A PDP Approach to Processing Center-Embedded Sentences

Jill Weckerly
Jeffrey L. Elman
Center for Research in Language
University of California, San Diego
weckerly @crl.ucsd.edu; elman@crl.ucsd.edu

Abstract

Recent PDP models have been shown to have great promise in contributing to the understanding of the mechanisms which subserve language processing. In this paper we address the specific question of how multiply embedded sentences might be processed. It has been shown experimentally that comprehension of center-embedded structures is poor relative to right-branching structures. It also has been demonstrated that this effect can be attenuated, such that the presence of semantically constrained lexical items in center-embedded sentences improves processing performance. This raises two questions:

(1) What is it about the processing mechanism that makes center-embedded sentences relatively difficult?

(2) How are the effects of semantic bias accounted for?

Following an approach outlined in Elman (1990, 1991), we train a simple recurrent network in a prediction task on various syntactic structures, including center-embedded and right-branching sentences. As the results show, the behavior of the network closely resembles the pattern of experimental data, both in yielding superior performance in right-branching structures (compared with center-embeddings), and in processing center-embeddings better when they involve semantically constrained lexical items. This suggests that the recurrent network may provide insight into the locus of similar effects in humans.

The Problem

It has been known for many years that not all embedded sentences are processed equally easily by listeners. Over a variety of measures, the comprehension and general processing of center-embedded structures has been found to be worse than that of right-branching sentences (Blaubergs & Braine, 1974; Blumenthal, 1966; Blumenthal & Boakes, 1967; Cairns, 1970; Fodor & Garrett, 1967; Larkin & Burns, 1977; Marks, 1968; Miller & Isard, 1964; Schlesinger, 1968). Thus, in

(1)(a) The woman saw the boy that heard the man that left. (RB)

(b) The man the boy the woman saw heard left. (CE)

Sentence (1a), which involves a right-branching structure (RB), is more readily processed than sentence (1b), which involves a center-embedding (CE).

The are various reasons why these two classes of sentences might differ with regard to intelligibility. These include (a) adherence to canonical word order, (b) difficulty of subject-verb matching in the matrix and embedded clauses, (c) distance between subjects and verbs, and (d) consistency of role assignments for nouns in both main and subordinate clauses. While canonical SV-O word order in (1a) is maintained through the matrix clause, in (1b), word order diverges considerably. The processor is faced with three adjacent nouns followed by three adjacent verbs. In (1a), the processor must be able to match the verb encountered in the first relative clause, *heard* with the previous noun, *boy*. This means the processor must "store" some notion of this noun until *heard* is reached. Once past the verb, it goes on to repeat the same action with the next relative clause.

In sentences such as (1b), these resolutions are not made as easily. The processor is required to simultaneously keep track of three nouns before it reaches the first verb, *saw*. It then must determine all the subject-verb-object relationships from representations of items occurring very early in the sentences. As each verb is encountered, the noun that serves as its subject was encountered progressively further back in the sentence, making the distance between the last verb and its subject considerable.

The difficulties of a subject-verb-object match in structures such as (1b) tax the storage capacity of the processing mechanism by requiring the simultaneous activation of a number of items and over a great distance. This is in contrast with sentences such as (1a) where the storage of information must span over one intervening item at most and whose S-V-O relationships can be determined one at a time. A difficulty that both sentences in

(1) share is that nouns serving as subject in one clause instantiate the object role in another. This shift in perspective is more difficult in (1b) because the nouns are adjacent, thereby increasing the chance of confusion.

A second finding of note is that, despite the problems posed by CE sentences, their comprehensibility may be improved in the presence of semantic constraints. Compare the following in (2)

(2)(a) The man the woman the boy saw heard left.

