Mechanisms of Temporal Integration and Knowledge Representation with Simple Recurrent Networks

Robert B. Allen
Bellcore, Morristown NJ
rba@bellcore.com

Knowledge representation was studied for simple recurrent networks which were trained to respond to the semantics of entire statements. In the first set of studies, networks were trained to recall category membership for objects. The representations showed clear differentiation by category. Moreover, simple inheritance of features was demonstrated. In other studies, networks were trained to respond to kinship relationships each of which was presented sequentially. The networks were shown to recognize indirect descriptions of people. An analysis of the hidden units showed high correlations for related questions. Moreover, analysis of individual bits showed that some seem to detect features.

Knowledge Representation with Neural Networks

While Simple Recurrent Networks (SRNs) have been shown to learn several interesting temporal integration tasks, including tasks that are relevant to natural language processing, little is known about how they perform those tasks. Although some previous work has explored the representations of SRNs, it is unclear what part, if any, the semantics of the entire statement played in these effects. Specifically, Elman (1990) has shown that networks predicted successive terms in pseudo-sentences, and terms which were likely to occur together were clustered together. In addition, Harris and Elman (1989) studied trajectories in principle components space for this task. Furthermore the predictions for these tasks are of a low level, and it seems likely that the computational complexity of the predictions would swamp any practical use of the higher-level effects. On the other hand, it seems possible that the human language system incorporates a variety of sources of feedback, including successive words, and it is possible that some combination of the different types of feedback work together in human language performance.

In this research, a variety of simple tasks were presented to modified SRNs and the activations of the hidden units were examined. In all of the studies, error correction was made for the entire statement. The activations were examined directly, without possibly obfuscating high-level statistics.

Network Architecture

Figure 1 illustrates the network employed in most of the following research. This network is different from the basic SRNs in a number of ways. First, a memory term has been added to the state units. Second, the entire question and one additional cycle with null inputs are presented (see also Sect. 9 of Allen 1990) prior to any error correction. Furthermore, as shown in Figure 1, the network is connected with fixed one-to-one weights between the inputs and hidden units. These weights were fixed because error correction occurred only after all of the input patterns had been presented and the inputs were "off". Parametric pilot studies showed that fixed weights of 6 were most effective for "on" inputs and -1 for "off" inputs. The network parameters were: momentum, α=0.9; learning rate, η=0.01; hidden-state copy, β=1.0; and state memory, μ=0.5.

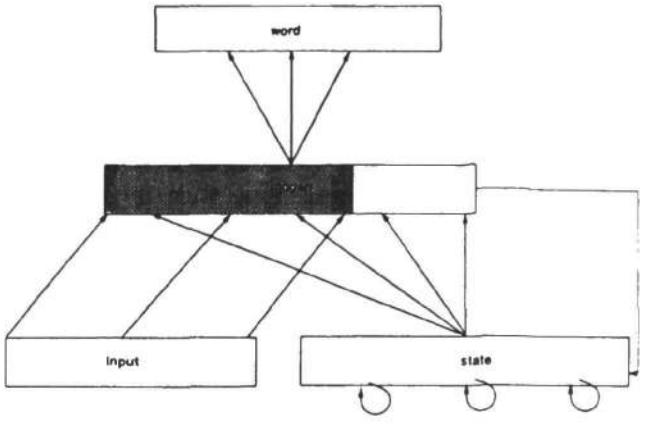

Figure 1. Basic Network Architecture

Simple Categorization and Inheritance

One area in which a hierarchical structure might be expected to develop in the hidden unit activations is with categorization of objects. Thus, a corpus was developed of true/false questions with 4 categories identified and 3 "objects" assigned to each. Moreover, each category was associated with a single "feature". Altogether there were 270 sentences in the corpus. The four of these describing the relationship between **object 1** and **category 0** were withheld and used for the transfer test. Networks with 25 input units, 40 hidden and state units, and 2 output units, 25-40*-2, learned this task without error after 325K pattern presentations.

