyo, Japan

Keywords: Attractor Neural Network, Reaction Time, Identification and Categorization Tasks, Orthography, Phonology and Semantics, Brain Damaged Patients.

Abstract: It was investigated that the ability of an attractor neural network. The attractor neural network can be applicable to various symptoms of brain damaged patients. It can account for delays in reaction times in word reading and word identification tasks. Because the iteration numbers of mutual connections between an output and a cleanup layers might increase, when they are partially damaged. This prolongation looks or behaves the delays of reaction times of brain damaged patients. When we applied the attractor neural network to the data of Tyler et al. (2000) for categorization task, it showed a kind of category specific phenomenon. In this sense, the attractor neural network could explain an aspect of the category specific disorders. In this sense the attractor network might simulate the human semantic memory organization. In spite of variations in data, and in spite of the simplicity of the architecture, the attractor network showed good performances. We could say that the attractor network succeeded in mimicking human normal subjects and brain damaged patients. The possibility of explaining the triangle model (Plaut & McClelland,1989; Plaut, McClelland, Seidenberg, and Patterson, 1996) also discussed.

1 INTRODUCTION

1.1 Category Specificity in Neuropsychology

Neuropsychological studies have revealed important insights such as an art and a structure of our semantic memories. In addition to this, Neuropsychological evidences from brain damaged patients might show the way of stores of semantic memory items. Among them, the category specificity are suggestive. Because the data of these patients often shows the double–dissociation between animate and inanimate objects. Warrington and her colleagues began to describe these kinds of phenomena in early 1980s. The discussion continues so far. This kind of patients often show the deficits of an identification, an naming, and a categorization task of animate objects, but the knowledge of inanimate objects (i.e. tools, outdoor objects, and tools, and so on) remain intact. On the other hand, there are another patients who are not able to identify, to name, and to categorize inanimate objects. However, these type of patients have an intact knowledge of animals. Based upon these evidences, Warrington and her colleagues have tried to explain that the structure of semantic memory and its nature. Why some types of brain damages patients cause animate specific category disorders, and the knowledge of inanimate objects is intact. On the contrary, another patients groups cause inanimate specific category disorders, and the show no decay of knowledge of animals. Would these data suggest that different contents of semantic memory are localized in the brain (maybe the left lateral inferior gyrus)? Might these data suggest that the information of these two categories are stored in a distributed manner in the brain? Or might these data emerge from the inter- and intra- correlations between items? In this paper, we intend to answer these questions.

In the literature, several kinds of category specific disorders were reported so far. These are fruits, vegetables, animals, and so on. Many researchers insisted that there existed at least two kinds of deficits; perceptual and functional knowledge (Warrington, 1981; Warrington and McCarthy, 1983; Warrington and Shallice, 1984a; Warrington and McCarthy, 1994). According to this theory, the category specificity cat be regarded as our semantic memories are organized by both perceptual and functional knowledge. (Warrington and Shallice, 1984b; Warrington and McCarthy, 1983; Warrington and McCarthy, 1987). War-

Asakawa S..
Attractor Neural Networks for Simulating Dyslexic Patients' Behavior.
DOI: 10.5220/0004170505820587
In *Proceedings of the 4th International Joint Conference on Computational Intelligence* (NCTA-2012), pages 582-587
ISBN: 978-989-8565-33-4

rington and her colleagues advocated that knowledge about musical instruments and jewelry were similar to animate objects. They also, on the other hand, insisted that inanimate objects and body parts could be identified as functional knowledge. According to their perceptional/functional hypothesis, the brain damages to the regions for dealing with perceptual semantic knowledge would cause the deficits of knowledge of animate objects. In other words, the difference between animate and inanimate objects might be different on the loci damaged. This hypothesis proposed by Warrington and her colleagues was also supported by the results of the neural network model (Farah and McClelland, 1991).

1.2 Representation of Semantic Memory

However, there exist studies that the semantic memory of animates had been damaged without lack of any perceptual knowledge. There are patients who show the deficits about animals without any specific disorders of perceptual knowledge(Caramazza and Shelton, 1998).

Can we say that the representations of perceptual and functional aspects of semantic memory would differentiate between animates and inanimate objects? Are the information of perceptual and functional knowledge stored separately in the brain? And therefore, are localized brain damages cause any category specificity? Can we say that that the category specificity suggests differences in the contents and the structures between categories?