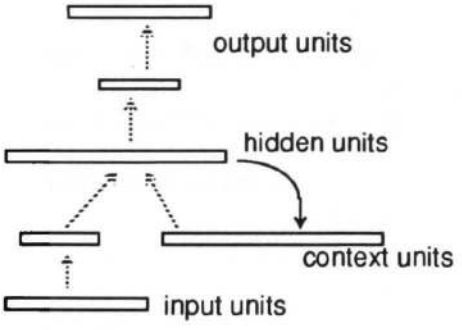
(b) The claim the horse he entered in the race at the last minute was a ringer was absolutely false.

In (2b), the three subjects nouns create strong—and different—semantic expectations about possible verbs and objects. This semantic information might be expected to help the hearer more quickly resolve the various subject and object lineups and as such aids in processing (Bever, 1970; King and Just, 1991; Schlesinger, 1966; Stolz, 1967). The verbs in (2a), on the other hand, provide no such help. All three nouns might plausibly be the subject of all three verbs.

We believe that such phenomena may provide valuable clues as to the nature of the processing mechanisms which subserves language. Our goal in this work has been to test a particular model of language processing (the simple recurrent network) in order to see whether it might provide an explanation for these effects.

The Model

Elman (1990, 1991a) showed that simple recurrent networks (SRNs) were able to develop internally structured representations which provide the basis for abstract, productive, and systematic behavior as required in syntactic processing. Such networks (shown in Figure 1) were able to use cooccurrence statistics to develop representations which captured type/token distinctions, lexical category distinctions, and aspects of grammatical structure.



The SRN model of language processing suggests that the processing differences between center-embedded and right-branching structures arise as a basic consequence of the processing mechanism itself, rather than from limitations in a memory system which is separate from (although used in) language processing. The state machine metaphor embodied in a PDP model delineates a more plausible role for the capacity of memory based on the nature of the representations used in the processing of a sentence. The PDP processor in the time course of a sentence creates a representation which integrates previous context with present input and can be thought of as a state vector that reflects the processor's current position in the sentence. As this vector is continually passed through a "squashing" function, it has only finite precision. Finite precision and degradation over time are also qualities of the human processor.

Using a PDP network, it will be shown how a processor that employs this type of representation degrades, and hence, is finitely precise in such a way that mimics the pattern of processing center-embedded and right-branching sentences by human processors. As the state vector cannot hold information for an infinitely long period of time, it is suggested that the representations used by the human processor are captured best by the state vector metaphor and are similarly limited.

As it has been experimentally demonstrated that the processing of center-embedded sentences is aided by the use of verbs that are semantically constrained, the simulation results of the processing of these structures will show that the architecture and mode of representation in the network support this behavior as well. It can best be understood by considering what types of information are available to the processor and how they are stored. Different types of information interact in a PDP model and influence the output such that it is the product of multiple constraints. The fact that a model such as this uses information other than "purely" syntactic information is nothing new; many theories posit the interaction of this sort as a cornerstone to processing and comprehension. What a PDP model suggests is that with regard to the processing of embedded sentences, the representations used are ones where information present in the various levels of embedding is simultaneously visible and allowed to interact either to facilitate or encumber processing. Unlike a stack-device metaphor for storage, representation in a state vector is not encapsulated and unavailable. The simulations will demonstrate how a PDP model with these properties produces behavior similar to that of the human processor and how viewing processing and representation in this way accounts for behavioral patterns in a straightforward way.

Simulations

For the purposes of the simulation, a small vocabulary was created consisting of 26 words: 10 nouns, 14 verbs, complementizer "that", and an end-of-sentence marker, ".". Since one of the behaviors of interest is processing

sentences with semantic bias, some notion of meaning must be represented. The network can never be semantically grounded in the sense that it knows what words mean; semantic relatedness is captured in the co-occurrence restrictions of the verbs. Classes of nouns serving as subjects and objects fall into classes of humans (NH), animals (NA), documents (DOC), and inanimate objects (INOBJ). The semantic structure of the artificial language is shown below:

VERB	POSSIBLE SUBJECTS	POSSIBLE OBJECTS
walk	NH, NA	—
live	NH, NA	—
write	NH	DOC
send	NH	DOC
love	NH	NH, NA
kick	NH	NH, NA
bite	NA	NH, NA
chase	NA	NH, NA
see	NH, NA	NH, NA, DOC, INOBJ
hear	NH, NA	NH, NA
advise	NH	NH
thank	NH	NH
own	NH	NA
tame	NH	NA

Table 1

The preceding words were constituents of an artificial grammar that generated both simple and complex sentences. Sentence types were produced with the basic pattern of NOUN-VERB-NOUN with verbs equifrequent and every instance of noun able to serve as head of an object- or subject-relative. In this way many different sentence types were generated including center-embedded/object-relative, right-branching, and subject-relative constructions. Sample sentences are shown in (3).

(3)(a) Wizard that advises dorothy tames lion.
(b) Dog that dorothy loves bites witch.
(c) Tiger chases lion that hears dorothy that kicks witch that sees slippers.
(d) Tinman thanks wizard.

Simple sentences were produced by restricting noun phrases to simple nouns and thus followed the strict NOUN-VERB-NOUN pattern. In every case, semantic restrictions were observed. All subject and object relatives were constructed with the appropriate verbs. A subject relative for animal nouns was instantiated only by those verbs for object- relatives which an animal subject is possible. For object relatives, only verbs which specified the head noun as possible object type were used to fill out that relative clause construction. Hence, sentences of the form shown in (4) did not occur.

(4)(a) *Wizard that bites dorothy tames lion.
(b) *Dog that dorothy advises bites witch.

Using the back propagation learning (Rumelhart, Hinton & Williams, 1986), an SRN of the form shown in Figure 1 was trained in a prediction task on data sets varying in composition. The network was presented with sentences, a word at a time. Each word was represented with a 26-bit vector in which a single bit was turned on. As a result, input representation contained no explicit information about the semantic or grammatical characteristics of lexical items. This information had to be learned by the network based on cooccurrence facts.

An incremental training strategy was used, based on the results reported in Elman (1991b), which indicated that the successful induction of hierarchical grammatical structures requires incremental learning. Accordingly, the network was trained on an initial data set of simple (monoclausal) sentences; over time, the percentage of complex sentences was increased until a final ratio of 75% complex/25% simple was achieved. The network was trained on a total of 40,000 sentences, each of which was presented 10 times.

Results

Network performance was evaluated by seeing how closely the network predictions approximated the (empirically derived) likelihood of occurrence of possible next words, given the prior sentence context; optimal performance would be achieved if the network learned the conditional probability distributions. We measured this by computing the mean cosine of the angle between the output activation vectors and the empirical likelihood vectors based on the final training data set. By the end of training, the network was good at predicting the following word in a variety of sentence structures as well as predicting the semantically appropriate verbs and objects for subjects and verbs respectively. The average cosine between the two sets of vectors was 0.8784. Perfect performance would have been 1.0; i.e., the vectors would have been parallel).

Test 1: center-embedded and right-branching sentences

The network was tested on subsets of center-embedded and right-branching sentences. Performance was evaluated on 192 novel sentences, each containing two levels of embedding as shown in (5).

(5)(a) Tinman hears tiger that sees witch that tames lion. (RB)
(b) Witch that tiger that tinman hears sees tames lion. (CE)

Both the likelihood and network output vectors were computed from these 192 test sentences. A four-bit vector that gave the distribution of outputs and likeli-

hoods for each of the categories NOUN, VERB, THAT, and S (end of sentence) was calculated. The mean cosine of center-embedded structures was 0.7137. The mean cosine for right-branching constructions was 0.8484. We can conclude from this that, given the prediction task, the network is more successful at right-branching structures than center-embedded ones.