The networks correctly answered the transfer questions and thus appeared to show limited 'inheritance'. The reason for this effect, as shown in Table 1, is that similar patterns of weights were generated by the object and category questions. While this result seems simplistic it has some appeal, in that the effects of category names, prototypical objects, both types of questions produce similar transformations of the state vectors. On the other hand, other simple operations are not supported by this

model. For instance, networks were not able to transfer when training sentences linking the categories to features were withheld. The networks showed 'inheritance' but not abstraction. The addition of a small amount of jitter (e.g., 0.0000001 * *net input*) to the hidden unit activations reduced the number of cycles to reach low error (about 0.05) by about 10%. As a test of the importance of the memory term, μ, on the state units, a network was run without a memory term (Elman, 1990) and no learning was observed.

Table 1. Roughly Orthogonal State-Unit Weights

	category		object	
	bit 28	bit 29	bit 28	bit 29
0	-36	47	-27	37
1	-16	-20	6	-14
2	16	-41	11	-27
3	11	33	19	38

Scaling Features in Categories

SRNs have been applied only to toy problems. The issue of whether the technique will scale to large stimulus sets is crucial for practical applications. As a problem gets larger we would hope find that each component of the problem gets easier. Thus, the function of the number of presentations to convergence for large problems should be sublinear. The categorization task above may be scaled by increasing the number of objects per category and giving the network a surplus of hidden units. Figure 2 shows the scaling is S-shaped for 10 averaged runs with from 2 to 6 objects in each of 5 categories.

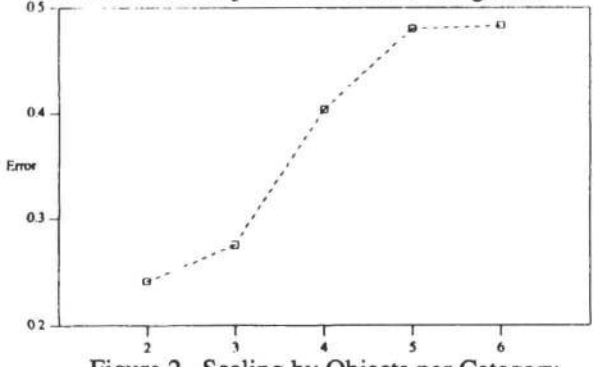

Figure 2. Scaling by Objects per Category

Kinship

The representation of kinship relationships can be a highly structured knowledge-representation task. Back-propagation training has been applied to the learning of kinship relationships (Hinton 1986) but only with simple three-term relationships which had no temporal processing of the inputs. Nonetheless, Hinton demonstrated that features relevant to the kinship relationships were observed in the hidden layer, at least for the highly constrained architecture he employed. In comparison to Hinton's approach, SRNs have the potential for arbitrary length descriptions. Moreover, the SRNs may be trained to answer questions about kinship or to respond to the truth of the assertions.

Procedure

Questions were developed based on the kinship relationships shown in Figure 3 which describe 12 people in one family. In the corpus, questions involving a second family, with an identical structure, were also included. The total lexicon was 40 input terms and 27 output terms. Some of the terms were compressed phrases; for instance, the phrase "the father of" was compressed to "fatherof". The kinship relations were "motherof", "fatherof", "sonof", "daughterof", "sisterof", "brotherof", "wifeof", and "husbandof".

Table 2. Examples of Training Statements

type	number	example question	answer
whois	80	whois motherof mA2	fA0
kinrel	40	kinrel fB2 fB3	motherof
agerel	46	agerel fA3 fA2	youngerthan
gender	24	gender mA5	male
true	205	daughterof fA2 fA3	true
false	205	wifeof mA5 fA5	false

The corpus included both questions and true or false statements. As shown in Table 2, four types of base questions were employed. This resulted in 190 base questions for each of the two families. Moreover, individuals could be specified indirectly. For instance, "mA2" (i.e., male 2 in family A) could also be described as "fatherof fA3". Thus, the questions "kinrel motherof fA3 mA2" and "kinrel fA2 fatherof fA3" are equivalent to asking for the kinship relationship between fA2 and mA2 , and the answer is that they are married. The expanded corpus included all possible indirect descriptions of the individuals and had a total of 4391 questions. Presumably this is because the feedback from the true/false responses was inadequate to develop the relatively complex discriminations required.

A 40-80*-38 network (40 locally encoded input units, 80 hidden units, 80 state units, and 38 locally encoded output units) was trained for 650K patterns on the basic corpus (without variations). Indirect specifications for the people were introduced and the network was trained for an additional 1250K patterns.

