Allen, R.B., IEEE First International Conference on Neural Networks (San Diego, June 1987), v.2, pp.335-341.

Several Studies on Natural Language and Back-Propagation

Robert B. Allen
Bell Communications Research
Morristown, NJ 07960 U.S.A.
rba@bellcore.com

Recent developments in neural algorithms provide a new approach to natural language processing. Two sets of brief studies show how networks may be developed for processing simple demonstratives and analogies. Two longer studies consider pronoun reference and natural language translation. Taken together, the studies provide additional support for the applicability of these algorithms to natural language processing.

1. Introduction

The processing of natural language and models of language use are among the most difficult problems for artificial intelligence. Many of the difficulties revolve around the issues of context[1] and exceptions. Context effects appear in pronunciation, in semantics, and even in the social meaning[2] given to an utterance. Neural networks and parallel distributed processors are particularly well suited to natural language processing because they are sensitive to context and exception.[3] [4] In addition, they are self-adapting, allow complex representations of information,[5] and are consistent with psychological models which have been proposed for natural language processing.[6]

In the following four sections, the back-propagation algorithm is used to demonstrate the learning and performance of a variety of tasks related to natural language. Networks were trained to complete analogies (Section 2), to select items in a perceptual field (Section 3), to demonstrate pronoun reference (Section 4), and to translate from English to Spanish (Section 5).

2. Analogies

The performance of analogies is related to cognitive ability.[7] Analogies are also closely related to metaphor which has been suggested to be crucial to language use.[8] [9] In the procedure employed here, analogy is considered to be the learning of systematic patterns operating on input patterns. While a similar process may be used on analogies involving internal representations of complex concepts, that was not directly tested.

In the first study, elements of the analogy were constructed of binary codes, which could be considered strings of features. The **a** and **b** elements differed by two features, the **a** and **c** elements differed by two other features. The network was to generate the **d** element which differed from the **c** element on the same bits that distinguished **b** from **a**. Some typical patterns are shown below; for instance, the first two elements of the first example differ in the 2nd and 4th bits. The network must generate an output element **d** which differs from the third element on just those two bits. This turns out to be related to the parity problem.

440 patterns were generated with these constraints and 400 of them were used to train a 18-30-6 network in which the 18 input units were fully interconnected with 30 hidden units and the 30 hidden units were connected to 6 output units. A learning rate (η) of 0.1 and momentum (α) of 0.9 were used here and in the studies in the following section. Input/output pairs were randomly selected from the training set for 80K

training presentations. The average error, $\Sigma abs(target-actual)$, reached 0.002. During a transfer test to the 40 additional analogies, all bits were correct when a threshold of 0.5 was applied to each output unit.

input			output
a	b	c	d
0011011	0110011	1001011	1100011
1111001	1100001	1111010	1100010

In a second study elements of the analogies were used, each composed of three real-valued members in the range 0-1. If each set of values is thought of as a point in a 3-dimensional space, then solving an analogy may be thought of as finding the distance and direction of movement between the final two elements that is the same as the movement between the first two elements.[10] In this study, the network generated a fourth set of points given the first three. 420 patterns were generated and 20 saved for the transfer test. A 9-15-3 network, trained for 60K pattern presentations, completed the transfer task with an average error of 0.02.

3. Demonstratives and Descriptions of Relative Position

Some essential functions of language are to select objects and to describe the relationships between objects in a perceptual field. In this section, the capability of networks to learn these types of links is demonstrated. In the first study, items in one part of the input field were selected by a code in a second part of the input field. Very roughly, this may be related to the use of demonstratives. For instance, the net might determine which of two objects was to the right (or left) of the other. This could be stated as 'name *this* object,' and 'name *that* object.' Some examples of the coding are shown below. The first two parts of the input field are localist encodings, the third part of the input selects left/right, and the output is a 3-bit distributed encoding representing the correct response.

examples			coding			
input		output	input			output
A B	'that'	a	00000001	00000010	10	110
D C	'this'	c	00001000	00000100	01	010

84 input patterns were generated from pairing 8 objects. From these, 10 were reserved for transfer. An 18-10-3 network trained for 40K cycles reached a error rate of 0.005. There were no errors on the transfer task.

