C E N T E R F O R R E S E A R C H I N L A N G U A G E

July 1995 **Vol. 9, No. 3**

The Newsletter of the Center for Research in Language, University of California, San Diego, La Jolla CA 92039
Tel: (619) 534-2536 • E-mail: crl@crl.ucsd.edu • WWW: http://crl.ucsd.edu/newsletter.html

• • •

FEATURE ARTICLE

Connectionist Modeling of the Fast Mapping Phenomenon

Jeanne Milostan
Department of Computer Science and Engineering, UCSD

ANNOUNCEMENTS

WWW: The CRL Newsletter is now on World Wide Web at URL: http://crl.ucsd.edu/newsletter.html

EDITOR'S NOTE

This newsletter is produced and distributed by the **CENTER FOR RESEARCH IN LANGUAGE**, a research center at the University of California, San Diego that unites the efforts of fields such as Cognitive Science, Linguistics, Psychology, Computer Science, Sociology, and Philosophy, all who share an interest in language. We feature papers related to language and cognition (1-10 pages, sent via e-mail) and welcome response from friends and colleagues at UCSD as well as other institutions.

SUBSCRIPTION INFORMATION

If you are currently receiving a hardcopy of the newsletter and have access to e-mail, please help us save printing and mailing costs by forwarding your e-mail address to CRL. If you require a hardcopy in addition, please request it and we will be happy to send you one.

If you know of others who would be interested in receiving the newsletter, please forward the e-mail or postal mailing address. Please forward correspondence to:

Jay Moody, Editor
Center for Research in Language, 0526
9500 Gilman Drive, University of California, San Diego 92093-0526
Telephone: (619) 534-2536 • E-mail: crl@crl.ucsd.edu

Back issues of this newsletter are available from CRL in hard copy as well as soft copy form. Papers featured in previous issues include the following:

Language and the Primate Brain
Martin I. Sereno
Department of Cognitive Science, UCSD
vol. 4, no. 4, August, 1990

The Segmentation Problem in Early Language Acquisition
Kim Plunkett
University of Aarhus, Denmark
vol. 5, no. 1, November, 1990

Neo-structuralism: a commentary on the correlations between the work of Zelig Harris and Jeffrey Elman
Peter Bensch
Department of Computer Science, UCSD
vol. 5, no. 2, March 1991

Preposition Use in a Speaker with Williams Syndrome: Some Cognitive Grammar Proposals
Jo Rubba and Edward S. Klima
Department of Linguistics, UCSD
vol. 5, no. 3, April 1991

Middle-Subjunctive Links
Ricardo Maldonado
Department of Linguistics, UCSD
vol. 5, no. 4, May 1991

Zai and Ba Constructions in Child Mandarin
Ping Li
Center for Research in Language, UCSD
vol. 5, no. 5, July 1991

Hitting the Right Pitch: A Meta Analysis of Sentence Context on Lexical Access
Mark St. John
Department of Cognitive Science, UCSD
vol. 5, no. 6, August 1991

What is Who Violating? A Reconsideration of Linguistic Violations in Light of Event Related Potentials
Marta Kutas
Departments of Cognitive Science and Neurosciences, UCSD
Robert Kluender
Department of Linguistics, UCSD
vol. 6, no. 1, October 1991

Learning to Recognize and Produce Words: Towards a Connectionist Model
Michael Gasser
Department of Computer Science, Indiana University
vol. 6, no. 2, November 1991
PDP Learnability and Innate Knowledge of Language
David Kirsh
Department of Cognitive Science, UCSD
vol. 6, no. 3, December 1991

Why It Might Pay to Assume That Languages Are Infinite
Walter J. Savitch
Department of Computer Science, UCSD
vol. 6, no. 4, April 1992

Mental Spaces and Constructional Meaning
Claudia Brugman
Center for Research in Language, UCSD
vol. 7, no. 1, October 1992

Is Incest Best? The Role of Pragmatic Scales and Cultural Models in Abortion Rhetoric
Seana Coulson
Department of Cognitive Science, UCSD
vol. 7, no. 2, January 1993

Marking Oppositions in Verbal and Nominal Collectives
Suzanne Kemmer
Department of Linguistics, UCSD
Vol. 7, no. 3, September 1993

Connectionist Representations: The State of the Art
Daniel Memmi
LIMSI-CNRS
Orsay (France)
Vol. 8, no. 1, November 1993

Abstract Better Than Concrete: Implications for the Psychological and Neural Representation of Concrete Concepts
Sarah D. Breedin, Eleanor M. Saffran, and H. Branch Coslett
Center for Cognitive Neuroscience, Department of Neurology, Temple University
Vol. 8, no. 2, April 1994

Analogic and Metaphoric Mapping in Blended Spaces: Menendez Brothers Virus
Seana Coulson
Department of Cognitive Science, UCSD
vol. 9, no. 1, February 1995

In Search of the Statistical Brain
Javier Movellan
Department of Cognitive Science, UCSD
vol. 9, no. 2, March 1995

Connectionist Modeling of the Fast Mapping Phenomenon

Jeanne Milostan
Computer Sceince and Engineering, UCSD

1 Introduction

The average child learns some 14,000 words before the age of 6, which represents the daunting task of acquiring 9 new words per day, or about one each waking hour [2]. Researchers examining the process by which this is accomplished have time and again encountered an interesting effect: often the child can acquire a new word from only one or a small number of exposures to that word. Susan Carey has dubbed this phenomenon "fast mapping."