Especially, there exists a kind of category specificity without any semantic memory damages. A hypothesis based upon this hypothesis is that each concept in out memory has been represented by the activation patterns of micro features, i.e. multidimensional vectors. A similar relationship between concepts could be regarded as overlapped activation patterns in the micro features.

We tried to represent data on the basis of the discriminability. The correlation matrix among items could be explained the category specificity. This method of memory representation was originally described by Devlin et al. (Devlin et al., 1998). The characteristics they adopted were enumerated as follows.

1. Specificity of features: Representation of semantic memory varies based on how they can be retrieved within the same category. Animates stored many perceptual features shared in the brain. On the other hand, inanimate objects have more discriminative features than animates.
2. Correlation: The co–occurrence of features strengthens in accordance with the correlation matrix of each item The concept of animates has higher feature correlations than that of inanimate objects.

Fig.1 shows the correlation matrix of each item calculated from data of Tyler et al. (2000). A comparison between an upper left square and a lower right square in the figure would be found that the upper left square (inanimate objects) have less mutual correlations than that of the mutual correlations within animate objects (the lower right square). Tyler et al. could be considered that they could control the stimuli.

Neural networks accounts of the brain damage insists that the correlations patterns among micro features are important in order to understand the category specific disorders. The researchers regard the neuropsychological evidences as verifications of these correlation patterns of micro features.

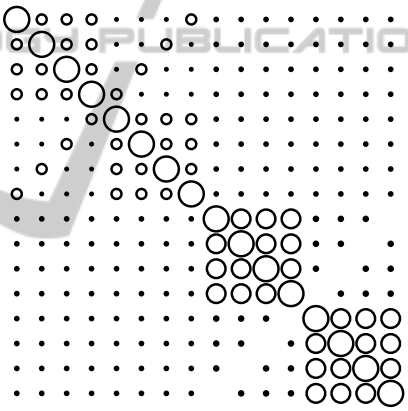

Figure 1: The correlation matrix calculated from the data by Tyler et al. (2000). The open circles mean positive (○) correlation coefficients, and the filled circles mean negative (●) correlation coefficients. The size of circles shows the correlation strengths. The upper left side of this figure shows inanimate objects, while the lower right side shows animates.

.It was able to be

2 ATTRACTOR NEURAL NETWORKS

Tyler et al. (Tyler et al., 2000) adopted a three–layered–network (perceptron) as a model dealing with the data described above. Although this type of network architecture is sufficient to account for the double dissociation between animate and inanimate objects, the attractor network seems to have more advantages than the perceptron describing some aspects of characteristics of semantic memory. For example, the

times of iterations between output and cleanup layers (Fig.2) until reaching the threshold of output criteria can be regarded as the prolonged reaction times of the brain damaged patients.

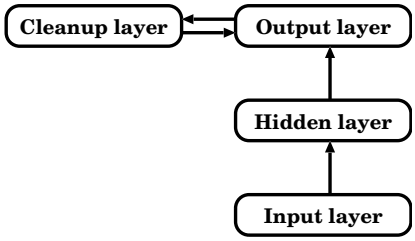

Figure 2: An attractor network originally proposed by Plaut and Shallice (1993).

Plaut et al.(Plaut et al., 1995; Plaut, 2001) adopted the attractor networks and tried to account for semantic dyslectic errors and compound errors from both visually and semantically. In the neural network, basic processing units are connected mutually. Upon this multidimensional space consisted of the values of processing units, the networks can change and retrieve the contents of adequate memories. In other words, when the network was given random initial values, the activation values of each processing unit would transit from the values to the value on the semantic memory space. The behavior of this network could be absorbed in an 'attractor'. There are many attractors corresponded to each memory item. If the initial values may be altered, the attractor network can be absorbed in the correct 'point' attractor. Thus, it is postulated that the 'basins' of each attractor are different.

Plaut et al. tried to explain the semantic errors, visual errors, and compounded both semantic and visual errors by using attractor networks. In neural networks, in general, units are connected mutually causing interactions among units. This interaction of activation patterns of each unit can be identified as the states of activation patterns of units. The activations of the units are transited from one to another as the memory retrievals. The transition from arbitrary initial states to some attractors are called the 'absorbability' of attractors. Therefore, we can consider that there are different basins of each words.

In case of the attractor networks, each attractor corresponds to each concept, and its basin represents its range to be absorbed in. Even if the state of the network defined by the activations of each unit would be changed on influences either noises or perturbations to the network, the state would stay within its basin. This means that we could get to the correct concept no matter how high the noises or perturbations are.

