Selective attention in the acquisition of the past tense

Dan Jackson
Cognitive Science & Linguistics 0108
University of California, San Diego
La Jolla, CA 92093
jackson@ling.ucsd.edu

Rodger M. Constandse
Computer Science & Engineering 0114
Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093
rconstan@cs.ucsd.edu

Garrison W. Cottrell
Computer Science & Engineering 0114
Institute for Neural Computation
University of California, San Diego
La Jolla, CA 92093
gary@cs.ucsd.edu

Abstract

It is well known that children generally exhibit a "U-shaped" pattern of development in the process of acquiring the past tense. Plunkett & Marchman (1991) showed that a connectionist network, trained on the past tense, would exhibit U-shaped learning effects. This network did not completely master the past tense mapping, however. Plunkett & Marchman (1993) showed that a network trained with an incrementally expanded training set was able to achieve acceptable levels of mastery, as well as show the desired U-shaped pattern. In this paper, we point out some problems with using an incrementally expanded training set. We propose a model of selective attention that enables our network to completely master the past tense mapping and exhibit U-shaped learning effects without requiring external manipulation of its training set.

Introduction

It is well known that in the process of acquiring the past tense, children generally exhibit a "U-shaped" pattern of development. The first past tense forms produced are generally correct, regardless of whether or not those forms are regular. After this period of correct performance, children go through a period of overgeneralization in which irregular forms are incorrectly inflected (e.g. *goed*). Finally, children seem to identify some forms as exceptions to the general regular pattern, and the overgeneralization errors decrease. Plunkett & Marchman (1991) (P&M hereafter) showed that U-shaped learning effects can emerge in connectionist networks in the absence of any discontinuities in the training regime.[1] P&M showed that such networks go through "micro U-shaped development." This is contrasted with the idealized vision of "macro U-shaped development" that predominates in anecdotal descriptions of children's patterns of acquisition. Macro U-shaped development refers to a rapid and sudden change from the memorization stage, where regular and irregular forms are reproduced with relatively equal levels of error, to a stage where the /-ed/ suffix is applied indiscriminately, resulting in overgeneralization for all irregular verbs. Micro U-shaped development, on the other hand, is characterized by selective application of the /-ed/ suffix, resulting in a period in which some irregular verbs are regularized, while others are produced correctly. Although most anecdotal descriptions of children's acquisition of the past tense have implied macro U-shaped development, studies of naturalistic past tense production (e.g. Marcus et al. (1992)) and studies using elicitation procedures (e.g. Marchman (1988)) show that micro U-shaped development is a better description of how children learn the past tense.

Although P&M (1991) were successful in showing that connectionist networks go through a micro U-shaped pattern of development, none of the networks they trained achieved mastery of all of the past tense mappings. In particular, the mean performance on the regular (add /-ed/) mapping was 84% (P&M (1991), p. 71), which is well below the percentage of regulars that most adult humans are able to inflect correctly (near 100%).

P&M (1993) demonstrated that networks can achieve acceptable levels of mastery and still show U-shaped learning effects if their training set is expanded incrementally. Unlike Rumelhart & McClelland (1986), they did not introduce a discontinuity in the training regime. Rather, they trained their networks on a small number of verbs at first, and then gradually expanded the training set. Trained in this way, the networks described by P&M (1993) were able to master the given vocabulary (correctly inflecting 97-98% of the regulars) after a period of micro U-shaped development.

This is an interesting result, but the use of an incrementally expanding training set must be justified. P&M (1993) note that "verb acquisition in children is a gradual process which follows an *incremental* learning trajectory (p. 27)," and go on to mention Elman's (1991) application of incremental training to the acquisition of simple and complex syntactic forms. There are, however, some crucial differences between the account of language acquisition implied by Elman's model and that implied by P&M (1993).

Elman's recurrent network was unable to learn adequately if it was trained on the entire set of simple and complex sentences at once. He showed that it could learn if it was trained on an incrementally expanded training set, beginning with the simple sentences, and working up to the complex ones. Nevertheless, he argued explicitly against using an

[1] See Pinker and Prince's (1988) critique of Rumelhart & McClelland (1986). They argue that Rumelhart & McClelland's model exhibited U-shaped learning effects because of discontinuities in its training set.

incrementally expanded training set in models of language acquisition, claiming that "children hear exemplars of all aspects of the adult language from the beginning (p. 6)." He then tried expanding his network's memory capacity, rather than incrementally expanding its training set. During the first phase of training, the recurrent feedback was eliminated after every third or fourth word. As training progressed, the network's memory window was gradually increased until the feedback was no longer interfered with at all.

