Hysteresis Parameter Settings and the SRN Learning Curve

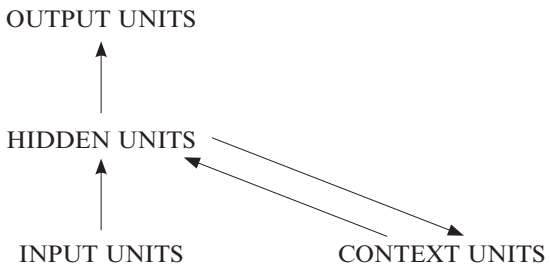
Isao Inoue

1. Introduction

In artificial grammar learning paradigm, Onnis, Christiansen, Chater, and Gómez (2003), and Onnis, Monaghan, Christiansen, and Chater (2004) have investigated how variability of the middle element *X* in the three-element string *AXB* affects the learnability of nonadjacent dependency relation between the elements *A* and *B*. It is found that high and zero variability of *X* get adults and infants to detect nonadjacent dependencies but the low level variability of *X* yields poor acquisition results. Thus, when subjects in these experiments are trained under high and zero variability condition, they can successfully discriminate between three-element strings involving correct *AB* pairings (A_1__B_1, A_2__B_2, A_3__B_3) and those with incorrect pairings (e. g. A_1__B_2, A_2__B_1, A_3__B_1), but no such successful performance is exhibited in the case of the low level variability.

This paper is an attempt to simulate the human performance observed by Onnis et al. (2003), and Onnis et al. (2004) by means of the Simple Recurrent Neural Network (SRN) (Elman (1990)). The SRN used in all our simulations has the following network architecture involving 6 input/output units and 30 hidden units which are recurrently connected to 30 context units (see Figure 1).

Figure 1.



Context units function as dynamic memory, copying the hidden unit activation pattern at a particular time step *t-1* and feeding it back to the hidden units at the next time step *t*. Each context unit receives its activation value from the corresponding hidden unit on a one-to-one basis, which is specified by the downward one-to-one connections from the hidden units to the context units. The upward connection from the context units to the hidden units is set up according to a standard full connectivity pattern, each hidden unit receiving connections from every context unit.

In our experiments, we used the SRN equipped with the hysteresis function, which is implemented in PDP++ software package (Dawson, C. K., R. C. O'Reilly, and J. L. McClelland (2003)). By changing the hysteresis parameter (hyst) in the following equation, we can control the updating of the context units:

(1)　$\mathrm{ContextUnit}_j(t) = (1\text{-hyst})*\mathrm{HiddenUnit}_j(t-1)+\mathrm{hyst}*\mathrm{ContextUnit}_j(t-1)$

The hysteresis parameter specifies, in effect, how much context units reflect their previous activation state. A large hysteresis value makes wide view of context available for processing sequential events, and a smaller value should be selected to capture short distance dependency relations. We examined whether or not the SRN incorporating the hysteresis function (1) could successfully simulate the human performance in question by trying different hysteresis settings. Five hysteresis values (0.1, 0.3, 0.5, 0.7, 0.9) were selected to cover a possible range of 0 to 1, and ten simulation runs from different random initial conditions were carried out under each hysteresis parameter setting.

2. Learning Non-Adjacent Relations

In all our simulations, the identical weight updating parameters were used: learning rate 0.25; momentum 0.9. Since the present computer simulations were conducted to replicate human learning performance reported in Onnis et al. (2004), the three set-sizes of *X* (1, 2, 24) adopted in Onnis et al. (2004) were employed as zero, low, and high variability condition in the context of three nonadjacent pairings A_1__B_1, A_2__B_2, A_3__B_3. Each element in *AXB* strings was represented as a randomly selected distributed pattern of activation across a 6-bit vector. The order of presentation of *AXB* strings to the network was randomized, and each string was presented once per epoch. In the case of set size 24, 72 *AXB* strings (3 *AB* nonadjacent pairings involving 24 different *X* elements) constituted one epoch, and this epoch process was iterated 200 times in the training process. In the case of set size 2, 6 *AXB* strings (3 *AB* nonadjacent pairings involving 2 different *X* elements) formed one epoch and

this epoch process was iterated 2400 times, and 4800 iterations were carried out for set size 1. With this training scheme, the frequency of occurrence of nonadjacent *AB* pairings was held constant across the three variability conditions.