Results

The values of hidden units were inspected during the responses to individual questions. Because the procedure in which hidden-unit values were copied in a one-to-one correspondence with individual state units, a large positive weight between the state unit and the corresponding hidden unit effectively latched those units. The network had 32.0% such weights (37

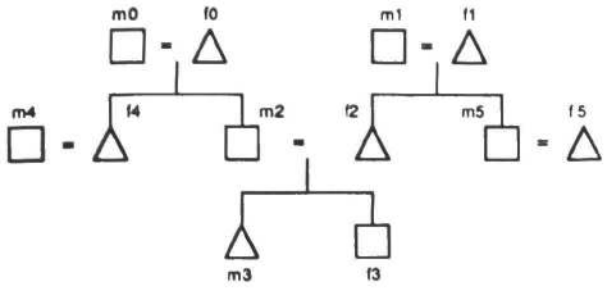

Figure 3. Template for Kinship Relationships

out of a possible 116) which were above an arbitrary threshold (2.5) while only 1.3% of the other weights (79 of 6320) exceeded that threshold.

It is also of interest to compare the representations across related questions. The Pearson *r* correlation statistic was used as a measure of the similarity of sets of hidden units. Correlations were computed between hidden unit vectors at the point when the response was made. Correlations for a typical question are shown in Table 3. It can be seen that many, but not all of the most highly correlated sentences are semantically related to the target.

Table 3. Examples of Question Correlations with Targets

correlation	question	answer
target	mA2 husbandof fA2	true
0.9752*	sonof mA0 husbandof fA2	true
0.9652	mA4 husbandof fA4	true
0.9508	husbandof fA2 mA2	true
0.9432	sonof husbandof fB1 mB5	true
0.9408	sonof mB1 husbandof fB5	true
0.9365	daughterof fB1 sisterof mB5	true
0.9306	mB4 husbandof daughterof mB0	true
0.9264	daughterof fB1 daughterof fB1	true
0.9198	daughterof fB1 daughterof mB1	true
0.9184*	mA2 husbandof daughterof fA1	true

The starred correlations are for questions which are identical in meaning to the target question. While the starred statements suggest that there is some similarity between the target and its variations, there are many variations and it is worth compiling summary statistics. The mean rank across all variations of the target statement is shown in the center of Table 4. In addition, the corpus contains other statements that are essentially identical in meaning with the target sentence, and the means of the variations of those sentences are also shown in the center column of Table 4. Because the total number of true statements is 1576 it can be seen that the statement forms are above the mean rank. Because the SRN successively compounds the input activations, it might be speculated that the similarity results were not due to training, but only to compounding of the input patterns. Thus, the weights of the network were scrambled within layers and the mean rank of sentence variations are shown on the right side of Table 4. The ranks are generally lower; hence there is more to the similarity results than simply compounding of initial inputs. However, the effect is not overwhelming, especially for the last two statements, which begin with the female rather than the male.

Table 4. Ranks of Variations of Question Set

question template	trained rank	scrambled rank
mA2 husband fA2	191.7	460.3
husband fA2 mA2	160.8	539.3
fA2 wifeof mA2	731.6	1065.9
wifeof mA2 fA2	592.0	357.7

Rather than looking at an entire statement, it is possible to correlate activations for partial statements. The activations for all person descriptions were obtained. The hidden unit vector correlations for mA0 were highest with other members of the same generation. and smallest with males of lower generations and females in the other family.

Figure 4 shows the hidden unit activations for a simplified stimulus. In the upper section of the figure are activations averaged across descriptions of males and females in families A and B. The lower portion of the figure shows activations for individual relationship terms (e.g., fatherof). Examination of the figure suggests that bit 26 (marked with a tick) codes for gender. Moreover, the relationship operators are differentiated in the same positions. Thus, the statement 'fatherof mA2' would set the generation bit in 'mA2' to produce a representation similar to 'mA1'. Beyond individual terms, it is possible to consider how the truth of an entire statement is determined.

A simplified model of what the network is doing to process these statements is summation of activations. If the same person is being described as both the left and right sides of the statements and the activations sum, then the veracity of the statement can be determined by checking the threshold of the activations. However, that is not an adequate model since the network is able to take temporal order into account and process statements with the same words in different ways. For instance, "mA2 husbandof fA2" is responded to as "true" while "fA2 husbandof mA2" is correctly responded to as "false".