Discussion

If we equate the network's error as measured by the cosine of the output and likelihood vector with a general processing difficulty then we have results that closely model the human data. It is not the case that center-embedded sentences are impossible; they are simply more difficult relative to other constructions. Another aspect of the model's behavior is reminiscent of the human data: the network's performance decreases drastically at three embeddings which is the limit to comprehension reported in the literature as well.

The network's performance can be understood if we consider the way it represents grammatical structure at the hidden layer. As each new word is presented, the hidden units receive input from both the current word and the previous hidden unit state. Thus, a given word's internal representation always reflects the prior context. Among other things, this context indicates where, in the space of possible grammatical sentence trajectories, the processor is; this context also indicates what may be expected before a sentence-final state is reached. For example, if the current word is *hears*, and the previous hidden unit state reflects the network having seen *dog*, for instance, then the internal state will be such that the network will not predict the end of the sentence. The grammatical structure it has inferred demands that the object of the verb be present before a final state can be achieved.

The representation of relative position in a sentence makes certain demands on the processor regardless of the structure being processed. However, processing structure type also makes its own demands on the representational capacity of the processor. The difference in processing of center-embedded vs. right-branching sentences very much depends on the amount of information that must be stored for further processing in the sentence. As each THAT clause is introduced in center-embedded sentences, the information about the head noun as well as its position relative to the matrix sentence must be represented and stored until the verb of its clause is found. Consider (6):

(6) Dog that dorothy that tiger chases loves bites witch.

After it has seen *tiger*, the network must "remember": (1) that it has seen three nouns, two animals, one human; (2) the fact that the human noun came between the two animals; (3) that the verbs that "go with" these nouns will be of a certain class; and (4) that it must find three verbs in order for the sentence to complete. This places heavy demands on a processor whose actions are executed via a state vector representation.

In right-branching constructions, the representational demands are not as extreme. Consider 7.

(7) Tiger chases dorothy that loves dog that bites witch.

The initial noun that the network encounters is followed immediately by a verb. After seeing this verb, the network can forget about the initial noun because its verb has been found. For the verb, the network need only store information about an appropriate object in generating its predictions. As it encounters the object of the matrix sentence, the processor expects that the sentence be resolved or that the previous noun be the head of another relative clause. In the case of right-branching structures, the processor need only keep information about one noun after encountering the relative pronoun. Thus, there is less information to be stored and over a much shorter distance.

This disparity is clear in the behavior of the network. With right-branching constructions the state vector need only contain representations of two previous words as well as the general position in the sentence. No level of embedding need be stored, because no resolution crucially depends on it. In contrast, with center-embedded sentences, the state vector must reflect sentence position *and* current level of embedding within the sentence. Furthermore, it must also keep information about the previously introduced nouns without having the verbs to advance it into the next state. This "state of suspension" imposes a significant tax on the representational capacity of the hidden unit layer, and approaches its limits of precision.

Researchers have often cited the limitations of working memory to explain certain processing biases of the human parsing mechanism, and specifically, to explain the difficulty in processing center-embedded sentences. In that view, working memory is seen as distinct from the mechanism which contains the grammatical information. The current account provides a somewhat different way of thinking about the asymmetry in processing center-embedded vs. right-branching structures. The account also appeals to the notion of representational storage capacity. However, the representational limitations are seen as intrinsic to the grammatical processor itself, rather than arising from a separate working system. If we view the process of sentence parsing/comprehension as movement from one state to another as in a connectionist network, then memory limitations are not an arbitrary number, but due to the nature of representations in human memory in sentence processing. This capacity specifies that a state like representation can only hold so much information over a certain distance. A reduction in

the amount or information or in the distance to be stored would facilitate processing, as in right-branching structures.

Test 2: Semantically biased and unbiased structures

The other major finding of interest to us was the fact that not all center-embedded sentences are equally difficult to process. We thus proceeded to test the network on different types of center-embedded sentences. Two sets of center-embedded sentences were created: one set of 192 sentences with semantically biased verbs, and another set of 192 sentences with semantically unbiased verbs. Bias in this case means there is some information in the verb that uniquely links it with either its subject or object or both. For instance, in sentence (8a) each verb encountered can only be resolved with one noun as subject whereas in (8b), any subject is compatible with any verb.