In this paper we examine the research which has been done on the manifestation of fast mapping in children and explore how this may be explained in terms of a general cognitive model of language acquisition. We then examine a number of basic and advanced connectionist models and systems and weigh how each stands in relation to describing and explaining the fast mapping behavior. We then speculate on what is missing from the constellation of models available and propose directions for future research in this area.

2 Fast Mapping

2.1 Empirical Demonstrations

Susan Carey [2] began by asking the question "What is learned when a word is added to a child's vocabulary? Where does the process of word learning begin?" In her study, she examined the preschool child's limits on word learning capacity. The study tested the acquisition of a novel word representing a color -- *chromium*. After demonstrating that none of the children in the study (age 3 to 4) had a separate name for the color olive (each identified it as green or brown), the experimenter presented the word *chromium* to each child in the context of a task request: "Please hand me the chromium cup; not the red one, the chromium one" where the choice was between one red cup and one otherwise identical olive (*chromium*) cup. Carey found that given only one exposure to the color name, upon comprehension testing one week later 9 of the 14 subjects successfully identified either an olive or a green color chip when asked to point to the *chromium* one. Additionally, during a production test 6 weeks later, 8 of the 14 subjects answered differently when asked to name the color chip than they had before the experiment began. That is, where they had originally named the chip green or brown, they now said they didn't know the name or used another unstable color referent from their vocabulary, thus indicating that they had learned and retained the knowledge that olive has its own color name.

Additionally, Carey found that the children who learned the name after the brief exposure could take two different tacks. For some, the False-Synonym group, *chromium* was used as another word for green. Other children adopted the Odd-Color-Odd-Name strategy; these children demonstrated comprehension of the word, but for production named another color from their lexicon which also did not have a stable referent, thus again demonstrating they knew that olive had a separate name.

Earlier, Nelson and Bonvillian [13] had performed a study in which children were exposed to 18 new concepts, of which 9 were made-up words and 9 were were actual English words which the children had not yet acquired (7 control children also did not acquire these words by the end of the study). In a series of 10 experimental sessions, the children were presented with examples while every third session an unnamed exemplar was used to test comprehension. Comprehension was tested both by asking for the object by name, and by holding up the example and asking "Bring me one of these." This study demonstrated that the child *could* acquire the name from a single example, but that learning was more likely when two or four named exemplars were encountered.

In examining the question of what characteristics of language are dissociable, Bates et al. [1] also performed a study examining the acquisition of a novel concept in young children. In this study, a novel object was given both a novel name ("fiffin") and a novel associated action ("glooping"). In an initial 5 minute exposure conducted in the home, the children were shown several fiffins and glooping was demonstrated. In a lab session 2-3 days later, comprehension was tested through a multiple choice test and in a play session: "Make the kitty gloop the fiffin." Of the 23 subjects, 9 performed the gesture successfully in the home, while 18 did so in the lab. 8 subjects also made successful verbal attempts at pronunciation in the home; 9 did so in the lab. During the multiple choice test, the average score was 75% correct, where 33% was chance. Additionally, 18 kitties successfully glooped. This study demonstrated again that children can obtain a concept after an extremely brief exposure, and that it was not necessary to perform imitation to obtain the concept, as many demonstrated lab comprehension without acting out in the home. Additionally, Bates showed that the type of knowledge the child demonstrated was correlated with language "style"; that is, fiffin comprehension was related to early

comprehension, while fiffin imitation was related to early production.

Mabel Rice [16] addressed word acquisition from television viewing, thus offering evidence that neither lexical acquisition nor fast mapping in particular are limited to interactive exchanges. In one study, Rice exposed a number of children to short cartoon segments which were designed to introduce new words. The test words in this case consisted of actual English which the children did not already have in their vocabularies, and which included a number of words which were not object names or attributes. In all, each subject was exposed to 20 new words in a brief time; each 12-minute cartoon presented several instances of a few new words, 114 presentations over all words total. From this exposure, the 5-year-old subjects gained an average of 4.87 words as compared to controls, while the 3-year-olds gained an average of 1.56 words. This study demonstrated that new words need not be contained in exaggerated, referent-matching contexts in order to be acquired, and that the new word need not be surrounded exclusively by familiar words. Additionally, this study demonstrated that words other than object names and attributes were also subject to fast mapping, and that the new words need not be presented in the exact same context each time in order to learn. Rice did additional work [17] in a more naturalistic home environment, where it was demonstrated that children learn new words rapidly from educational programs such as "Sesame Street" even with the environmental distractions associated with home television viewing.