Using this schedule of expanding memory, Elman was able to get the network to learn the entire training set. This is a reasonable account of language acquisition because we know that children have limited memory capacity early in development, and that this capacity increases as development continues. Furthermore, the network is exposed to the entire adult language, which is more realistic than using a subset of the language for training.

P&M's (1993) model did not have a limited memory--it was not a recurrent network, and did not have a memory in the sense that Elman's (1991) model did. P&M had to resort to limiting its training set, which was then gradually expanded. P&M (1993) claim that it is "unlikely that children attempt to learn an entire lexicon all of a piece (p. 27)." Perhaps what they had in mind was that children have access to the entire vocabulary, but only pay attention to a limited number of words. In this case, the way they have modeled attention is questionable. At the outset of training, the network was given 20 verbs, on which it is trained to 100% accuracy before expansion began. In effect, the network was being told which verbs to pay attention to at the outset, and trained on them to perfection before it could start attending to other verbs. By the end of training, the network had the entire vocabulary in its training set--it was paying attention to each element of the vocabulary to the same degree. Clearly, we need a better way to model attention.

In this paper, we examine the effect of selective attention on a network's ability to learn the past tense mappings. We do not specify the examples to which the network should pay attention, and we do not restrict the set of examples to which the network can be exposed. Like Elman, we believe that in order for our model to be realistic, the entire vocabulary must be accessible to the network from the start. We show that networks with this mechanism of selective attention master the past tense mapping and exhibit micro U-shaped learning effects in the absence of any external manipulation of their training set.

Selective Attention Model

Our model of selective attention is based on the method of *active selection* (Plutowski & White (1993)). This method was originally used for incrementally growing a training set by using a partially trained network to guide the selection of new examples. Plutowski et al (1993) introduced the idea of using maximum error as the criterion for selection. In our implementation, this criterion is used for selecting examples for weight adjustment (cf. Baluja & Pomerleau (1994), whose network ignores sections of the input with high prediction error). Instead of using active selection for incrementally growing the training set, we assume a fixed size training queue of size N corresponding to the child's working or perhaps episodic memory. As the child samples the environment, we assume the child computes his error on any verb, and then compares this error with what is currently in the training queue. If the error on the sampled example is worse than what is currently in the queue, the example is inserted in the queue and the best example in the queue is "forgotten." We may view this error as a measure of the novelty or salience of the verb.

To simulate this, at the beginning of an epoch the simulator randomly selects a window of W examples from the vocabulary and tests the network on them. The likelihood of any particular example being chosen for the sample window depends on its frequency in the vocabulary. The above procedure is applied to update the queue. Thus, the entire set of examples may change from one epoch to the next depending on N, W, and the error on the samples.

The network's initial exposure to a form results in its being placed in the sample window. Weight adjustment does not occur until the form has been put into the training queue. Training on a verb is therefore "off-line" in the sense that it occurs some time after the verb is initially encountered.

It is reasonable to suppose that children are not able to cycle through every verb in the language in order to choose the ones they need to pay attention to for the purposes of synaptic adjustment, so it was important to limit the size of the sample window. We might think of the window as the network's short-term memory (for recently heard verbs). It needs to hold a limited number of items in memory so that it can compare them to choose the queue elements when updating the queue.

Methods

Our input-output pairs were taken from the database used by P&M. The interested reader should refer to P&M (1991, 1993) for details about the representations. The network is given a verb stem as input and must produce the inflected verb as its output. The transformations from the stems to the past tense forms are classified into four possible classes: arbitrary, identity, vowel change, and regular. Each of these corresponds to a possible English past tense transformation.