(8)(a) Dog that dorothy that bear bites tames chases tiger.
(b) Dog that dorothy that bear sees hears walks.

The network's outputs in response to the two sets of 192 sentences were collected. Likelihood vectors were calculated based on the two sets of center-embedded sentences combined. Comparisons were made between biased and unbiased sentences with one embedding, and then with two levels of embedding. The results were as predicted. For sentences containing one level of embedding, the mean cosine between the activation and likelihood vectors for unbiased sentences was 0.5719; the mean cosine for the sentences with semantic biases was 0.6311. For sentences with two levels of embedding, the overall performance decreased but the same basic pattern remained. In the unbiased condition the mean cosine was 0.5385 and in the biased condition it was 0.5719[1]. It can be concluded that semantic information which uniquely linked a subject with its verb in center-embedded sentences aided the network.

Discussion

We see again that the network's performance parallels that of human listeners. The network benefits from the semantic constraints associated with words in order to represent embedded structures more clearly. The semantic information provided by the verb helps in two ways. First, it helps the network pinpoint the noun which serves as its subject by incompatibility of the other nouns in storage. Second, because this resolution can be made with higher probability, it puts the network in a more precise state of expectation for the next word. As the network goes through the sentence, word by word, there will be various points where it must be able to link words often separated a great distance with their conceptual dependencies.

As soon as it encounters the first verb it must be able to determine which noun is its subject and which noun served as its object. The resources of the network are heavily taxed at this point, because it has information about three nouns that it must keep active at some level. As it encounters the first verb, it is able to make a relatively easy match. As there no words intervening between the last mentioned noun and the first verb, this will be the easiest subject-verb resolution the network has to make. It is at this point where the network might be aided by the nature of the verb. If the verb provides some information by virtue of its co-occurrence restrictions, and to a lesser extent its argument structure, the next subject-verb-object resolution might be greatly aided. That is, if the first verb encounters is compatible with only the last mentioned noun, the resolution can be made quickly and puts the network in a state of awaiting the next word with stronger expectations.

When the next verb is encountered, the network is forced to make another subject-verb resolution. If the nature of this second verb is such that it is compatible only with the intermediary noun and not the first or last mentioned nouns, then the network will benefit greatly from this information. The noun instantiating the subject role of this verb will be determined with less chance of confusion with the other two nouns. The network has heavy representational demands at this point because the potential subjects of the current verb have occurred quite a long time ago and chance of confusion and intermixing of information are high. Thus, information that would clearly delineate a match will be used by the network, and again, will put the network in a more precise state of readiness for the next word.

As it encounters the last verb, the network will be aided by stronger expectations about this verb. Also if this verb is compatible with only the first noun, then the network will be able to match it with its noun, which at this point has occurred many words previous. Semantically biased words aids in putting the processor in a more precise state of readiness. A precise state in this case means that in predicting the word, the activations for incompatible verbs are lower and the activations for appropriate verbs are higher. For example, consider (9)

(9) Dog that tinman that bear chases tames bites witch.

After the network has seen the verb *chases* it must predict a verb that is compatible with a human subject. In

[1] To a large extent, the low values here are only an artifact of the measure used. The likelihood vectors are calculated specific to a data set. The test data set only contains one structure of the many that the network has mastered, and therefore skews the likelihood vectors in a way that makes the network's performance appear low.

general, the network is pretty good at this. It strongly activates human compatible verbs that humans can do compared to the very low activations that are present for bits corresponding to verbs which require animal subjects. Where one can see the effect of type of verb is in this pattern of activation. When the network is presented with a verb that has definite subject specifications relative to the other words, the activations for appropriate verb for the next word are higher and the activations for verbs that are incompatible, and would constitute mistakes, are lower.