In addition, if damages in attractor networks destroy positions of point attractors, the same stimuli fall in incorrect attractors due to transformations of size and shapes of basins. Therefore, it requires more time to fall in correct attractors than the normal attractor network does (Fig3).

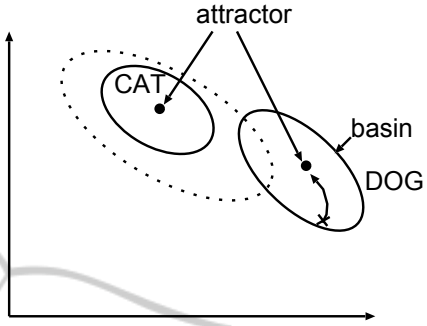

Figure 3: An attractor network originally proposed by Plaut and Shallice (1993).

2.1 Mathematical Notations

Each neuron, or unit, U_x has an output function $f(x)$, which is a sigmoid function, as follows,

$$U_x = f(x) = \frac{1}{1+e^{-ax}}. \tag{1}$$

Throughout the numerical experiments in this study, we fixed a constant $a = 4.0$. The units in the hidden layer (U_h) can be expressed as follows:

$$U_h = f\left(\sum_{i\in I} w_i U_i + \theta_h\right), \tag{2}$$

where, w_i means i–th connection weight, U_i means an output value of the i–th input unit, and θ_h means a threshold value in the unit h, the subscription I means the output values in the units of input layer.

A unit in the output layer (U_o) and a unit in the cleanup layer (U_c) are denoted as (3) and (4);

$$U_o = f\left(\sum_{i\in H} w_i U_i + \sum_{i\in C} w_i U_i + \theta_o\right) \tag{3}$$

$$U_c = f\left(\sum_{i\in O} + \theta_c\right) \tag{4}$$

where, θ_o and θ_c in the equations denote threshold values in the output and the cleanup layers respectively. The states in units both the output and the cleanup layers were updated repeatedly until the convergence criterion had been reached or until the maximum numbers of iterations ($\tau \leq 50$).

In the learning phase, we defined the mean square error as:

$$E = \frac{1}{2}\sum (u_i - t_i)^2. \tag{5}$$

where, t_i indicated an i–th teacher signal. Actual learning of connection weights of each unit can be obtained by partial differential as follows:

$$\Delta w = -\eta \frac{\partial E}{\partial w}, \tag{6}$$

where, η indicates a learning rate fixed as $\eta = 0.01$ throughout this study.

The initial values of w and θ were assigned in accordance with an uniform random value generator ($-0.1 \leq w, \theta \leq 0.1$).

2.2 An Application of Attractor Networks to Dyslexia

Plaut et al. showed that their attractor network could reproduce symptoms of deep dyslexia. According to their simulations, by means of the operation of semantic memory structure, they succeeded to account for the double–dissociation between concrete and abstract words (Plaut et al., 1995; Plaut, 2001). They constructed the semantic memory that the representations of concrete words have more micro features than those of abstract words. They postulated when the degree of the brain damages would be moderate, concrete words would show lighter deficits than abstract words. Further, if the degree of the brain damage would be severe, the concrete words would have more severe deficits than the abstract words.

In this study, we did not rely on the dichotomous taxonomy, such as animate–inanimate objects classification, but rather, we represented the data on the basis of the discriminability and correlation. Thus, we adopted attractor neural networks in order to account for more phenomena, such as confusion matrix among items, reaction times, categorization task. We then studied the validity of the data expressions as described above.

3 NUMERICAL EXPERIMENTS

3.1 The Data of Tyler et al., (2000)

Tyler et al. (Tyler et al., 2000) adopted the isomorphic mappings in order to train their networks. In other words, their networks had to learn the output pattern identical to the input patterns. In this condition, the network must acquire the reproduction of the input pattern. However, we can consider two other conditions (teacher signals in this case). One is that the target matrix (teacher signals) being the identity matrix, having 16 rows and 16 columns, all the diagonal elements being 1 and all the non–diagonal elements being 0. Another is that the matrix having 16 rows times 2 columns, where the elements of this matrix consisting (1, 0) when the item is an animate object, and (0, 1) when the item is an inanimate object. To summarize these three conditions;

Category Condition: the target matrix is a 16 rows $\times$ 2 columns matrix, where the targets to be learned are animate objects, the output vectors are (1, 0). Otherwise (inanimate objects) the output vectors are (0, 1).