	Arbitrary	Identity	Vowel Change	Regular
Type Frequency	2	20	68	410
Token Frequency	100	2	5	1

Table 1: Type and Token frequencies of the past tense mappings

For the arbitraries, there is no relation between the stem and the past tense form, e.g. 'go→went.' For the identities, the past tense form is identical to the verb stem. This mapping requires that the verb stem end in a dental consonant (/t/ or /d/), e.g. 'hit→hit.' For the vowel changes, a vowel in the stem may be replaced by a different

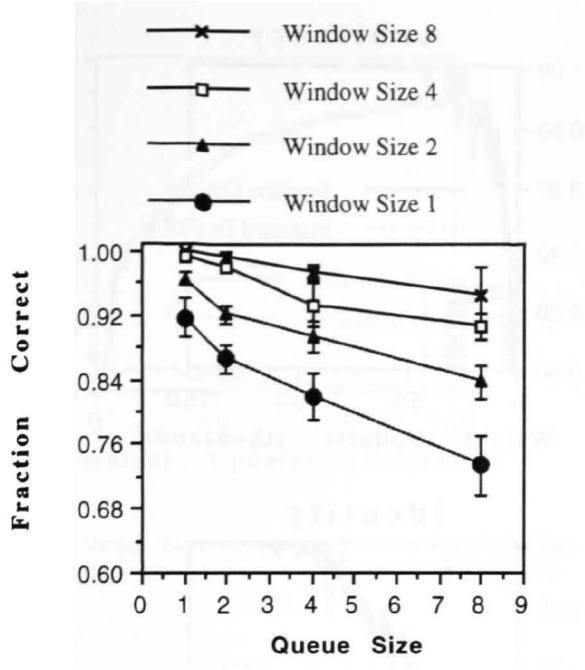

Figure 1: Performance on regular verbs after 120,000 weight updates as a function of queue and window size.

vowel in the inflected form of the verb, depending on the original vowel and the consonant that follows. We had 10 different types of vowel changes in our vocabulary, analogous to 'ring→rang,' 'blow→blew,' etc. Finally, for the regulars, a suffix is appended to the verb stem. The form of the suffix depends upon the final vowel/consonant in the stem. If the stem ends in a dental (/t/ or /d/), then the suffix is /-id/, e.g. 'pat→pat-id.' If the stem ends in a voiced consonant or vowel, then the suffix is voiced /d/, e.g. 'dam→dam-d.' If the stem ending is unvoiced, the suffix is unvoiced /t/, e.g. 'pak→pak-t.'

The type and token frequencies of each of these classes in our vocabulary are shown in Table 1. The type frequencies are identical to those used by P&M (1991), but the token frequencies are somewhat different. For each type of past tense mapping, we took the averages of a small, but representative sample of verb frequencies from Kucera & Francis (1967), and then normalized them by the frequency of the regulars.

Our networks were trained with the back propagation algorithm. The network architecture consisted of 18 input units (each verb stem was formed from 3 phonemes each requiring 6 units to represent), 30 hidden units and 20 output units (2 suffix units were needed in addition to the transformed stem). The choice of 30 hidden units was made to parallel the architecture used by P&M (1993). The learning rate and momentum were also set according to the values used by P&M (1993), namely a learning rate of 0.1 and a momentum of 0.0. To evaluate network performance, the output for each phoneme in the stem was mapped to the closest legal phoneme (using Euclidean distance). Then the output was compared with the target.

We investigated the effects of different sample window and training queue sizes by letting W and N take on the values 1, 2, 4 or 8 and training networks with all sixteen possible combinations. Five sets of networks were trained, with initial weight values the same within each set, but varying between them.

Results

Figure 1 shows the effect of using different training queue and sample window sizes. The average performance for each combination of W and N is plotted in Figure 1, with standard deviation indicated by error bars. The networks that performed best were the ones that had large sample windows and small training queues. The larger the sample window, the more examples the network has to choose from. Once an example is chosen and trained on, however, the network's error will change not only for that verb, but for other verbs as well. If the network trains on a regular verb, for example, we expect its error on other regular verbs to go down slightly, as well. Thus, it is better for the network to "pay attention to one thing at a time," because this allows it to choose its training example based on its error on that example immediately prior to training, rather than using an error value that may have changed due to training on another verb in the queue.

Figure 2 shows the average performance of 5 networks trained using the selective attention mechanism with sample windows of size 8 and training queues of size 1. Because of the method of training we are using, it is more meaningful to analyze the networks according to the number of weight updates they have undergone. This makes it difficult to compare our results with those of P&M, however, because they graph results in terms of epochs, and the size of the training set changes with each epoch. For the purposes of comparison, therefore, we ran 5 networks using the traditional method for selecting training examples (the one used by P&M (1991)) on the same data. The average performance of these networks on regular verbs is shown in figure 3.