If we consider the experimental human data in processing these same type of sentences, we can compare the network's decrease in error and activation of appropriate expectations with a general measure of comprehension. The pattern is basically the same. With the inclusion of semantic cues, error goes down and appropriate activation increases. This would find its correlate in better comprehension in human subjects.

We do not claim that poor comprehension in human subjects is solely due to imprecise predictions, and of course we recognize that the prediction task captures only a small part of natural language processing. Although not the only factor in sentence comprehension, there is evidence that comprehension is in part driven by the ability to anticipate (e.g., Grosjean, 198; Marslen-Wilson & Tyler, 1980). The present findings illustrate general processing characteristics of our PDP model, and we believe similar behaviors would be observed in a comprehension task as well.

The network, its architecture, and its representations suggest similar properties in the human processing mechanism. That the network uses semantic information in what would be considered syntactic parsing, is suggestive of a parallel, interactive system. Additionally, the nature of the interaction of semantic constraints points to a system that allows the simultaneous availability of all types of pertinent information up to that point in the sentence. In other words, information is also available not only across semantic and syntactic modules, insofar as they exist, but also across levels of embedding.

The network represents what it has seen in a sentence by a state vector. Within this vector, the network has "stored" information about properties of nouns and verbs as well as the number of embeddings. Contrary to a stack-like mechanism, information is simultaneously available from all levels. There is no encapsulation of information. Upon encountering a verb, the fact that the network has information about all the previous nouns from different levels of embedding and co-occurrence restrictions of the verb facilitates in the processing of that word and subsequent words. It is suggested that the human processing mechanism has the same properties in order for there to be better comprehension with semantically biased verbs. Any model which designates a traditional stack as its primary storage device is hard pressed to account for the processing difference observed in the experimental data.

References

Bever, T. (1970a). The cognitive basis for linguistic structure. In J.R. Hayes (Ed.), *Cognition and the development of language*. New York: Wiley.

Blaubergs, M.S. & Braine, M.D.S. (1974). Short-term memory limitations on decoding self-embedded sentences. *Journal of Experimental Psychology*, **102**, No.4, 745-748.

Blumenthal, A. (1966). Observations with self-embedded sentences. *Psychonomic Science*, **6**,453-454.

Blumenthal, A.L. & Boakes, R. (1967) Prompted recall of sentences. *Journal of Verbal Learning and Verbal Behavior*, **6**, 674-676.

Elman, J.L. (1990). Finding structure in time. *Cognitive Science*, **14**, 179-211.

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

Elman, J.L. (1991b). Incremental learning, or the importance of starting small. Technical Report 9101, Center for Research in Language, University of California, San Diego.

Foss, D.J. & Cairns, H.S. (1970). Some effects of memory limitation upon sentence comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, **9**, 541-547.

Grosjean, F. (1980). Spoken word recognition processes and the gating paradigm. *Perception & Psychophysics*, **28**, 267-283.

King, J & Just, M.A. (1991). Individual differences in syntactic processing: the role of working memory. *Journal of Memory and Language*, 30, 580-602.

Larkin, W. & Burns, D. (1977). Sentence comprehension and memory for embedded structure. *Memory and Cognition*, **5**, 17-22.

Marks, L.E. (1968). Scaling of grammaticalness of self-embedded English sentences. *Journal of Verbal Learning and Verbal Behavior*, **7**, 965-967.

Marslen-Wilson, W., & Tyler, L.K. (1980). The temporal structure of spoken language understanding. *Cognition*, **8**, 1-71.

Miller, G. & Isard, S. (1964). Free recall of self-embedded English sentences. *Information and Control*, **7**, 292-303.

Schlesinger, I.M. (1968). *Sentence structure and the reading process*. The Hague: Mouton.

Stolz, W.S. (1967). A study of the ability to decode grammatically novel sentences. *Journal of Verbal Learning and Verbal Behavior*, **6**, 867-873.