In a study intended to explore what aspects of a word are developed upon fast mapping, Chris Dollaghan [3] tracked acquisition and use of a nonsense word, "koob". This word was introduced in a naturalistic setting; the experimenter asked the child (age 2:1 to 5:11) to "Hide the koob under the bowl" rather than explicitly stating "This is a koob." The experiment was constructed so that the child could actually perform the task requested without forming any theory of the name of the intended object. After one exposure to the word, the subjects were later tested for comprehension ("Hand me the koob"), production ("What's this?"), recognition ("What is this? Is it a koob, soob, or teed?") and association with location ("Where did you hide this?"). In most cases, an immediate inference between the unfamiliar word and object was made, although the extent to which that knowledge was available for use varied considerably from child to child.

2.2 Manifestations, Modulations, Limitations

The above studies and many others demonstrate that the fast mapping phenomenon is a real, robust occurrence which appears over a variety of situations. The amount of fast mapping varies with age; in particular, subjects who are too young do not show much learning. Learning occurred more readily over a broad base of examples rather than a narrow base (only 1 example). Fast mapping is robust across method; children successfully acquired words from limited exposure whether the presentation was by an experimenter, the child's mother, or the television. It is robust over distraction, as demonstrated in the unfamiliar environment of a laboratory or in the distracting environment of television viewing in the home, with its associated sibling, parental and play distractions. Fast mapping is robust across linguistic method of presentation; the effect was present for words presented in incidental naming, explicit presentation ("This is a ... '), and in sentences both where the surrounding words were all familiar and when they contained other unknown words.

The amount and manifestation of the effect was seen to vary with gender, age, and cognitive style -- whether the child favors one word, telegraphic speech versus whole-phrase speech. Additionally, the effect varied with birth order and sibling constellation. Nelson and Bonvillian [13] found that children whose next-older sibling was less than 24 months older gained the most words, with first-born children close behind while lagging last were those children whose next-older sibling was more than 24 months older. Nelson and Bonvillian hypothesize that the first-born children have the added advantage of parents who have more time to spend, and thus are exposed to more explicit referential sentences and more parental time overall. Short-lag children lose the benefit of total parental attention, but are helped to a greater extent by the presence of an older sibling whose speech is more like their own than like the parents'. That is, the short-lag child receives more predigested and simplified examples of speech on which to bootstrap; some of the processing has already been done and the short-lag child can leverage off this benefit. Conversely, the longer-lag children do not have this benefit, and also do not receive parental attention to the extent that first-born children do.

2.3 Theory

Rare Event Cognitive Comparison Theory

In [14], Nelson explores how current language and cognitive levels facilitate and limit what will be learned next. The overall acquisition mechanism depends on cognitive comparisons between old and new structures in order for the child to determine when the current language structure is insufficient. The mechanism is seen as a "rare event" mechanism, as the attention to new input which leads successfully to the development of new structures for the child's

future use occurs only rarely. The development of a new structure occurs along the following lines:

1. Assignment of old structure to new input strings. As long as new input matches the structures already in use, the system need not change.

2. Tentatively Abstracted Foci. Something happens to draw attention to some area of the structure, to create a "hot spot" of attention. This may occur because a number of mismatches of new input strings have drawn attention, or simply because the child's existing structures have developed to a certain extent which prepares for a new structure. In this way, developments can bootstrap, as a child may not be ready for a particular structure until other supporting structures have been laid out first.

3. Finding input mismatches within Tentatively Abstracted Foci. Once attention is drawn, mismatches will be more readily noticed and attended to.

4. Selective Storage. Certain strings of interest will be stored, perhaps in episodic memory.

5. Selective Retrieval. With attention to mismatches, previously encountered examples can be retrieved for comparison with the detected mismatch. Note that language advances can thus be made during private thinking, as the child retrieves example strings from memory and mulls them over alone.

6. Selective Analysis. The child considers the newly collected data.

7. Selective Hypothesis Monitoring and Consolidation. A conclusion is reached and new structures are tentatively created. Previously encountered and new input strings are compared against the new structure, which is eventually consolidated into the child's language structure.

From this point of view, one can see how input exposure will affect the child's particular path to language mastery. Different types of input will cause individual children to call into question various structures at different times. The particular structures the child attends to will determine the path of acquisition the child takes. The issue of birth order mentioned above can be cast in this light; input from a slightly-older sibling is more like the child's own production, thus the differences are small and more easily attended to, allowing the younger child to ride on the coat tails of the older sibling's language efforts. Similarly, first-born children tend to get more explicit input from parents, and thus again attention is more readily drawn.