The networks with selective attention performed very well. By 125,000 weight updates, all 5 networks had mastered all of the past tense mappings (with a standard deviation of 0.0). When the networks without selective attention had reached 125,000 weight updates, they only inflected an average of 85% of the regulars correctly (with a standard deviation of 0.013). This is the level of performance reached by P&M's (1991) best network at the end of training (p. 71). Even after 500,000 weight updates, these networks only got an average of 95% correct (with a standard deviation of 0.007). In summary, the network with selective attention was better both in final performance and learning speed.

Figure 4 shows the networks' ability to generalize. The first graph shows the average performance of the five networks on novel verbs that did not fall into any of the vowel change classes or end in a dental consonant--the indeterminates. The error bars indicate standard deviation.

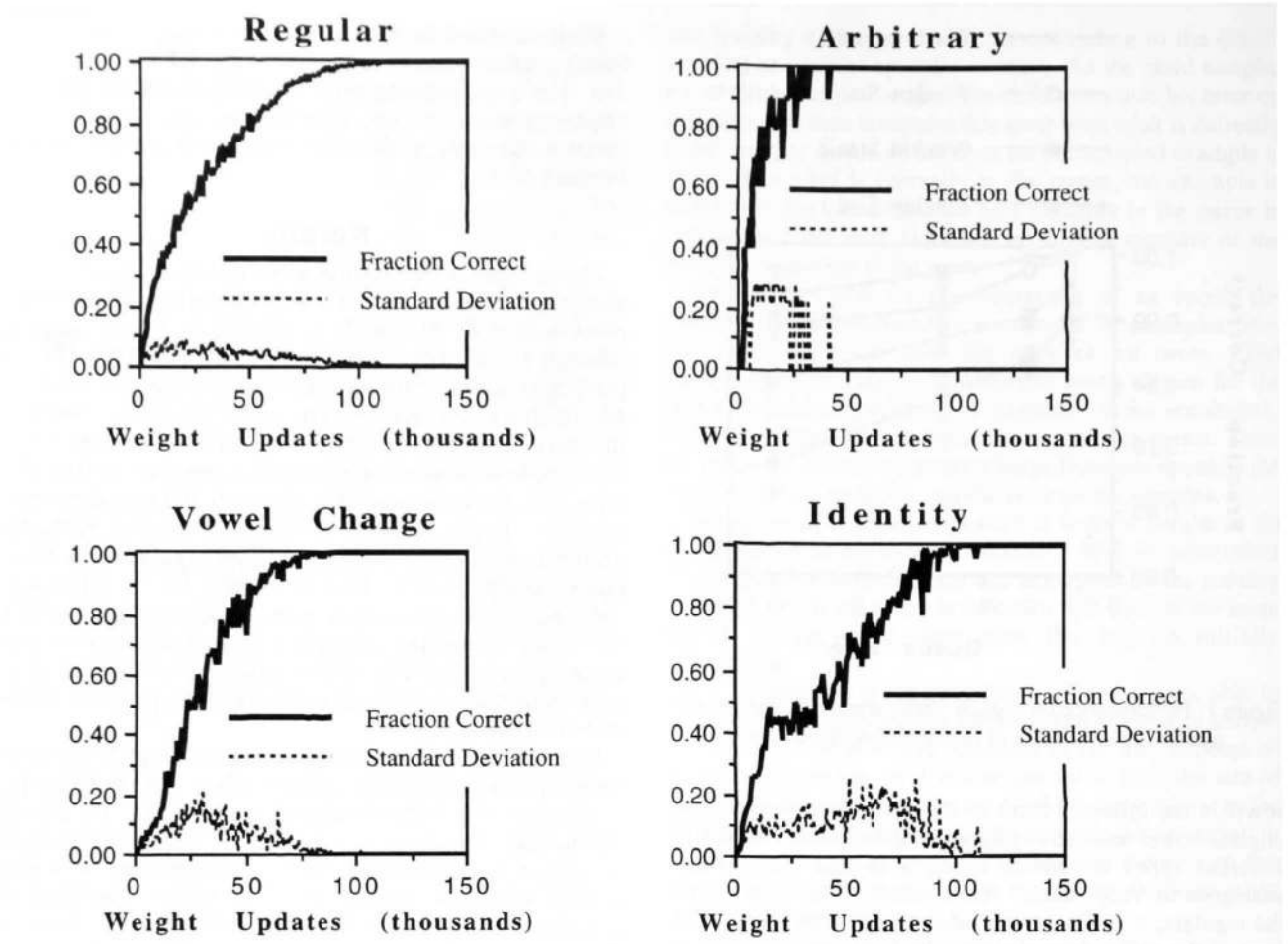

Figure 2: Average fraction correct and standard deviation for the five selective attention networks tested on all of the regular, arbitrary, vowel change and identity verbs in the training set.