Descriptions may also be matched to stimuli in a perceptual field. For instance, a person might be asked, "Is object a to the right of object b?" 4 objects were assigned localist encodings. Pairings of these objects were made and they were positioned in two of the four object positions in the perceptual field. The description of the objects used 3 bits for each object and the encoded relationship (left/right) used two bits. The position of the objects in the perceptual portion of the input was allowed to vary across 4 spaces within the perceptual field. For instance:

examples			coding		
input		output	input		output
AxxC	a l c	t	0001 0000 0000 0100	001 01 011	10
DxBx	d r b	f	1000 0000 0010 0000	100 10 110	01
BAxx	a l b	f	0010 0001 0000 0000	001 01 110	01

Using this format 120 patterns were generated, and 100 of these were used in training a net that had 20 hidden units for 20K trials. After training, the average error was essentially 0.0. In addition, on the 20 transfer items there were no errors. A more complex simulation was run in which objects were positioned on a 2×2 grid and in addition to left/right relations, over/under was also trained. Using 6 objects with 6-bit localist codes and 3-bit distributed description codes, 240 patterns were generated, and 30 of these were saved as transfer patterns. After 80K training trials on 45-30-2 net, there was an average error of 0.014. Accuracy on the transfer task was 80%.

4. Pronoun Reference

One of the content-dependent problems in natural language processing is the identification of the correct referent of a pronoun.[11] Because much of the difficulty in pronoun reference is with the constraints of context, this appears to be a promising topic for a connectionist approach. In human language understanding the pronoun may produce activation of the appropriate noun in a sentence.[12] The procedure employed here required activation in the output of a pointer to the correct noun referent. The clues for disambiguation were relatively simple such as gender and number. There was no sentential pronominalization (e.g., Jack kissed Jill and she liked it.).

A vocabulary of 90 terms (19 singular and 19 plural nouns, 16 past-tense verbs, 9 pronouns, 9 adjectives, 6 proper nouns, 6 prepositions, 4 conjunctions, and 2 determiners) was used. From these terms, 2276 unique sentences were developed, and 20 were set aside as a transfer set. To optimize the learning of context, sentences were generated with multiple pronouns and referents placed in various positions. The pronouns differed in gender and number, and could appear either before or after the terms to which they referred.

The terms in the sentences were assigned random 8-bit binary codes. The sentences were a maximum of 15 words long and included at least one pronoun. At the end of the sentence, a probe pronoun was presented and the network was required to activate bits in an output vector which mapped to the position of the noun(s) to which the pronoun referred. For instance, the pronoun "it" refers to the word in the 5th position of the following example:

sentence	probe	output
the boy took the computer and dropped it on the floor	it	000010000000000

A 128-150-15 network was trained for 500K input/output pairings with η=0.05 and α=0.9. As a test of performance, two sets of transfer sentences were developed in which the network had to identify pronoun reference based solely on features that it had learned with the training set. One test set contained sentence frames of all possible paired combinations of 5 edible with 5 inedible objects and the pronoun referred to the edible object.

alice put the cookie on the table and she ate it	it	000100000000000
alice put the table on the cookie and she ate it	it	000000100000000

Overall, the edible objects received higher activation than the inedible objects, $F(1,8)=21.51$, $p<0.002$. In a second test set, the association of names by gender was tested in sentences such as:

mary gave john the computer and he smiled	he	001000000000000
mary gave john the computer and she smiled	she	100000000000000
john gave mary the computer and he smiled	he	100000000000000
john gave mary the computer and she smiled	she	001000000000000

In this case there was a main effect of order of presentation on the relative activations, $F(1,8)=13.13$, $p<0.007$, but more interestingly there was also an interaction which demonstrated that the network was sensitive to the

gender, $F(1,8)=49.66$, $p<0.00001$.

5. Natural Language Translation

Translation is often cited as a difficult natural language processing problem because it is believed to require an understanding of two languages as well as their coordination.[13] Several machine translation systems exist but they are generally difficult to program and modify, and moreover, they often require human intervention and post-editing. The most successful systems are those that deal with a limited domain[14] such as translation of meteorological forecasts or maintenance manuals.[15]

If translation is thought of as mapping from one language to another, the possible applicability of back-propagation becomes apparent. As a test of this approach a large number of English sentences and their Spanish translations were generated using a highly limited vocabulary and format. All sentences included a subject, verb, direct object, and indirect object. The verb was either 'to give' or 'to offer' and three different verb tenses were used (present, past, and past perfect). In the English sentences, the order of direct object and indirect object was randomly selected, while in the Spanish sentences the preferred sentence structure always places the indirect object after the direct object. Nouns referring to people (including two first names) and animals were used as subjects and indirect objects, while nouns referring to things were used for direct objects. Nouns were randomly modified by one of two adjectives.