A Tentative Approach

The fast mapping phenomenon may then be cast in the light of the preceding information. One may envision a protracted "hot spot" of attention to word naming which the child encounters, perhaps driven by the root cause of the above elaborated system. Based on the data collected from the various studies, we may draw the following conclusion: Fast mapping in children and the resultant characteristics of the word which the child thus obtains are affected by area of attention and input amount and style, and the use of episodic memory to integrate and store the information. It is reasonable to hypothesize that a language acquisition model which incorporates these elements may also demonstrate the fast mapping phenomenon.

3 Episodic Memory

Human memory is not performed by a single mechanism, but consists of several different functionally and physically distinct components. To simplify, a distinction may be drawn between what can be termed declarative memory, that of explicit facts and events, and nondeclarative memory, which is involved in such things as habit formation and priming. Only the information in the declarative memory can be consciously recalled, and it is this part of memory which is of concern when addressing one-trial lexical acquisition.

Lesion studies point out the essential role of the hippocampus and surrounding structures in the operation of declarative memory. Again to simplify, the hippocampus is involved in processes which bind together previously unrelated events (represented in different parts of the brain), which then together constitute a memory of the event in question. Additionally, the hippocampus also participates in forming in neocortex an integrative trace of a newly formed memory, possibly through a feedback loop between neocortex and the hippocampus. That is, the hippocampus develops and maintains a temporary "trace" of the formed memory while a more permanent one is formed elsewhere in the brain.

Thus it becomes clear that in order to adequately model the process of lexical acquisition in a more realistic manner, and to thus develop a system which will naturally manifest the fast mapping behavior, it is necessary to develop a system which addresses the issues of attention, episodic memory, short-term memory to long-term memory conversion, and the

ability to generalize similarities and still handle very novel input gracefully.

4 Connectionism

In this section we look at previous work which has been performed, with an eye to systems which may result in fast mapping. We examine standard network architectures, followed by specific systems which have attempted to emulate episodic memory or lexical acquisition in general, or the fast mapping phenomenon in particular. This is followed by an examination of a handful of larger systems which attempt to more adequately address human performance issues.

4.1 Basic Connectionist Models

Backpropagation

The backpropagation neural network [19] is a multilayer architecture consisting of interconnected layers of processing units. Input vectors are presented to the elements of the input layer, and activation is propagated through the network to the output layer. During training, the values at the output layer are compared to the actual desired output associated with the given input vector. Any error in the network output is used to calculate adjustments to the network weights using a gradient descent technique. The overall effect is that over time, the network weights adjust to form a representation of the function described by the set of input-output vectors presented. The backprop network is often able to form a generalization of the function, rather than a simple mapping-and-recall of the input-output pairs. This generalization is often desirable, in that inputs which are similar to learned data will receive outputs which are similar to the learned patterns. Unfortunately, often truly novel inputs will also be given a generalized output, rather than the specific output to which it is matched. This has some utility in modeling overgeneralization in language acquisition, but does not function properly for the acquisition of novel concepts.

Unfortunately, as powerful as backprop is, it is not suitable for modeling the fast mapping phenomenon. If a new input is presented to a network which has already been trained, it is possible that the representation the network has developed is not suitable to generate the proper output for the input. In this case we would like to perform additional training on the network to incorporate the new data. Unfortunately, presentation of the new input-output pair to the network for only a few training cycles may not be sufficient to adjust the weights to properly represent the new information. Simply adding the new data to the existing training set and continuing training will require many training episodes for the network to develop a representation of the new data; it will not display the rapid learning desired. One also may attempt to force the weights in the network to make a large adjustment in the direction indicated by the new data; however, this technique runs the risk of losing previously learned associations as the network may move too far in that direction. Either way, the network will take too long to learn or will not learn well enough to model fast mapping. (But see Section 4.5 below on some possibilities afforded by recurrent networks, i.e. networks which allow self connections or backward connections.)

Autoassociative Memory

The autoassociative neural network is actually a large family of paradigms, all of which have in common the association of an input vector with itself. One very useful member of this class is the Kohonen network [7]. In this model, the network consists of a number of processing elements each of the same dimensionality as the input. Through training, the values in the element vectors are adjusted so that they come to represent the space of possible input vectors (as represented by the examples given during training). Training this network consists of identifying the processing element which lies closest to the input vector, and adjusting the element vector towards the input vector by some fraction of the distance between them. With the addition of "neighborhood" links, in which each element is connected to additional processing elements which will form its neighbors, the network will form a topological map of the space of input data. A common use of the "neighbors" is to adjust all those elements in the closest element's neighborhood toward the input vector also, by a smaller amount than the winner adjustment. Through a very large number of training presentations the processing elements come to reflect the spatial representation and extent of the training input. An example representation of a network which has been trained to represent an even distribution of points in the unit square is shown in Figure 1. This type of network, frequently called a topological map or feature map, is often used as a memory of examples seen, and as such may be a candidate for representing lexical memory in humans. Since the network is self-organizing and topological, it will develop areas of common information which can be seen as representing a lower dimensional projection of the main information represented by the network. However, since the mapping is continuous, the

File Name : topomap.ps
Title : noname.ps
Creator : pnmtops
Pages : 1

Figure 1: Autoassociative Topological Map

precise boundaries of the various categories developed are not specified.