As can be seen in the graph, the networks generalize fairly well. The dashed line shows the fraction of indeterminate novel verbs the networks inflected with a suffix (around 90%), irrespective of whether the form of the stem was correct. The dashed line shows the fraction of indeterminate novel verbs the networks inflected as regulars with no changes to the stem (around 70%).

The novel dental and vowel change graphs show that the regular mapping is not applied indiscriminately to novel forms--the fraction of verbs inflected as regulars is lower in these graphs. The networks have learned something about the phonological regularities inherent in the vocabulary. In particular, the novel vowel change graph shows that verbs that are phonologically similar to the vowel change verbs in the training set are as likely to be inflected with a vowel change as they are to be regularized.

Figure 5 shows the number of times each verb token was in the training queue for a particular simulation. Note that some regular verbs never make it into the queue, i.e. are never trained on. Since we happen to know all verbs were sampled, the network must have had low error on these verbs when they were in the window. This is further evidence that the network has learned the regular rule, and shows that our procedure avoids unnecessary computation.

Discussion

We have presented a model of selective attention which chooses training examples from a random sample of the training set. The size of the window from which the training example can be chosen is limited and the training queue itself is limited. Whether only one or both of these should be considered "memory" is a question of interpretation, but here we have suggested that the queue can be considered the memory. One could also break the processing down into two stages, one where samples are put into memory for later processing, and then a stage in which they are organized according to salience, and then practiced.

The idea that children process a significant amount of the linguistic input they receive after the fact is corroborated by data concerning crib speech--monologues and language practice (including grammatical modifications and imitation/repetition) that children engage in when they are alone in their bed before going to sleep (Jespersen (1922), Weir (1962), Kuczaj (1983)). Crib speech is characterized by a freedom (because of the lack of communicative intent) to use free association to generate sequences of sounds and words, the associations being either phonological, syntactic

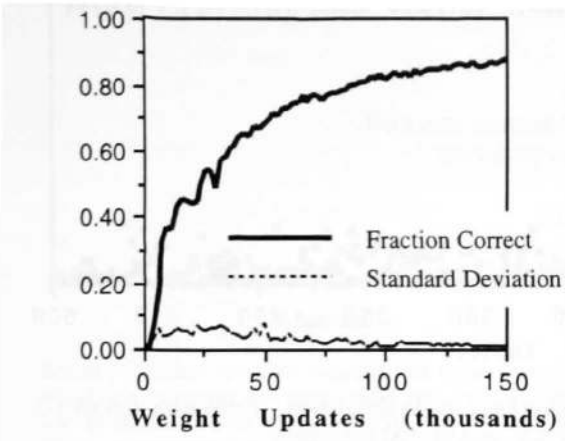

Figure 3: Average fraction correct and standard deviation for five networks without selective attention tested on regular verbs.

or semantic (Britton (1970)). Probably because of this freedom, children are more likely to engage in language practice in crib speech than in social-context speech (in terms of relative frequency) (Black (1979), Britton (1970), Kuczaj (1983)).

As Kuczaj writes:

> ...children process linguistic information at (at least) two levels: (a) the level of initial processing, which occurs in short-term memory shortly after children have been exposed to the input, and (b) the level of post-initial processing, which occurs at some later time when children are attempting to interpret, organize, and consolidate information that they have experienced over some longer period of time...children are most likely to notice discrepancies between their knowledge of language and linguistic input at the level of post-initial processing, and...crib speech is a context in which children may freely engage in overt behaviors that facilitate both post-initial processing and the successful resolution of moderately discrepant events. Although older children and adults may be able to notice discrepancies during the initial processing of linguistic information, it is unlikely that young children are able to do so...Children may initially store new forms and new meanings and later compare these new acquisitions with previous ones in post-initial processing (Kuczaj (1983), pp. 167-168).

