Boundary effects in the linguistic representations of simple recurrent networks

Ronan Reilly
Department of Computer Science
University College Dublin
Belfield, Dublin 4, Ireland
rreilly@ccvax.ucd.ie

Abstract

This paper describes a number of simulations which show that SRN representations exhibit interactions between memory and sentence and clause boundaries reminiscent of effects described in the early psycholinguistic literature (Jarvella, 1971; Caplan, 1972). Moreover, these effects can be accounted for by the intrinsic properties of SRN representations without the need to invoke external memory mechanisms, as has conventionally been done.

Introduction

Several well attested phenomena in psycholinguistics relate to the interaction of short-term memory and sentence and clause boundaries. These results come from some of the earliest research in modern psycholinguistics. For example, Jarvella (1971) presented listeners with short stories which were interrupted at various points. The listener's task was to recall verbatim as much as possible of the preceding material. The results showed that recall was best for words in the clause immediately preceding the interruption, and that it dropped off markedly for words prior to the clause boundary. In a related study, Caplan (1972) looked at the time taken for subjects to judge whether or not a probe word had been present in a two-clause sentence which they had just heard. If the probe word had occurred prior to the clause boundary, reaction times were slowed significantly.

The conventional explanation for these effects is that "the completion of a clause is the condition under which lexical material is transferred from the most accessible memory system to one that is less accessible " (Fodor, Bever, & Garrett, 1974, p. 344). A similar explanation has been articulated more recently by Garnham (1985), who proposed that information is "transferred into a more permanent memory store at clause boundaries, and detailed representation of surface form is erased from immediate memory" (Garnham, 1985, p. 204). The goal of the study described here is to discover if these phenomena can be accounted for within a connectionist framework, and without invoking an external memory management system.

Simple Recurrent Networks

Language is encountered in speech as a temporal sequence of phonemes. The consequence of this for connectionist models is that the input must be integrated over time. The main technique used until quite recently was to transform the temporal dimension into a spatial one. Thus, the input to a network consisted of a set of units where each of the units (or group of units) represented the input at a different point in time. Adopting this approach forces the modeler to set an arbitrary limit on the temporal memory of the system. (see Fanty, 1985 for an example of the application of this approach to natural language processing).

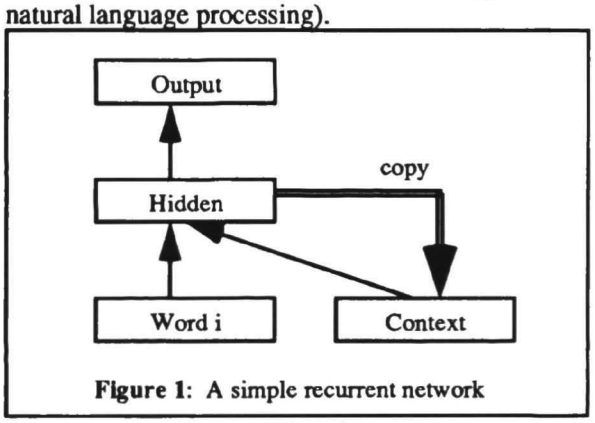

Figure 1: A simple recurrent network

A more appealing solution is to use a recurrent network. Recurrence in a connectionist network is implemented by taking the state of some part of the network at time *t* and using it as input (in addition to

the external input) at time $t+1$. A recurrent network variant, the simple recurrent network (SRN), was first applied to natural language processing by Elman (1990). The structure of a typical SRN is given in Figure 1. It is a standard feedforward network, with an additional set of "context" units. These units store a copy of the activations of the hidden units from the previous time step which are then used as input at the current time step. The weights from the context units are modifiable like any other weights in the network. The context units serve as a form of memory in which aspects of preceding input item are represented with decreasing precision.

Elman (1991) explored the ability of simple recurrent networks to carry information over a long sequence of inputs. The input to Elman's SRN were words in sentence-like sequences, and the task of the network was to output word $n+1$ given word n as input. Take the following sentence from Elman's corpus: *Dog who cat chases sees girl.* There is a dependency between the number of the noun *Dog* and the number of the verb *sees*. If a listener were presented with this sentence and asked, for example, to judge whether or not it was grammatical, then s/he would have to carry information about the number of the noun and use it in assessing whether the verb number was correct. This would have to be done irrespective of the number of intervening words. In reality, the greater the number of intervening words, the more difficulty people have in performing this task. What Elman demonstrated was that given the task of anticipating the next word in a sentence, a simple recurrent network was able to utilise information, such as number, encountered several time-steps previously. However, just as with people, the network did not demonstrate perfect performance, rather the performance degraded as the number of intervening words increased.