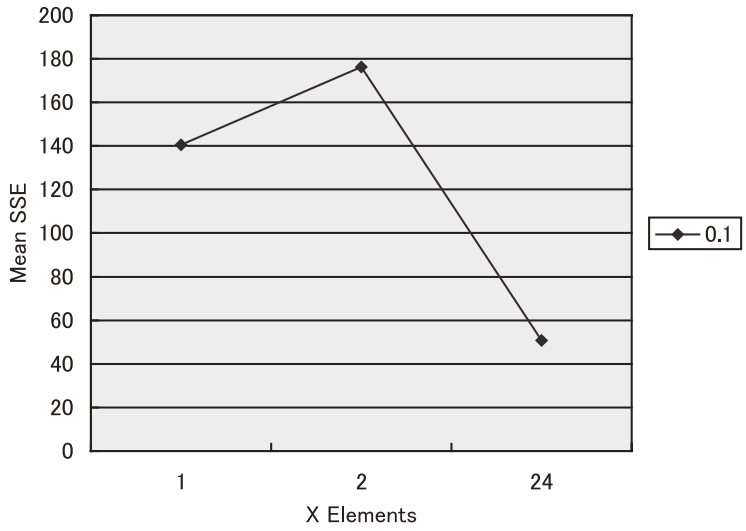
After training SRNs, they were tested on *AXB* strings involving novel *X* elements, not used in the training phase. The *AXB* strings used in the testing process were formed by inserting 24 new *X* elements into the following three non-adjacent *AB* pair frames, A_1__B_1, A_2__B_2, A_3__B_3, totaling 72 strings. In the testing process, we checked whether or not the trained network could make a correct prediction about the *B* element corresponding to the paired *A* element when the network received an *AX* sequence in this order. We evaluated the network performance by calculating the summed squared error (SSE), which is the summed difference between the actual output activation (*o*) and the target value (*t*) of six output units:

(2) $SSE = \sum(t - o)^2$

3. The Result of Simulations

For each hysteresis parameter setting (0.1, 0.3, 0.5, 0.7, 0.9), ten simulation runs from different random initial conditions were carried out under three variability conditions (namely, 1, 2, 24 *X* elements), and only under the hysteresis parameter setting of 0.1, the successful simulation result was obtained:

Figure 2. SRN performance under the hysteresis setting of 0.1



A steep rise in SSE was observed when the *X* variability factor only just changed from 1 to 2, leading to the highest SSE error rate at *X* elements 2, as shown in Figure 2. This unique performance pattern agrees closely with the crucial human behavioral characteristic observed in Onnis et al. (2004).

But no such result was gained under other hysteresis parameter settings:

Figure 3.　SRN performance under the hysteresis settings of 0.3, 0.5, 0.7, and 0.9