The model we have presented is completely compatible with these observations, if one assumes both the queue and the window are part of the memory. The "discrepancies" in this case are the error signals the network generates for each verb in the sample window. The network can generate the form it expects to see in a particular context, and compare this with what it actually heard. In this way, the network supplies itself with indirect negative evidence (Elman (1991)), which is used in the adjustments of its weights. As Kuczaj suggests is true for children, our networks could not

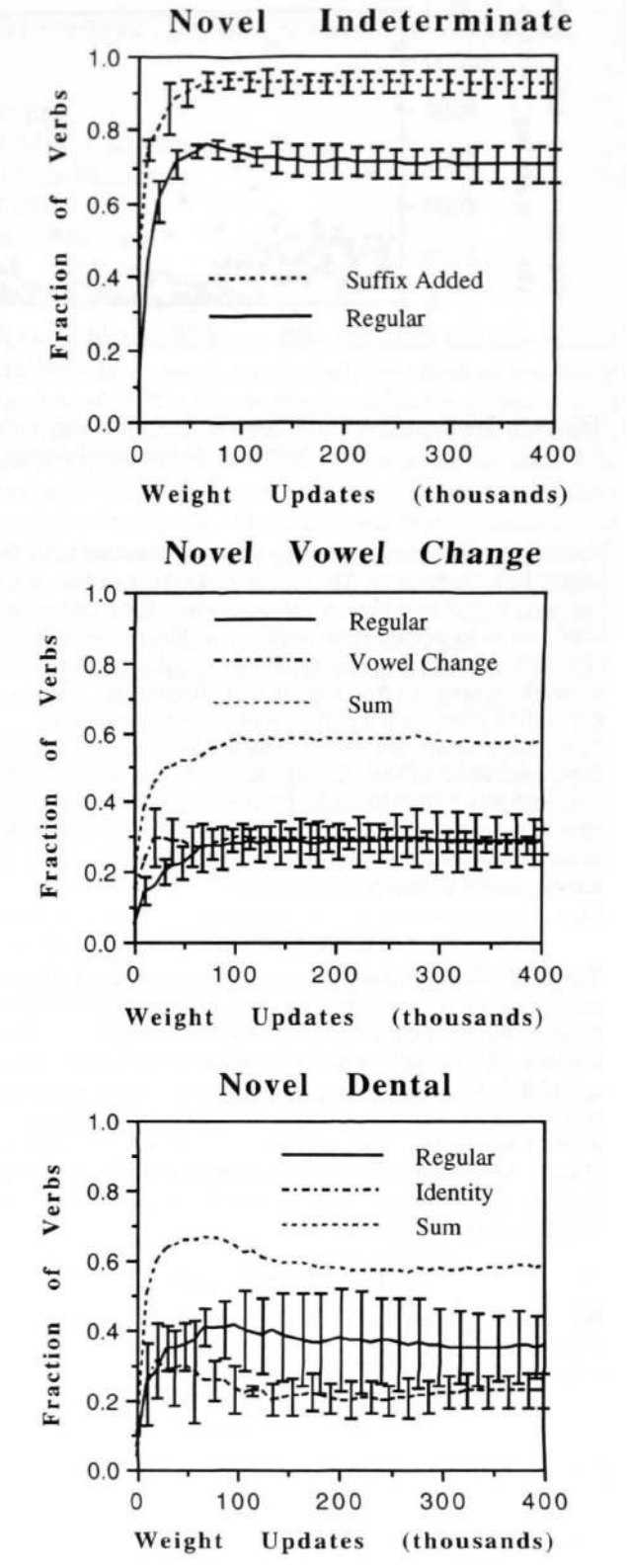

Figure 4: Average generalization on 132 novel indeterminates, 62 novel vowel change verbs and 28 novel dentals.