Now, the task Elman set his network was by no means as demanding as having to perform a full parse, but it did entail operations that are thought to occur in parsing. The main facet of language processing captured by his network is the impact of expectancies at a given point in a sentence. Such expectancies are known to be an important part of natural language understanding (Marslen-Wilson & Tyler, 1980).

Representation in SRNs

The internal representations developed in the hidden units of the network mediate the sensitivity of SRNs to the structural properties of sentence-like sequences. These representations are quite different from conventional forms of representation used in natural language processing. They are vectors of activation values, typically of high dimension, which vary as a function of time. Consequently, the processing of an input sequence can be thought of as the traversal of a trajectory through a sequence of states in this representational state space.

Elman (1990, 1991) and others (Servan-Schreiber, Cleeremans, & McClelland, 1991) have explored the capacity of these representations primarily from a linguistic point of view. To a large extent this work has been in response to the agenda set by Fodor and Pylyshyn's (1988) critique of connectionism, in which they claimed, among other things, that connectionist representations were inadequate for representing the compositionality of language.

Apart from their representational adequacy, an equally interesting question is the psychological adequacy of SRN representations. Do any interesting psychological consequences arise from viewing mental representation as fixed-width vectors of activation values, and from viewing language processing as the following of a trajectory through a high-dimensional state-space? As an initial step in answering this question it is important to show that SRN language representations demonstrate certain well-known psychological properties of language representations, such as the boundary effects described above.

Generating SRN Representations

The first stage of this study involved the development of representations that could be used in a subsequent study of boundary effects. To this end, a replication of the learning experiment described in Elman (1991) was carried out.

The architecture of the network used is shown in Figure 2. It consists of 24 input units and 24 output units in which words are represented in a localist manner. The two layers of 10 units are designed to create a bottleneck at the input and output to encourage the creation of distributed word representations and also to help reduce the overall connectivity of the network. The context units serve as the temporal memory of the system. The single lines represent full connectivity, and the double line represents one-to-one connectivity (the weight in

this case is 1.0). Note, also, that the *copy* links are not modifiable.

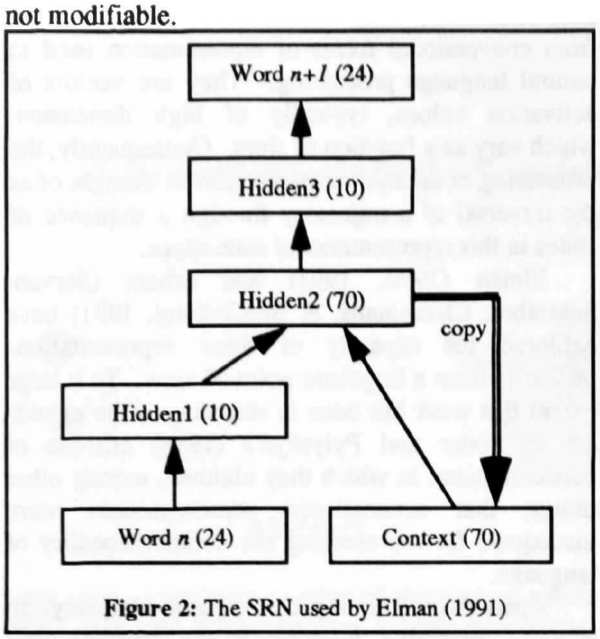

Figure 2: The SRN used by Elman (1991)

Table 1: Elman's (1991) grammar used for generating the SRN training corpus

S	→	NP VP "."
NP	→	PROPN | N | N RC
VP	→	V (NP)
RC	→	who NP VP | who VP (NP)
N	→	boy | girl | cat | dog | boys | girls | cats | dogs
PropN	→	John | Mary
V	→	chase | feed | see | hear | walk | live | chases | feeds | sees | hears | walks | lives

The network in Figure 2 was trained using a corpus of sentences of varying complexity generated from the grammar given in Table 1, identical to that used by Elman (1991). During generation, number agreement between nouns and verbs within a clause was maintained, and between head nouns and verbs in relative clauses, where relevant. Note that the localist word representation did not preserve morphological similarity.

