ELSEVIER

Cognition 73 (1999) 297–300

COGNITION

www.elsevier.com/locate/cognit

Discussion

Simple recurrent networks can distinguish non-occurring from ungrammatical sentences given appropriate task structure: reply to Marcus

Douglas L.T. Rohde*, David C. Plaut[1]

School of Computer Science, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213-3890, USA

In Rohde and Plaut (1999) we reported on a series of connectionist simulations in which simple recurrent networks (SRNs) were trained to successively predict the next word in sentences generated by a simple stochastic grammar with some of the complexities of English, including number agreement, variable verb argument structure, embedded clauses, and semantic biases on noun-verb co-occurrences. Our main goal was to demonstrate that connectionist networks inherently ''start small'' in language learning without the need for any external manipulation of either the training environment or internal memory resources (*contra* Elman, 1991, 1993). We also argued that the results support a perspective on language learning – certainly not original with us – that any lack of explicit negative evidence provided to children need not implicate innate, domain-specific learning constraints because implicit prediction within a stochastic language environment can provide sufficient implicit negative evidence.

In his commentary, Marcus points out that our modeling work did not address a number of issues that he considers to be critical to understanding language learning. Although we might quibble over some of the issues, we are in complete agreement that any simulation that only generates predictions over word representations could not possibly constitute a fully comprehensive model of language acquisition and processing. Indeed, as stated in our article, our view is that:

> Although word prediction is a far cry from language comprehension, it can be viewed as a useful component of language processing to the extent that learning a grammar is useful, given that the network can make accurate predictions only by learning the structure of the grammar (p. 71).

* Corresponding author. Fax: +1-412-268-5060.

E-mail address: dr@cs.cmu.edu (D.L.T. Rohde)

[1] Co-corresponding author. Fax: +1-412-268-5060; E-mail: plaut@cmu.edu.

PII: S0010-0277(99)00055-4

Along these lines, we are currently developing a connectionist model of sentence processing in which implicit prediction plays an important role in linking comprehension and production. We hope this work goes further in addressing some of the issues that Marcus mentions, particularly with regard to accounting for detailed behavioral data.

The more substantive challenge raised by Marcus concerns what he takes to be a ''serious, principled limitation'' in using word prediction by an SRN to learn grammatical knowledge. Put briefly, every time a word does *not* occur in a given context, learning reduces its likelihood of being predicted in that context, which seems problematic for cases in which non-occurring words are nonetheless valid (i.e. grammatical) continuations. Marcus gives an example involving a novel verb *fleedle* in which, although Smith only ever fleedles Jones, sentences in which Smith fleedles other individuals are still grammatical. Marcus has also carried out simulations with SRNs in which the likelihoods of non-occurring continuations are reduced to near zero, implying that the network treats them as ungrammatical rather than simply unlikely (see Marcus, 1998, 1999; as well as the URL in Marcus' commentary).

However, these arguments and demonstrations fail to adequately take into account a fundamental property of distributed connectionist networks: learning and processing are strongly influenced by the similarity among *all* of the items and contexts in the training environment, where the relevant similarity is not only in terms of surface forms but also in terms of underlying functional relationships (as reflected by learned, internal representations). Thus, in developing a model of how people behave with particular items in particular contexts, it is rarely adequate to train a network on only those items in those contexts. Rather, a network would be expected to generalize the way people do only if its training environment adequately approximated the full range of relevant surface and functional similarities that people experience in the domain.

To make this point concrete, we carried out a very simple simulation of Marcus' fleedle example under two conditions (see http://www.cnbc.cmu.edu/ ~ dr/Fleedle for details). In the first, an SRN using localist input and output units was trained on SVO (Subject-Verb-Object) sentences with four people (Smith, Jones, Ripken, Bellanger) and one verb (fleedled), where if S was Smith then O was always Jones. In the second condition, the network was first trained on SVO sentences using the same four people but seven other verbs, with no constraint on S and O selection. We then introduced the fleedle sentences, in which Smith only ever fleedled Jones, while continuing to train on the sentences with the other verbs 75% of the time. Prediction performance following ''Smith fleedled…'' as a function of training experience with fleedle sentences under each condition is shown in Fig. 1.