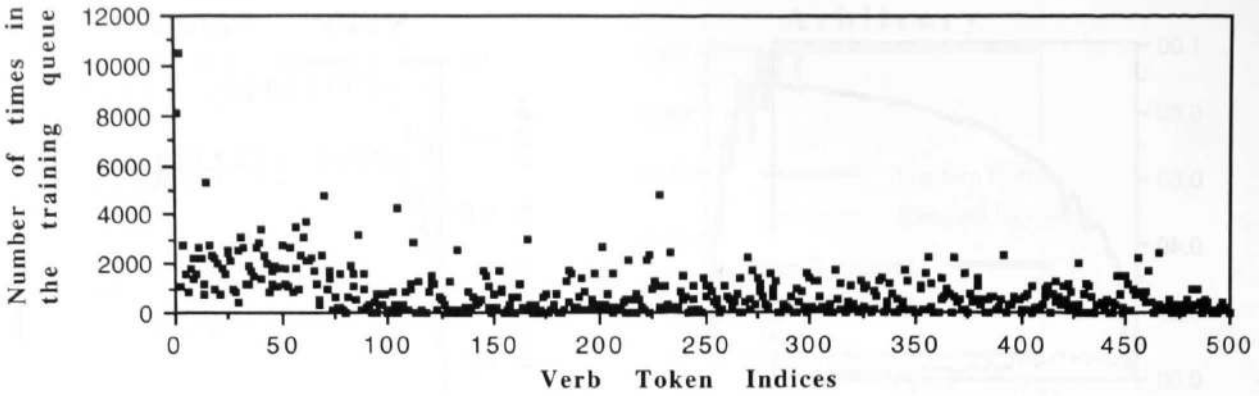

Figure 5: Total number of times each verb token was in the training queue for a single simulation. Verbs with indices 1-2 are arbitrary, 3-70 are vowel change, 71-90 are identity and 91-500 are regular.

"notice" the discrepancies during the initial processing of the linguistic information. The initial processing occurs when the verb is put into the sample window. Later, when the time comes to update a network's weights, the mechanism of selective attention comes into play. At this point, the network generates error signals and chooses the verb with the highest error for the purposes of weight adjustment.

In future work, we may try to develop a connectionist implementation of the training queue. We would also like to investigate other strategies for deciding what the network should pay attention to. Finally, we plan to use our model of selective attention in more primary tasks, such as learning word meaning.

Conclusion

The mechanism of selective attention we introduced allowed the networks to guide their own training. The networks focused on the examples for which they needed the most training. As a result, they performed extremely well. They completely mastered the regular, identity, vowel change and arbitrary past tense mappings and showed the ability to generalize. They also showed micro U-shaped learning effects. Most importantly, our networks achieved their high level of performance without requiring us to externally manipulate their training sets.

Acknowledgments

We would like to thank the GEURU research group for helpful comments. This research was supported in part by NSF grant IRI 92-03532.

References

Baluja, S. & Pomerleau, D.A. (1994) "Using a saliency map for active spatial selective attention: implementation & initial results." In: Advances in Neural Information Processing Systems 7, (Tesauro, Touretsky & Leen, eds.). Cambridge, MA: The MIT Press.

Black, R. (1979) "Crib talk and mother-child interaction: A comparison of form and function. Papers and reports on child language development, 17, 90-97.

Britton, J. (1970) Language and learning. London: Penguin Books.

Elman, J. (1991) "Incremental learning, or the importance of starting small." CRL Technical Report No. 9101. University of California, San Diego.

Jespersen, O. (1922) Language: its nature, development and origin. New York: Allen and Unwin.

Kucera, H. & Francis, W.N. (1967) Computational analysis of present-day American English. Providence, RI: Brown University Press.

Kuczaj, S.A. (1983) Crib speech and language play. New York: Springer-Verlag.

Marchman, V. (1988) "Rules and regularities in the acquisition of the English past tense." Center for Research in Language Newsletter, 2 (4).

Marcus, G.F., Ullman, M., Pinker, S., Hollander, M., Rosen, T.J., & Xu, F. (1992) "Overregularization in language acquisition." Monographs of the Society for Research in Child Development, 57 (4), Serial No. 228.

Plunkett, K., Marchman, V. (1991) "U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition." Cognition 38, 43-102.

Plunkett, K., Marchman, V. (1993) " From Rote Learning to System Building: Acquiring Verb Morphology in Children and Connectionist Nets." Cognition 48, 21-69.

Plutowski, M., White, H. (1993) "Selecting concise training sets from clean data" *IEEE Transactions on neural networks*, 3:1.

Plutowski, M., Cottrell G.W., White, H. (1993) "Learning Mackey-Glass from 25 examples, plus or minus 2." In: Advances in Neural Information Processing Systems 6, (Hanson, Cowan and Giles, eds.). San Mateo, CA: Morgan Kaufmann.

Rumelhart, D. E., McClelland, J.L. (1986) "On Learning the Past Tense of English Verbs", PDP: Explorations in the Microstructure of Cognition, Vol 2.

Weir, R.H. (1962) Language in the crib. The Hague: Mouton.