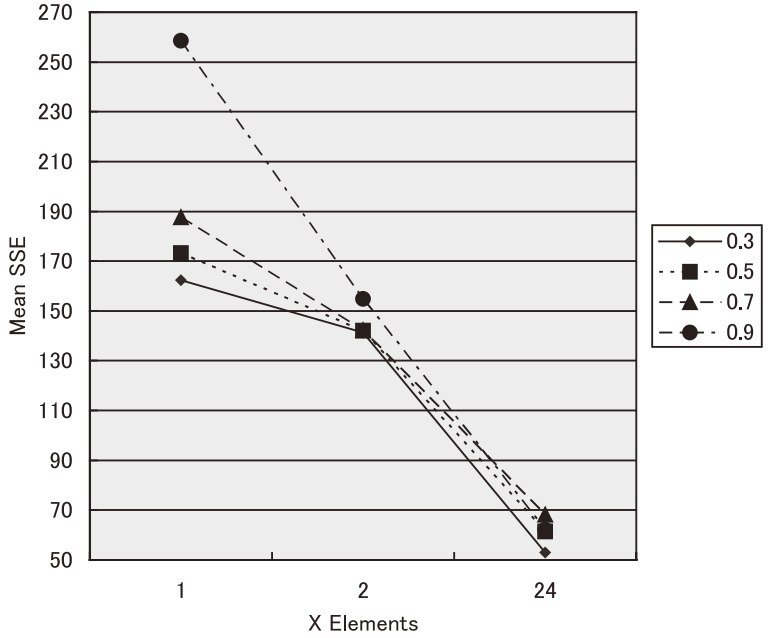


Figure 3 shows that the highest SSE error rate always occurs under the zero variability condition (namely, 1 *X* element), which contradicts the human experimental data reported in Onnis et al. (2004).

The successful SRN performance shown in Figure 2 was evaluated by means of ANOVA. The independent variable included the three levels of *X* variability factor (1 vs. 2 vs. 24) and the dependent variable was SSE. The means and standard deviations for SSE scores are shown in Table 1:

Table 1.

X variability	Mean SSE	Std. Deviation	N
1	140.6	25.4	10
2	176.3	9.3	10
24	50.6	1.8	10

The results of the ANOVA indicated a significant *X* variability effect, $F(2, 27) = 171.1$, $p < 0.001$. We conducted three pairwise comparisons among the means, using the Games-Howell procedure, because the assumption of homogeneity of variance was violated (Levene's Test, $F(2, 27) = 26.8$, $p < 0.001$). There were significant differences among the three variability conditions and a polynomial trend analysis showed a significant quadratic effect, similar to human experimental data.

Table 2.

(I) Variability	(J) Variability	Mean Difference (I-J)	Significance
1	2	–35.7	.004
	24	89.9	.000
2	24	125.7	.000

Although the overall trend of SRN performance successfully mimicked the human behavioral pattern, the SRN failed to attain the same level of accuracy that human learners have achieved. The biggest difference between SRN and human learners arose in the case of zero variability. The SRN performance under zero variability was relatively poor (roughly 68% correct) compared to the very high score (reaching almost 90% level) achieved by human learners. A follow-up study in search for appropriate hysteresis value found that fine-tuning centered around the hysteresis value of 0.1 did not bring about a sufficient improvement in accuracy. Thus, the simulations reported in this paper suggests that additional mechanisms should be incorporated into the SRN system to produce desirable simulation results.

References

Dawson, C. K., R. C. O'Reilly, and J. L. McClelland (2003) *The PDP++ Software Users Manual*, Carnegie Mellon University.

Elman, J. L. (1990) "Finding Structure in Time," *Cognitive Science* 14, 179–211.

Gómez, R. L. (2002) "Variability and Detection of Invariant Structure," *Psychological Science* 13, 431–436.

Gómez, R. L. (in press) "Dynamically Guided Learning," in Y. Munakata and M. Johnson (eds.) *Attentional and Performance XXI: Processes of Change in Brain and Cognitive Development*, Oxford, Oxford University Press.

Gómez, R. L. and J. Maye (2005) "The Developmental Trajectory of Nonadjacent Dependency Learning," *Infancy* 7(2), 183–206.

Newport, E. L. and R. N. Aslin (2004) "Learning at a Distance I. Statistical Learning of Non-Adjacent Dependencies," *Cognitive Psychology*, 48, 127–162.

Onnis, L., M. H. Christiansen, N. Chater, and R. Gómez (2003) “Reduction of Uncertainty in Human Sequential Learning: Evidence from Artificial Language Learning,” *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, pp. 886–891, Lawrence Erlbaum, Mahwah, New Jersey.

Onnis, L., P. Monaghan, M. H. Christiansen, and N. Chater (2004) “Variability Is the Spice of Learning, and a Crucial Ingredient for Detecting and Generalizing in Nonadjacent Dependencies,” *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, Lawrence Erlbaum, Mahwah, New Jersey.

O’Reilly, R. C. and Y. Munakata (2000) *Computational Explorations in Cognitive Neuroscience*, Cambridge, Massachusetts, The MIT Press.