English	Spanish
The grandfather offered the little girl a book	El abuelo le ofrecio un libro a la nina pequena

A vocabulary of 31 English words was used and each word was encoded in 5-bits. The longest English sentence was 10 words and shorter sentences were padded with nulls. The Spanish vocabulary consisted of 40 terms, which was more than the English primarily because differences in gender were required for adjectives and determiners. They were coded with 6-bit codes, and the longest sentence was 11 words long. Thus there were 50 input bits and 66 output bits. 3310 unique English sentences and their Spanish translations were generated. From this collection 33 sentence pairs were randomly selected as a transfer set and the remaining sentence pairs were used as a training set. Although most of the sentences were straightforward, a few seemed contrived.

The translation was not a simple mapping, because of differences in the number of words and reversals of adjectives/nouns. Three models were tested; the first was a normal back-propagation net with the hidden layer removed. This 50-66 net asymptoted quickly at an average error rate of 0.180. The second model was a 50-150-66 back-propagation with $\eta=0.01$ and $\alpha=0.9$. The average error on this model reached 0.070 after 1M pattern presentations. The third model was a multi-hidden-layer 50-150-150-150-66 back-propagation network also with $\eta=0.01$ and $\alpha=0.9$. The error rate on the multi-hidden-layer net fell below the level of the no-hidden-layer net (0.180) after 17K presentations. After 100K presentations the error rate was 0.102, after 1M presentations the error reached 0.036, and after 2M presentations the error was 0.027 and was still gradually decreasing. On the transfer set, the net without a hidden layer was incorrect on an average of 9.3 bits and 4.6 words per sentence. By comparison, the multi-hidden-layer net was incorrect on an average of 2.5 bits and 1.3 words, on the transfer set.

6. Summary

The first three studies show that these models can perform analogical reasoning, integration of information across different types of coding schemes, and context effects, while the fourth study suggests a new domain of application. Along with other research,[16] [17] these studies provide strong support for the applicability of connectionist approaches to linguistics.

In the present studies, especially the last two, the fact that the networks started with no semantic or syntactic knowledge and learned at least some of the regularities of language is remarkable. However, the results raise a number of issues that should be examined further. One set of issues concerns the encoding strategies that were employed. For instance, the encodings used in the translation study were extremely compact to

minimize the number of discriminations that needed to be made by the network. However it is possible that a coding scheme with more bits would have allowed greater flexibility of the network and would have resulted in fewer errors. In addition, the codes used in these studies were randomly paired with the terms they represented. This approach may be distinguished from work in which terms are encoded with explicit microfeatures, e.g.[4] While random encoding was employed here to match the essentially random pairings of words and their meanings in natural language, it would have been possible to use a hierarchical or similarity-based encoding scheme. For instance, the output side of the translation net could have employed an encoding based on parts of speech, gender, number, and so forth.

While the some success was achieved in learning with these relatively small problems, a crucial question is how performance would be affected when more complex patterns are used as inputs. One possibility is that the nets would be rapidly overloaded; however, another possibility is that exposure to a wider variety of examples may provide greater opportunities for generalizations and stability.

Extensions of the procedures described in this paper, especially the translation study, might investigate network architectures. For instance, the results in Section 5 suggest that the multi-hidden-layer net is superior to nets with either 0 or only 1 hidden layer. However, the multi-hidden-layer net also has many more model parameters and it would be of interest to investigate whether the layering itself or just the large number of parameters account for the difference. Other network architectures could also be considered. For instance, auto-associator networks[18] might be applied to sentences in each of the two languages and, hopefully, these would extract the essential features of the sentences. A back-propagating link could then be built between the hidden layers of the auto-associators so that features in one language could be mapped to the features of the other language. In a sense, this is similar to the usual approach to machine translation (and presumably to human translation) in which each language is mastered before translation is attempted[19] However, preliminary results with this type of multi-net have been poor; apparently the types of features extracted in the two auto-associator networks are not easily coordinated. Variations of the procedure in which learning on the networks is interleaved are presently under investigation.