Unitary Condition: the target matrix is a 16 rows $\times$ 24 columns matrix. This target matrix is identical to the matrix of the input signals.

Identical Condition: the target matrix is an unitary matrix of 16 rows $\times$ 16 columns, where diagonal elements are 1 and other elements in this matrix are 0.

In the unitary condition, the network is required to learn the precise knowledge of each member in the input patterns. In category condition, the network is required to learn the higher concepts of both animate and inanimate objects, as Tyler et al. (Tyler et al., 2000) suggested. In the homo condition, the unitary matrix means that each item can play a roll to form the identical matrix.

3.1.1 Network Architecture

We adopted the number of units in the hidden layer to be 10, and the number of units in the cleanup layer to be 1. The reason for determining the number of units in the cleanup layer to be 1 is based on the preliminary experiment.

3.1.2 Procedure

We decided the maximum iteration numbers between the output and the cleanup layers to be 50 for each item. If the error of this attractor network did not reach the convergence criteria, defined by the sum of squared errors being less than 0.05 for each item. Within the maximum number of iterations between the output and the cleanup layer, the program gave up to let the networks learn this item, and was given the next item to be learned. The order of the items to be learned was randomized within each epoch. This procedure was repeated until the network learned all the items.

The convergence criteria was set that all the sum of squared errors are below 0.05 throughout in this study. The network was given the input signals and teacher signals at a time to learn the output patterns.

At first, the output values were calculated from the input patterns to the units in the output layer. Then iterations between the output and the cleanup layers started until the output values have reached the criteria, or the iteration numbers have been exceeded 50 times.

3.1.3 Results

We investigated the mean iteration numbers between the output and the cleanup layers. These numbers indicate the times that the initial value is absorbed in an attractor when the initial value was located within a basin of an attractor (Fig. 3).

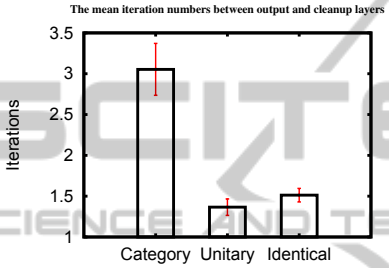

Figure 4: The mean mutual iteration numbers (n=20) between the output and the cleanup layers, when we train the attractor network by the data of Tyler et al., (2000).

In the category condition, the iteration numbers were nearly equal to 3. The attractor network may be able to utilize the connections between the output and the cleanup layers in the category condition. On the other hand, the mean iteration numbers both the identical and the unitary conditions were from 1.0 to 1.5. These results might indicate that the attractor network did not always utilize the connections between the output and the cleanup layers, when these two conditions were given. Therefore the attractor network might be considered that it behaved as a three–layered–perceptron. The attractor networks include three–layered–perceptions in the special case.

3.2 The Triangle Model

3.2.1 Procedure and Conditions

The triangle model of word reading (Sidenberg and McClelland, 1989; Plaut et al., 1996) was developed to mimic human performance of word reading. The triangle model includes phonology, orthography, and semantics (Fig.5). The attractor neural network model can be applied to this model. One arrow in Fig.5, for example, from orthography to semantics, can be regarded as an attractor network. Because there are 6 arrows in Fig.5, we can apply 6 attractor networks to the

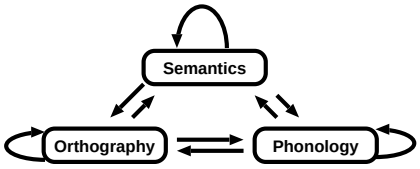

Figure 5: A schematic drawing of the triangle model proposed by Seidenberg and McClelland, 1989; Plaut et al., 1996.

triangle model. In addition to these 6 arrows, we can add 3 self recurrent arrows, from orthography to orthography, from semantics to semantics, from phonology to phonology. These self recurrent arrows can be also regarded as attractor networks. Furthermore, we can add 2 more mappings from 3 elements, orthography, semantics, and phonology, to concept and identical conditions (see the section of Tyler et al.). Therefore, we can obtain 15 conditions in total. We

Figure 6: Two hypothesized concepts of semantics Plaut & Shallice (1993). The upper left part and the lower right part of the matrix, similar to Fig.1.

adopted the attractor neural networks to all 15 conditions respectively. We set the number of units in the cleanup layer to 1, and set the number of units in the hidden layer to 50 for every condition.