The task of the network was to anticipate the next word in a sequence given the current word. Obviously, at any point in a sentence, the next word is not always completely predictable. Therefore, the performance of the network was judged on whether it could anticipate the right *class* of the next word. Training took place in four distinct phases using a slightly modified version of the back-propagation learning algorithm (Rumelhart, Hinton, & Williams, 1986), in which the cross-entropy rather than the standard sum-of-squares error function was used (cf. Hinton, 1989). During the first phase the network was trained on five epochs of a corpus of 10,000 simple SVO and SV sentences. In the second phase, 25% of the corpus comprised complex sentences (i.e., contained one or more relative clauses). During the third and fourth phase, the percentages of complex sentences rose to 50% and 75%, respectively. A schedule of varying learning rates was also used: 0.09, 0.08, 0.06, and 0.04 for each successive phase. As is usual with SRNs, a momentum term of 0.0 was used.

In Elman's (1991) study, the context units were reset at the end of each sentence. This resulted in loss of information across sentence boundaries. Since, boundaries (both clausal and sentential) are of interest here, a reset was not used. As a consequence, the network learned the prediction task somewhat less well than when a reset was used (82% accurate prediction[1] as compared to 76% with a reset).

The Simulation Experiments

The assumption underlying the following series of simulation experiments is that at any point during the input of a sentence to the SRN, the pattern of activation on the hidden units is a representation of the sentence structure up to that point. It contains information about the preceding words in the sentence and their organisation which should, in principle, be accessible. If the representation has characteristics that are similar to mental representations, then the accessibility of lexical information should be affected by the position of clause and sentence boundaries.

Experiment 1: Sentence boundary effects

The purpose of Experiment 1 was to verify the existence of a sentence boundary effect in a simulated recall experiment. A network was trained to generate a yes/no response when given a sentence representation (i.e., a hidden unit vector from the SRN) and a probe word as input. This type of network will be referred to as a *recall network*, to distinguish it from the SRN (see Figure 3). If the probe word was present in the sentence up to the

[1] An output vector was considered accurate if each element was within 0.1 of its corresponding element in a vector based on transition probabilities derived from the corpus.

point at which the hidden unit vector was extracted (the *interruption point*), the network was trained to output a "yes" response, otherwise a "no" response. To represent the probe word, the 10 unit distributed representation taken from the first hidden layer of the SRN was used. A learning rate of 0.01 and a momentum of 0.9 were used in the training of all recall networks.

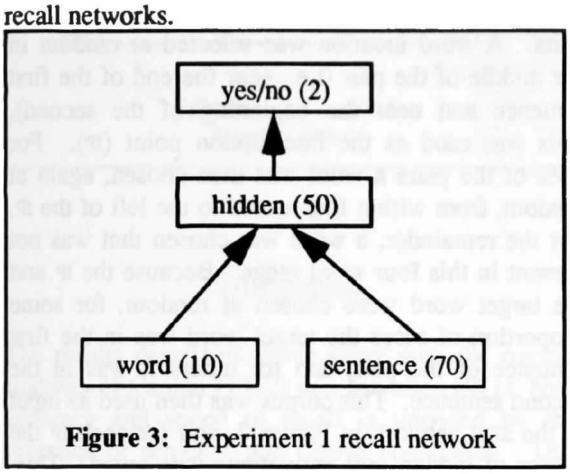

Figure 3: Experiment 1 recall network

The training corpus for Experiment 1 was selected in the following way: A subset was selected of the 10,000 sentence SRN training corpus comprising sentences of length four (including the full stop). This subset consisted of 1,264 sentences and they were made into 632 sentence pairs. These pairs were then used as input to the trained SRN network. For half of the pairs the hidden unit activations were saved after the input of the first word of the second sentence. For the remaining half, hidden unit activations were saved following the input of the third word in the second sentence. The task of the recall network was to decide whether or not a given probe word had been encountered in the preceding input sequence. For those cases in which the target had indeed been present (two-thirds of the cases), the distance between the interruption point and the target was two words. The sentence boundary intervened for half of these cases and a word intervened for the remainder. This, therefore, was the simplest test for boundary effects.

The network was trained for 30 replications of 10 epochs each. A replication involved resetting the weights to small random values and randomly selecting a new set of targets and distracters from 632 sentence pairs. Thus, both "subjects" and materials were treated as random factors. The results of this simulation demonstrated, as hypothesised, that the speed and accuracy of response were impeded by the presence of a sentence boundary. A "yes" response was deemed to have been made if the activation of the "yes" unit exceeded 0.5. A similar level of activation on the "no" unit was taken to indicate a "no" response. Mean squared error (MSE) was used as an analogue of response time (Seidenberg & McClelland, 1989). Table 2 summarises the results of the simulation averaged over the 30 replications. The difference between the boundary/no-boundary conditions was significant for both the accuracy (t=7.12; df=29; p<0.001) and MSE data (t=-2.28; df=29; p=0.03).