This feature map paradigm has several properties which make it less than adequate for representing fast mapping. The most obvious weakness is that the network always returns as the winner the vector of the element which is closest to the presented input. For generalization, this is a desired trait in that one will always be presented with a representative vector which will be identical or similar to an actual input from the training set, or some blended combination of inputs. The problem comes when a novel input is presented which is very much unlike those previously seen. In normal operation, that element vector which is closest to the input vector will be returned as a "memory" of the input, regardless of the actual distance to the input. The network does not take into consideration the actual distance from the nearest vector, nor the typical distance between vectors in the trained network. One may wish to use the distance between the input vector and the closest processing element as an indication of whether the input is correctly categorized by the network. However, absolute error is not an adequate measure because there is no threshold for a decision of "don't know." The distance threshold will vary between individual processing elements; an element in a densely populated area of input space will encompass a much smaller area for its valid inputs than an element in a sparsely populated area. The network has no notion of representing "don't know" or of flagging that the input is extremely different.

Additionally, a problem still lies in modifying the network to incorporate the new information. Addition of a truly novel input may deform the network severely, with performance returning only gradually through continued training over the entire training set. This is clearly not an adequate model of fast mapping. The standard implementation of autoassociative feature maps will not adequately model the fast mapping phenomenon.

4.2 Attempts to Address "Fast" Mapping

Fast Weights

Hinton and Plaut [6] modified a standard backpropagation network to have two connections between each unit: one with a slow, stable weight and one with a fast, elastic weight. The slow weights function much as they would in a regular connectionist model: they change slowly and hold the

long-term knowledge of the network. In contrast, the fast weights change rapidly and continually decay toward zero, and thus reflect only the recent past. The effective connection between two units is the sum of the fast and slow connections. At any time, the system's knowledge can be thought of as the slow weights with a temporary overlay of the fast weights.

This system could be used for for rapid temporary learning. In other words, when presented with a new association, the network could conceivably store the information in one trial. Although this addresses the issue of a backprop not being able to rapidly integrate new information, this setup does not address the possibility of previous knowledge being obscured by the new addition. This solution is obviously better than forcing a single set of weights in the direction of the new information, as the previous knowledge is not lost; however previously learned associations may still be unavailable while the temporary weights are in place.

Additionally, although it is easy to train the fast weights for the desired one-shot learning effect, it is not clear how to incorporate the new associations gracefully into the slow weights for long-term storage without the traditional drawbacks of continued training with the entire training set plus the additions. Thus, this system does not adequately meet our needs for fast mapping as seen during language acquisition.

CHARM

The CHARM (Composite Holographic Associative Recall Model) system developed by Janet Metcalfe [10] [9] [11] uses a mathematical technique similar to that used in holography to form an associative system which can be rapidly updated through the operations of convolution and correlation. The use of convolution for association results in the interaction of all of the parts of one item with all of the parts of another. The system is presented input/output pairs represented as feature vectors to be associated. Through various mathematical transformations, the input is associated with the output, and their total is combined with the results of other pairs into a large system representation. For recall, the input is presented to the whole system, and further mathematical machinations are performed, resulting in a vector intended to represent the output of the original pair. Due to the nature of the mathematics, this system shows one-shot learning. That is, upon one presentation of a pair, the association is contained in the system. This shows much more promise in the modeling of fast mapping than the models considered thus far, but as noted above, in practice the fast mapping phenomena does not occur every time, nor does a successful fast mapping imply that the concept has been obtained in its entirety. If the CHARM model were an accurate representation, more cases would be seen of concepts springing fully-formed from the little wizards' minds, as it were.

However, CHARM does have to its credit the ability to model quite a number of other psychological phenomena, including generalization and a number of memory interference and failure effects. This system holds much promise in its future applicability as a model of fast mapping.