Another network architecture that might be applied to language processing models, such as translation, would incorporate time-averaging on some unit activations, e.g.[20] This approach would appear to be more like natural human language use than the parallel input employed in sections 4 and 5. However while temporal proximity may be a factor in some aspects of language use,[11] the possibility of embedded phrases and clauses in sentences suggests that some type of buffering or intermediate representation must be used, and that simple time-dependent models are inadequate for natural language.

A general issue for training connectionist networks concerns the importance of multiple constraints for language learning. Perhaps it is surprising that the networks studied here learned some of the regularities of language given only one, or at most two, types of input codes. Of course, most human language learning combines many kinds of information.[21] Thus, multiple types of input information (see Section 3) may improve the robustness of learning of language by networks as well. A final, perhaps rhetorical, issue concerns the point at which artificial neural networks could be said to 'have language.' This may be related to enduring arguments about the definition of language.[22]

A great deal of research remains in integrating, refining, and extending results such as those reported in this paper for the development of a connectionist linguistics. Clearly, however, connectionist approaches appear quite promising.

Acknowledgment

I thank the members of the Bellcore neural net group for a stimulating environment in which these ideas were developed.

References

1. Barwise, J. & Perry, J. *Situations and Attitudes.* MIT/Bradford, Cambridge, MA, 1983.

2. Searle, J.R. *Speech Acts: An Essay in the Philosophy of Language.* Cambridge University Press, Cambridge, 1969.

3. Cottrell, G.W. Parallelism in inheritance hierarchies with exceptions. *Proceedings IJCAI,* 1985, 194-200.

4. McClelland, J.L. & Kawamoto, A.H. Mechanisms of sentence processing: Assigning roles to constituents of sentences. In: McClelland, J.L. & Rumelhart, D.E. (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition (Vol. 2).* Cambridge, MA, Bradford Books/MIT Press, 1986.

5. Rumelhart, D.E., Hinton, G.E., & Williams, R.J., Learning representations by back-propagating errors. *Nature,* 1986, *323,* 533-536.

6. Collins, A.M. & Loftus, E.F. A spreading activation theory of semantic processing. *Psychological Review,* 1975, *82,* 407-428.

7. Sternberg, R.J. *Intelligence, Information Processing, and Analogical Reasoning.* Erlbaum, Hillsdale, NJ. 1977.

8. Cottrell, G. Toward a connectionist semantics. *Proceedings of Theoretical Issues in Natural Language Processing - 3.* Las Cruces, NM, 1987.

9. Lakoff, G. & Johnson, M. *Metaphors We Live By.* Chicago, University of Chicago Press, 1980.

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

11. Hobbs, J.R. Resolving pronoun reference. *Lingua,* 1978, *44,* 311-338.

12. Corbett, A.T. & Chang, F.R. Pronoun disambiguation: Accessing potential antecedents. *Memory and Cognition,* 1983, *11,* 283-294.

13. Slocum, J. A survey of machine translation: Its history, current status, and future prospects. *Computational Linguistics,* 1985, *11,* 1-17.

14. Kittredge, R. *Sublanguage: Studies of Language with Restricted Semantic Domains.* New York, Walter de Gruyter, 1982.

15. Isabelle, P. & Bourbeau, L. TAUM-AVIATION: Its technical features and some experimental results. *Computational Linguistics,* 1985, *11,* 18-27.

16. Hanson, S.J. & Kegl, J. PARSNIP: A connectionist network that learns natural language grammar from exposure to natural language sentences. *Proceedings of the Cognitive Science Society,* 1987.

17. St. John, M.F. & McClelland, J.L. Reconstructive memory for sentences: A PDP approach. *Proceedings Ohio University Inference Conference, 1986.*

18. Ackley, D.H., Hinton, G.E., & Sejnowski, T.J. A learning algorithm for Boltzmann machines. *Cognitive Science,* 1985, *9,* 147-169.

19. Macnamara, J. & Kushnir, S.L. Linguistic independence of bilingualism: The input switch. *Journal of Verbal Learning and Verbal Behavior,* 1972, *10,* 480-487.

20. Jordan, M.I. Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Cognitive Science Society*, 1986, 531-546.

21. Miller, G.A. & Johnson-Laird, P.N. *Language and Perception.* Cambridge, MA, Harvard, 1976.

22. Premack, D. *Gavagai!* MIT Press, Cambridge, MA, 1986.