Table 2: Results from the recall network of Experiment 1 averaged over 30 replications.

	Proportion correct	Mean squared error
No Boundary	0.94	0.072
Boundary	0.69	0.087

Experiment 2: Clause Boundary Effects and Memory for Relevant Information

The purpose of Experiment 2 was to determine if the effects found for sentence boundaries also held for clause boundaries. An additional goal of this experiment was to explore what kind of information survives boundary transitions, and what information does not.

The training corpus for the recall network used in this experiment was derived in a similar way to the previous one: A set of 128 sentences was constructed, which was input to the SRN, and a set of hidden unit vectors were saved at a number of interruption points. The input to the SRN comprised three-word sentence fragments, each fragment having one of two forms: N *who* N and N V N. Where N could be *boy(s), girl(s), dog(s),* or *cat(s),* and V could be *see(s), hear(s), chase(s),* or *feed(s).* The first noun in each fragment was the target, and the second noun, the interruption point. Note that the end of a who-fragment is still within a clause, but that the end of a verb-fragment is at the end of a clause. If clause boundaries behave in a similar way to sentence boundaries, the target should be less accessible in the verb-fragments than in the who-fragments.

The task of the recall network in this experiment was varied slightly from the previous ones in order to ensure that the boundary phenomenon was not an artefact of the particular type of recall network used. Rather than indicate whether or not a target word had occurred in the fragment, the recognition network was trained to

carry out two tasks simultaneously: (1) to output the number of the target word (i.e., singular or plural), and (2) to indicate whether the target was human or animal. If the SRN is to perform its task successfully, it must be sensitive to verb and noun number. Consequently, number is the type information that should persist over a several input elements. The human/non-human distinction, on the other hand, is one which is of no relevance to the SRN, and is less likely to be encoded in the sentence representation.

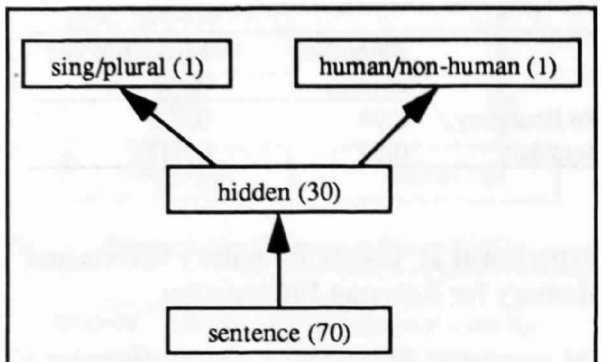

Figure 4: Experiment 2 recall network

The recall network for this experiment was trained for 30 replications of 25 epochs each. The results indicated that number information was retained, but to a lesser degree for the verb-fragments than for who-fragments. This difference was statistically significant for both the accuracy (t=19.23; df=29; p<0.001) and MSE data (t=-25.41; df=29; p<0.001). However, the network behaved at chance level in the retention of the human/non-human distinction for both types of fragment (t<1.0). Details of the analysis of the number data are given in Table 3.

Table 3: Recall of number information in Experiment 2 averaged over 30 replications.

	Proportion correct	Mean squared error
Within clause	1.00	0.078
End of clause	0.61	0.228

Experiment 3: Replication of Caplan's study

The purpose of Experiment 3 was to extend the results of the previous experiments by carrying out a more detailed replication of Caplan's (1972) findings. The architecture of the recall network was similar to Experiment 1, but the training corpus was more extensive.

The training corpus was devised by selecting a subset of the 10,000 sentences used in the fourth phase of the SRN training described above. On this occasion, the subset consisted of the 716 sentences that were exactly nine tokens long (including the full stop). These sentences were then made into 358 pairs. A word location was selected at random in the middle of the pair (i.e., near the end of the first sentence and near the beginning of the second). This was used as the interruption point (IP). For 66% of the pairs a word was then chosen, again at random, from within four words to the left of the IP. For the remainder, a word was chosen that was not present in this four word range. Because the IP and the target word were chosen at random, for some proportion of cases the target word was in the first sentence of the pair, and for others it was in the second sentence. This corpus was then used as input to the SRN network in Figure 2, and for each IP the vector of hidden unit activations was saved. This, along with the 10-unit word representation vector, was used as input to the recall network.