3.2.2 Results

As we predict from the data of Tyler et al., (2000), attractor networks could solve all the data of 15 conditions. Table 1 shows the examples of iteration times for convergence.

4 DISCUSSION

If we could consider the attractor network as a concept formation model of human brains, then the identical condition could be regarded as the model to rec-

Table 1: The examples of iteration times for convergence.

	from		
to	Orthography	Phonology	Semantics
Category	1547.1	1965.3	48.4
Unitary	302.3	171.8	228.1
Orthography	78.2	359.8	378.20
Phonology	966.4	136.1	294.1
Semantics	577.9	288.5	74.4

ognize the shape of a dog exposured in one's retina as 'dog'. The category condition might be considered that subjects and patients could recognize this visual image of a dog as an animal. We could interpret the unitary condition as that subjects and patients would have been recognized a 'dog' per se. In this way, we could interpret the three conditions we adopted in this paper as the human models of recognition described here. The category condition could be considered as the one which utilized the loop between the output and the cleanup layers the best among three conditions. Thus, it seems that the category condition might be the model of category judgement. In addition to this, we observed one of the category specific disorders in the destruction experiment which destroyed the mutual connections between the output and the cleanup layers. This results should not be considered as accidental artifacts of the computer simulations.

Attractor networks could be applied to the triangle models of word reading as well. Although the original triangle model (Sidenberg and McClelland, 1989) has self recurrent arrows, each arrow could not be regarded as an attractor network. This study tried to interpret that the triangle model would be consisted of all 15 attractor networks in total.

5 CONCLUSIONS

In spite of the simplicity, the attractor neural network could describe several symptoms of neuropsychology. This is one of major advantages of this model. The possibility to explain the double–dissociation between animate and inanimate objects should be discussed further in separate papers. However, there still are possibilities for this model to account for the double–dissociation between animate and inanimate objects. The difference between intra– and inter– correlations shown in fig. 1 might cause the category specificity, because one category has higher inner–category correlations than that of the other category. In this study, we adopted non–dichotomous memory representations whose correlation matrix of microfeatures such as Fig.1 and Fig.6. These representations can be considered as one of the major advantages of the model.This kind of object representations might emerge as one aspect of category specific memory disorders.

REFERENCES

Caramazza, A. and Shelton, J. (1998). Domain specific knowledge system in the brain: The animate–inanimate distinction. *Journal of Cognitive Neuroscience*, 10(1):1–34.

Devlin, J., Gonnerman, L., Andersen, E., and Seidenberg, M. (1998). Category-specific semantic deficits in focal and widespred brain damage: A computational account. *journal of cognitive neuroscience*, 10(1):77–94.

Farah, M. J. and McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, 120(4):339–357.

Plaut, D. C. (2001). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. In Chirstiansen, M. H. and Charter, N., editors, *Connectionist Psycholinguistics*, chapter 8, pages 244–278. Ablex Publishing, Westport, CT.

Plaut, D. C., McClelland, J. L., and Seidenberg, M. S. (1995). Reading exception words adn pseudowords: Are two routes really necessary? In Levy, J. P., Bairaktaris, D., Bullinaria, J. A., and Cairns, P., editors, *Connectionist Models of Memory and Language*, pages 145–159. University College London Press, London.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., and Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103:56–115.

Sidenberg, M. and McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96:523–568.

Tyler, L., Moss, H., Durrant-Peatfield, M., and Levy, J. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, 75:195–231.

Warrington, E. and McCarthy, R. (1983). Category specific access dysphasia. *Brain*, 106:859–878.

Warrington, E. and McCarthy, R. (1994). Multiple meaning systems in the brain: A case for visual semantics. *Neuropsychologica*, 32:1465–1473.

Warrington, E. and Shallice, T. (1984a). Category-specific semantic impairment. *Brain*, 107:829–854.

Warrington, E. K. (1981). Neuropsychological studies of verbal semantic systems. *Phil. Trans. R. Soc. Lond. B*, 295:411–423.

Warrington, E. K. and McCarthy, R. (1987). Categories of knowledge further fracitonations and an attempted integration. *Brain*, 110:1273–1296.

Warrington, E. K. and Shallice, T. (1984b). Category specific semantic impairment. *Brain*, 107:829–854.