In both conditions, as ''Smith fleedled…'' is always followed by Jones, Belanger and Ripken become progressively less likely as continuations. Of course, it makes sense that they are treated as much less likely than Jones; after all, this distinction reflects a fundamental contingency in the environment. However, Belanger and Ripken are quickly treated as ungrammatical (near zero likelihood) when the

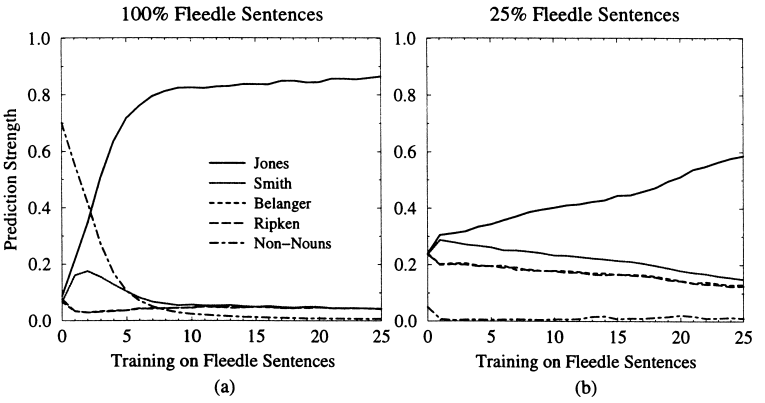

Fig. 1. Prediction accuracy following ''Smith fleedled…'' for a network (a) trained only on sentences containing ''fleedled'', or (b) also trained on sentences containing other verbs. Results are averaged over 50 runs with different random initial weights.

network is trained only on the fleedle sentences, analogous to what Marcus has found in other contexts. By contrast, when the system has even meager experience with other verbs, it treats Belanger and Ripken as unlikely but clearly grammatical (by comparison with non-nouns, for example). The reason is that Belanger and Ripken behave just like Jones in the context of the other verbs, and the internal similarity induced from those contexts influences how the fleedle sentences are processed.

Note that the same point applies to the second domain that Marcus considers: the tendency for infants habituated to ABA versus ABB syllable sequences to generalize to analogous sequences composed of novel syllables (Marcus, Vijayan, Bandi & Vishton, 1999; also Gomez & Gerken, 1999). The syllables used in the testing phase may have been novel in the context of the experiment (and were completely novel in the simulation Marcus mentions), but they were certainly not novel in the context of the full range of auditory experience of the infants. The generalization behavior of the infants presumably depends heavily on representations derived from the similarity structure of auditory experience outside the laboratory (see McClelland & Plaut, 1999, for further discussion).

To be clear, Marcus is right that, from a connectionist perspective, each occurrence of ''Smith fleedled Jones'' is a very small amount of evidence against the grammaticality of ''Smith fleedled Ripken''. What he fails to consider is that this small influence is overwhelmed by the indirect evidence for the grammaticality of the sentence based on the similarity of the behavior of Smith and Ripken (and all other nouns) in a vast range of other contexts. Marcus' simulations do not exhibit the appropriate generalization behavior because their training environments lack the appropriate task structure. If there is a ''serious, principled limitation'' in using sequential connectionist networks to learn grammatical knowledge through implicit prediction, it has yet to be identified.

Acknowledgements

We thank the CMU PDP research group for helpful comments and discussions.

References

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

Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.

Gomez, R. L., & Gerken, L. (1999). Artificial grammar learning by 1-year-olds leads to specific and abstract knowledge. *Cognition*, *70*, 109–135.

Marcus, G. F., Vijayan, S., Bandi Rao, S., & Vishton, P. M. (1999). Rule learning by seven-month-old infants. *Science*, *283*, 77–80.

Marcus, G. F. (1998). Can connectionism save constructivism? *Cognition*, *66*, 153–182.

Marcus, G. F. (1999). *The algebraic mind: integrating connectionism and cognitive science*, Cambridge, MA: MIT Press.

McClelland, J. L., & Plaut, D. C. (1999). Does generalization in infant learning implicate abstract algebra-like rules? *Trends in Cognitive Sciences, 3. With reply by G. Marcus*, 166–168.

Rohde, D. L. T., & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, *72*, 67–109.