4.3 The DISCERN Model

Description

The DISCERN model (DIstributed SCript processing and Episodic memoRy Network), developed by Risto Miikkulainen [12], is a distributed artificial neural network system which learns to process simple stories which follow a stereotypic framework. As such, it combines the traditional symbolic artificial intelligence paradigms of scripts and frames with more realistic cognitive modeling and neurocomputation methodology. This model combines the issues traditionally associated with script-based story understanding and adds to it the idea of episodic memory. There are several issues which the symbolic approach to script theory does not address. For instance, the architecture, processing mechanisms and knowledge embedded in symbolic systems are hand-coded with a specific domain and data in mind. Inferences are based on handcrafted rules and representations of the scripts. Such systems cannot utilize the statistical properties of the data to enhance processing.

One thing which the DISCERN model brings to the story understanding task is the idea of episodic memory. Narratives are stored in the model one at a time as they are read in, with only a single presentation. The new story is recognized as an instance of a familiar sequence of events and attention is paid only to the facts specific to the story, even though the system has not gone back and explicitly reactivated all the stories previously encountered. This parallels human episodic memory, which seems to be structured to support classification based on similarities and storing the differences, with the particular structures being developed by experience.

The episodic memory structure of DISCERN also supports associative retrieval. As in humans, a question supplies only partial information about the story to which it refers, yet the story is retrieved with only the question as a cue. The DISCERN model has been developed to address these issues. It is the

File Name : pyramid.ps
Title : noname.ps
Creator : pnmtops
Pages : 1

Figure 2: Hierarchical Feature Map

implementation of episodic memory which is of interest for the purposes of this paper.

Episodic Memory Implementation

The DISCERN model implements episodic memory as a collection of traces on a hierarchical feature map system. As described above, a self-organizing feature map (autoassociative network) is a (biologically-motivated) method for unsupervised learning and for organizing information. The feature map representation has many properties which make it well-suited for modeling memory. Classification performed by a feature map is quite robust, even in the presence of noise or incomplete inputs. Categorical perception can be thus modeled, since inexact input often results in the recovery of the exact representation of previously stored data. In contrast, since the feature maps tend to be continuous with intermediate states, it is possible in some cases to recover a blend of a number of items. However, as also mentioned above, the feature map representation suffers from the drawback that boundaries of related areas are not specified on the map. Additionally, feature maps created from high-dimensional input vectors take a long time to train.

These drawbacks can be addressed to some extent with hierarchical feature maps. In this case, the hierarchical nature of the input features are represented by a pyramid of feature maps. This speeds the learning of the system, and makes categorization easier. In this setup, the input features are initially classified by the uppermost map. The vector is then passed down to subsequent maps for more detailed classifications (Figure 2).

The episodic memory storage and retrieval is implemented in the system as trace feature maps on the hierarchical map structure. Trace feature maps differ from ordinary feature maps by creating a memory trace at the location of classification on the map. The map remembers that at some point it received an input item which was classified at that point. The traces can be stored one at a time, and the whole of the traces over an episode constitute the memory of events. The traces are modeled by using the "neighborhood" links of the feature map as activity links to develop basins of activation. The attraction bubbles created by the various memory traces are then superimposed and blended. Upon memory recall, a partial or noisy input is presented to the system. If it falls within an attraction bubble, the activation will be drawn towards the center of the bubble, and the stored vector associated with the center will be returned. In this case, the input vector could represent a question for the system, with the unspecified features representing the unknown roles, which would then be filled in through returning the center vector of the specific instance activated.

File Name : gestalt.ps
Title : noname.ps
Creator : pnmtops
Pages : 1

Figure 3: Sentence Gestalt Network Model

In terms of representing episodic memory, this system performs well. New stories presented to the system develop a memory trace which is robust in a small number of presentations, and thus models the "fast" part of fast mapping without resorting to an artificial, "guaranteed one-shot" learning mechanism. The system demonstrates a number of memory phenomena such as interference effects and generalization. The structure of the system does not overly constrain how the information in the memory is to be organized, and thus the system with use comes to reflect the statistical properties of the data it has seen.

However, like the Kohonen Feature Maps reviewed earlier, the system suffers from the limitation that it cannot learn truly novel information. That is, although it can successfully represent stories on which it was not originally trained, stories which are extremely unlike those seen during training will not be handled correctly. Several suggestions for extensions to the system (including those suggested by the author) addressing this limitation will be examined in Section 5 below.

4.4 Attentional Mechanisms

The "sentence gestalt" model of St. John and McClelland [20] was developed as an attempt to create a model which learns to convert a sentence to a conceptual representation of the event which the sentence describes. The model is intended to disambiguate ambiguous words, instantiate vague words, assign thematic roles, and elaborate implied roles. In addition, it is required to learn to perform these tasks, and perform them on-the-fly as the sentence is presented, rather than waiting until the sentence is finished and then performing calculations. The model is a mostly feed-forward network with a number of hidden layers and a small amount of recurrence (Figure 3).

The model performs rather well at its assigned tasks, and is able (through the "probe" inputs) to answer questions about the representation of a sentence it has developed. It also demonstrates a number of appropriate phenomena such as generalization, interference and priming effects, and frequency effects.