The recall network was trained 30 times for 300 epochs each, using the modified version of the back-propagation learning algorithm discussed in Experiment 1. For each replication a different set of targets and interruption points was chosen so that again there was randomisation over materials as well as "subjects". This yielded a dataset that could be compared with data from the first experiment in Caplan's study. Caplan's data were reaction times for target words as a function of the presence of a clause boundary and distance in syllables between the boundary and end of sentence. A similar set of data were derived from the simulation, again using MSE as a reaction time analogue, and with the interruption point being treated as the end of sentence. Both sets of data are graphed in Figure 5. As in Caplan's study, the boundary condition was significant, with reaction time and accuracy significantly impeded by the presence of a boundary (F(1,29)=31.29, p<0.001, for MSE, and F(1.29)=65.28, p<0.001, for the accuracy data). There was, however, a significant interaction between the presence of the boundary and the distance of the boundary from the IP (F(1,29)=11.76, p=0.002, for MSE, and F(1.29)=5.74, p=0.23, for accuracy). This type of interaction was *not* found in Caplan's data. Note, however, that the apparent trend in the opposite direction in his data was not statistically significant.

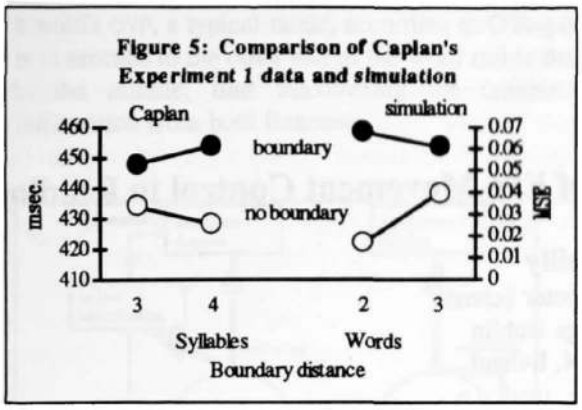

Figure 5: Comparison of Caplan's Experiment 1 data and simulation

Discussion

The results described above agree with the overall findings of Caplan (1972) and Jarvella (1971): both accuracy and speed of recall is impeded by the presence of sentence and clause boundaries. This suggests that the types of distributed representation generated by SRNs have some psychological plausibility. The significant interaction between boundary distance and boundary presence in Experiment 3 may be due to SRNs having a more limited capacity memory than people.

These results are significant because they arise as a side-effect of the sentence representation, and not from an interaction between the representation and some external memory mechanism, as is the conventional explanation. It could be argued that SRN representations are created in a psychologically implausible manner (i.e., by backprop), and consequently the work reported here is of limited relevance. But the manner in which the representations are created need not invalidate the simulation results, since these arise from the distributed and *dynamic* nature of the representations, and are not dependent on how the representations are created.

The findings also suggest a new way of thinking about representation in language processing. Information in a sentence is preserved on a "need-to-know basis": only information that is necessary is maintained. This property could provide a natural way of capturing some of the strategic effects in language processing that have complicated the interpretation of many psycholinguistic experiments in the sentence and discourse area (see Garnham, 1989, for an excellent analysis of this problem).

References

Caplan, D. 1972. Clause boundaries and recognition latencies for words in sentences. *Perception and Psychophysics* 12:73-76.

Elman, J. L. 1990. Finding structure in time. *Cognitive Science* 14:179-212

Elman, J. L. 1991. Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:195-225.

Fanty, M. 1985. Context-free parsing in connectionist networks, TR-174, Dept. of Computer Science, University of Rochester.

Fodor, J. A., Bever, T. G., Garret, M. 1974. *The psychology of language*. New York: McGraw-Hill.

Fodor, J. A., & Pylyshyn, Z. 1988. Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3-71.

Garnham, A. 1985. *Psycholinguistics: Central topics*. New York: Methuen.

Garnham, A. 1989. Integrating information in text comprehension: The interpretation of anaphoric phrases. In G. N. Carlson & M. K. Tannenhaus (Eds.), *Linguistic structure in language processing*. Dordrecht, The Netherlands: Kluwer.

Hinton, G. E. 1989. Connectionist learning procedures. *Artificial Intelligence* 40:185-234.

Jarvella, R. 1971. Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior* 10:409-416.

Marslen-Wilson, W. D., & Tyler, L.K. 1978. The temporal structure of spoken language understanding. *Cognition* 8:1-71.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. 1986. Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland and the PDP Research Group (Eds.), *Parallel distributed processing. Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Seidenberg, M., & McClelland, J. L. 1989. A distributed developmental model of visual word recognition and naming. *Psychological Review* 96:523-568.

Servan-Schreiber, D., Cleeremans, A., &, McClelland, J. L. 1991. Graded state machines: The representation of temporal contingencies in simple recurrent networks. *Machine Learning* 7:161-193.