Convergence of Kinship Representations

A direct test was devised of whether associations of the hidden unit activations reflect similarity in

semantics of the input. Specifically, two input units were given exactly the same meaning and the representations generated by those units were compared. The procedure was similar to the previous experiment except that a simpler family network was employed (with only 8 members) and only the base questions were used. Hence, the network required 33 input units (32 units for the family relationships and 1 extra for the redundant input). Duplicate "kinrel" questions were generated. The networks were trained for 600K pattern presentations. Across five runs, the average correlation of the hidden units for the two versions of kinrel was 0.56. As a control, the correlations of the activations in five untrained networks were obtained; the average correlations had a mean of 0.15. Clearly, the training of the network has caused the activations for the semantically related terms to become more similar.

Implications and Challenges

SRNs were shown to be able to generate responses for highly structured knowledge representation tasks. The development of bits which code for specific features was striking. Moreover, there is considerable other evidence that the representations reflected the semantics of the questions and statements. In addition, the results show clearly the state-like nature of recurrent network processing. The results also have implications for models of the semantic space. High correlations for semantically related statements may imply that surface features are lost during processing.

Bit Overlays

Similar mechanisms of simple bit operations were found in both the categorization and kinship studies. While the representations identified in this work were not always clear, the most common mechanism appeared to be operations on bit vectors. The process is unlike the usual slots of frame-based approaches to representation; however it is consistent with many psychological models. For instance, the vectors shown in Table 1 are reminiscent of psychological spaces derived from multivariate statistical methods.

Analogy

The view of cognitive processing as feature bit activation and masking is consistent with models of analogy (see Rumelhart and Abramson, 1973). Thus an analogy may be seen as applying similar operators to similar, though meaningfully different, vectors.

Information Retrieval

Information retrieval (IR) has received only cursory attention with neural networks (Mozer, 1984). However, it might be expected that networks would be effective for IR because many of those procedures are statistically based, and there are similarities between several of these IR models and neural models. Moreover, there is a variety of ways in which neural networks may be applied to IR. For instance, judgments about the relevance of a document may be used as a source of feedback for error correction of a neural network. However, it is difficult to obtain enough relevance judgments to make this practical. Few of the existing statistical IR techniques take advantage of the temporal integration of terms and SRNs have that potential. Indeed, since SRNs have been shown to be able to handle simple grammars (Allen 1990; Servan-Schreiber, et al., 1989), it is possible that they could process some aspects of

Figure 4. Hidden Unit Activations

Hidden Unit Vectors By Gender with Operators.

```
MA  0113020000226030000100101210510010000000210030011103001001000100000000271071002 6
MB  0112110000226120000000010110410020000000320020001004011101000100110000162072101 6
FA  0104010000166030000200011010200010000010100020001004011000000100200000181080003 6
FB  0202010000226020000100021010400030000000210020000005012100000100400000172080103 7
                                  |                                                   |
fa  0000010000225020000000002610400041000000040040010004000000001000200000063081000 1
mo  0003010000107020000400026110510320000030130040253102081000000100800000392080100 5
br  0100030000148130000000100610000221000020040010000002030500001100000000083082303 6
si  1212010000207020000001001210000100000000510200010050060000000010000009208010 37
hu  0119030000237020000000017001101300000100700200110140200030001000000004810821026
wi  0102010000136030000100011000400020000000100030000004001000000100000000182080102 6
so  0001000000267030000000020010000021000010250030000103000200000100910000353081200 5
da  0111000000387020000001110040020010000000373020001003004000000100000000302070001 7
```