For the purposes of this paper, the most interesting property of the sentence gestalt system is that it effectively develops an attentional mechanism. That is, the system must learn through example which parts of the sentence are important for providing which types of information. The system learns to make appropriate balances between word order and semantic constraints for determining the meaning and roles of words in a sentence, for example, without this knowledge being otherwise coded into the system.

4.5 Generalization and Novelty

Although the linguistic processing model developed by Plaut et al. [15] focuses mainly on learning to read (bold connections in Figure 4), the system they have developed demonstrates some interesting behavior which may be applicable to modeling fast mapping. Plaut and his co-authors develop a recurrent network which learns to map orthography, the printed letters of a word, to phonology, the phonetic representation of the word. Their effort has produced a system which not only performs the mapping task, but successfully demonstrates the frequency versus consistency effects shown by human subjects and additionally shows performance following damage which parallels the language difficulties of surface dyslexic patients.

The interesting behavior of the model in terms of the fast mapping task is the behavior of the recurrent network in the face of novelty. It is of some concern when using recurrent networks that due to the dynamics of the attractor surface represented by the system weights, novel inputs will be treated as "incomplete" or "noisy" data and subjected to the generalization behavior of the network. However,

File Name : fig1.ps
Title : /tmp/xfig-fig000210
Creator : fig2dev
CreationDate : Thu Nov 17 20:30:35 1994
Pages : 1

Figure 4: Linguistic Processing Framework

their network develops basins of attraction which interact like ripples in a pond to create additional attractor basins for data which has not actually been presented to the network. For instance, even if the network has only seen evidence for *by* (mapping to the sound /bI/) and *no* (mapping to the sound /no/) the network may also form an attractor basin into which *bo* would naturally fall (i.e. /bo/). These extra basins can be shown to be a natural consequence of having a highly connected, high dimensional dynamic space.

Note that in this case, the network demonstrates fast mapping. That is, even though the network had not been trained on the mapping between the letters *bo* and the sound /bo/, the network correctly made the mapping. The network has an appropriate attractor basin for this mapping, and the system provides just enough of a nudge to enter the basin and converge to the mapping. At this point, any training on this specific example will serve to deepen and expand the attractor basin, thus ensuring that the mapping will be made more readily (in this case, fewer steps until convergence) in the future. This rapid initial mapping followed by subsequent strengthening of the learned associations is exactly the phenomenon which we seek. The use of this type of network in the processing of learning to speak has been anticipated by Plaut et al. as represented by the dotted line in Figure 4, although this use was not addressed directly in their paper.

5 What's Missing; What's Promising

None of the models explored adequately model fast mapping (nor linguistic acquisition in general) in a way which is satisfactory to represent a model of human performance. However, several of the systems show promise, which may be exploited through various changes. Using these modified systems, an overall connectionist system can be developed which may indeed display the desired fast-mapping phenomenon, while still producing overall behavior which is consistent with other aspects of human language acquisition. The proposed system combines aspects of the DISCERN model to represent episodic memory, the sentence gestalt network to provide an attention mechanism, and a recurrent network as described above to represent long-term memory.

The DISCERN model [12] representation of episodic memory has as its largest drawback an inability to represent truly novel inputs due in part to its basis in symbolic script theory but mostly due to the nature of the autoassociative networks used. However, as suggested by the author, modifying the episodic memory to provide dynamic recruitment of new units to the network as needed would address this problem, with additional reorganization training conducted between input episodes. This can be seen as an implementation of the structure building theory examined in section 2.3, with offline restructuring paralleled by language development which occurs during the child's private play. We propose also that if the input feature vectors, rather than being handcoded to represent the scripts, were learned by an additional network system, this network would become a more accurate model of episodic memory.

The sentence gestalt model developed by St. John and McClelland [20] is an ideal candidate for the role of just such an additional network. As described above, this network has demonstrated the ability to develop a form of attentional mechanism. We propose that a network similar to that of St. John and McClelland be used to determine which input features deserve the most attention. These feature vectors may then be used as the basis for a system similar to the DISCERN model.

Finally, we propose a recurrent system similar to the one used by Plaut et al. [15] to represent the long term memory. This type of system provides the necessary generalization and ability to represent novel inputs which is necessary for the representation of memory.

In this model, the sentence gestalt network/attention mechanism would steer the focus of the network to features of interest, where the interest would itself be defined by the attention mechanism and would evolve over the course of the simulation. The gestalt of the inputs and features of focus would then be sent to the episodic memory network, where the incoming information would be incorporated into the episodic representation, recruiting units as required to

represent novel information. As mentioned above, the episodic memory will be in a state of continual reorganization; once the episodic memory "settles down" in its representation of a new concept, that representation can then be incorporated slowly into long-term memory. If the representation for a particular concept developed by the episodic memory does not exist or is substantially different from that stored in the long-term memory, the new representation will, through gradual training, be incorporated into the long term memory. If the episodic representation is consistent with that already in long-term memory, the representation will be consequently strengthened in long-term memory through additional training. Finally, feedback from both the episodic and the long-term memory can interact with the attentional mechanism to provide the basis from which to detect novelty and discrepancy worthy of attention.

It is hoped that the system proposed will adequately model the process of lexical acquisition in a more realistic manner, and thus will naturally manifest the fast mapping behavior. The proposed system is intended to address the issues of attention, episodic memory, short-term memory to long-term memory conversion, and the ability to generalize yet still handle novel input gracefully.

6 Conclusions

The prodigious rate at which young children acquire language has led some to dub them "linguistic wizards." The task of acquiring thousands of words, along with semantics and syntax and learning to tie their shoes all within a few short years, requires fast mapping, or the acquisition of a word through extremely limited exposure. This effect has been studied by a number of researchers, and has been found to be quite robust.

The field of connectionist modeling, in its quest for insight into human language acquisition, has thus far failed to develop a feasible system which adequately mimics human performance in language acquisition, including the fast mapping phenomenon so prevalent in children attempting the task. However, as this paper has shown, several current research efforts show promise in addressing these issues. In light of this, a model has been proposed consisting of a combination of various system components described in this paper, which is intended to more closely model episodic memory, attention, and generalization. It is argued that this model would then display the characteristics associated with human performance, including the fast mapping phenomenon.

References

[1] Bates, E., Bretherton, I. & Snyder, L. (1988). Acquisition of a novel concept at 20 months. *From First Words to Grammar: Individual Differences and Dissociable Mechanisms*, 124-134. Cambridge, NY: Cambridge University Press.

[2] Carey, S. (1978) The child as word learner. *In* M. Halle, G. Miller & J. Bresnan (Eds.), *Linguistic Theory and Psychological Reality*, 264-293. Cambridge, MA: MIT Press.

[3] Dollaghan, C. (1985). Child meets word: "Fast Mapping" in preschool children. Journal of Speech and Hearing Research, 28, 449-454.

[4] Hecht-Nielsen, R. (1990). Neurocomputing. Reading, MA: Addison-Wesley.

[5] Hertz, J. A., Krogh, A. S. & Palmer, R. G. (1991). *Introduction to the Theory of Neural Computation*. Reading, MA: Addison-Wesley.

[6] Hinton, G. E. & Plaut, D. C. (1987). Using fast weights to deblur old memories. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, 177-186.

[7] Kohonen, T. (1984). *Self Organization and Associative Memory*, Second Edition. Berlin: Springer-Verlag.

[8] McClelland, J. L. & Rumelhart, D. E. (1986). *Parallel Distributed Processing* (Vol. 2). Cambridge, MA: MIT Press.

[9] Metcalfe, J. (1991). Recognition failure and the composite memory trace in CHARM. *Psychological Review, 98*, 529-553.

[10] Metcalfe Eich, J. (1982). A composite holographic associative recall model. *Psychological Review, 89*, 627-661.

[11] Metcalfe, J. & Murdock, B. B. (1981). An encoding and retrieval model of single-trial free recall. *Journal of Verbal Learning and Verbal Behavior, 20*, 161-189.

[12] Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory*. Cambridge, MA: MIT Press.

[13] Nelson, K. E. & Bonvillian, J. D. (1978). Early semantic development: Conceptual growth and related processes between 2 and 4 1/2 years of age. *In* K. E. Nelson (Ed.), *Children's Language* (Vol. 1), 467-556. New York: Gardner Press.

[14] Nelson, K. E. (1987). Some observations from the perspective of the rare event cognitive comparison theory of language acquisition. *In* K. E. Nelson & A. van Kleek (Eds.), *Children's Language* (Vol. 6), 289-331. Hillsdale, NJ: Lawrence Erlbaum Associates.

[15] Plaut, D. C., McClelland, J. L., Seidenberg, M. S. & Patterson, K. E. (1994). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. Submitted to *Psychological Review*.

[16] Rice, M. L. & Woodsmall, L. (1988). Lessons from television: Children's word learning when viewing. *Child Development, 59*, 420-429.
[17] Rice, M. L., Huston, A. C., Truglio, R. & Wright, J. (1990). Words from "Sesame Street": Learning vocabulary while viewing. *Developmental Psychology, 26*, 421-428.
[18] Rumelhart, D. E. & McClelland, J. L. (1986). *Parallel Distributed Processing* (Vol. 1). Cambridge, MA: MIT Press.
[19] Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning internal representations by error propagation. *In* D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel Distributed Processing* (Vol. 1), 318-362. Cambridge, MA: MIT Press.
[20] St. John, M. F. & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217-257.