Hidden Unit Vectors By Age with Operators.

```
oA  0004000000137020000100000300300000000001030040000004020000000000100010142070000 2
oB  0003000000118020001000000200300000000002120030000004010000000000100020042070001 2
mA  0003000000157020000000001000000000000001200300000060111000000000000000310700003
mB  0002000000137010000000000100100010000000130020000006012100000000200010022070000 3
yA  0002000010168020000000000000000000000000130020000006002100000000100010031070000 5
yB  0000000000127010001000000110001000200000240010001007012400000000210000012080000 5
                                  |                   |                             |
fa  0000010000225020000000002610400041000000040040010004000000001000200000063081000 1
mo  0003010000107020000400026110510320000030130040253102081000000100800000392080100 5
br  0100030000148130000000100610000221000020040010000002030500001100000000083082303 6
si  1212010000207020000001001210000100000000510200010050060000000010000009208010 37
hu  0119030000237020000000017001101300000100700200110140200030001000000004810821026
wi  0102010000136030000100011000400020000000100030000004001000000100000000182080102 6
so  0001000000267030000000020010000021000010250030000103000200000100910000353081200 5
da  0111000000387020000001110040020010000000373020001003004000000100000000302070001 7
```

natural language syntax. In addition, SRNs also seem well suited for IR because of their ability to model users (Allen 1990).

A compact and efficient knowledge representation could be useful for IR. Thus, it is possible to imagine that a representation for a query could be developed and this could be matched with representations for statements from a document. If SRNs are trained to make responses based on the meaning, they may have that potential (Allen 1990). For the generation of responses to questions, (as is the case with the studies reported in this paper) the activations would tend to converge. Declarative statements would simply generate activations without resolution; thus at the end of a statement or document, the residual feature activations might reflect the overall meaning of the document.

Unfortunately, the results in these studies do not provide much encouragement. For instance in the kinship studies, not all of the semantically related statements showed high correlations with the targets. Of course, it would be desirable to find ways to improve the quality of the performance. Clearly, for a practical applications, such as IR, a much larger lexicon would be required and in those, facts will be sparser than in the training sets used here. On one hand that would require greater training time; however, it might also make retrieval much easier since there would be fewer easily confused cases. Of course, there is still the problem of getting a corpus of statements which had an appropriate level of feedback (e.g., true/false) for the type of training employed here.

Massive Knowledge Bases

If clear-cut representations are not able to be identified, it may be more useful to think of developing massive question-answering systems (Allen, 1990). Of course, there are also substantial difficulties in that direction. For instance, better techniques are needed for integrating sequential output with sequential inputs. Another problem is finding a corpus with sufficient feedback for training.

Architectural and Procedural Issues

The studies here demonstrate the advantage of the modifications to the standard SRNs. Indeed SRNs without state memories were not able to complete the categorization task. Furthermore, additional variations of the basic network proved advantageous in some cases. In addition, small amounts of noise apparently helped the network in the categorization task avoid temporal local minima.

Extensions

Previous work and the present results suggest that there is a great deal that is right with this model for language processing. For instance, representations were found which were similar to those proposed for other models of psychological space; scaling, at least in one case, was found to be sublinear; limited inheritance was demonstrated; and feature-detecting bits were identified in the representations. However there are also substantial limitations, the representations often seemed contorted and the processing mechanisms seemed too simple for complex language. There are two directions to consider for moving to practical models of language processing with neural networks. On one hand, we could change expectations as to what is required for a language processing system and attempt to get networks to learn with fairly simple modifications to the existing mechanisms. On the other hand the network mechanisms may be embellished. For instance ways might be sought to integrate them with perceptual processes (see Allen 1990), or to train them to develop more complex structures such as stacks and variable bindings. The present results suggest that the latter strategy will probably be more successful that the former.

References

Allen, R.B. 1990. Connectionist language users. *Connection Science*, 2, 279-311.

Elman, J.L. 1990. Finding structure in time. *Cognitive Science, 14*, 179-211.

Harris, C. and Elman, J. 1989. Representing variable information with simple recurrent networks. *Proceedings of the Cognitive Science Society*, Ann Arbor, 635-642.

Hinton, G. 1986. Learning distributed representations of concepts. *Proceedings of the Cognitive Science Society*, Amherst, MA. 1-12.

Mozer, M.C. 1984. *Inductive information retrieval using parallel distributed computation.* ICS Technical Report 8406, La Jolla, UCSD.

Rumelhart, D.E. and Abramson, A.A. 1973, A model for analogical reasoning. *Cognitive Psychology, 5*, 1-28.

Servan-Schreiber D., Cleeremans, A., and McClelland, J. 1989. Learning sequential structure in simple recurrent networks. In: D. Touretzky (ed.) *Advances in Neural Information Processing Systems, 1*, Morgan-Kaufmann, San Mateo, CA, 643-652.