A Connectionist Model Of Speech Act Prediction

Shane S. Swamer, Arthur C. Graesser, Stanley P. Franklin, Marie A. Sell, Robert Cohen, & William B. Baggett
Department of Psychology
Institute for Intelligent Systems
Memphis State University
Memphis, TN 38152
swamerss@msuvx1.memst.edu

Abstract

We developed a connectionist architecture that accounts for the systematicity in the sequential ordering of speech act categories. That is, to what extent can the category of speech act n+1 be successfully predicted given speech acts 1 through n? Three connectionist architectures were contrasted: Elman's recurrent network, a single-entry backpropagation network, and a double-entry backpropagation network. The recurrent network fit the speech act sequences in naturalistic conversation better than the backpropagation networks. Most of the systematicity was captured by the network's use of 2 to 3 prior speech acts of context.

Introduction

How predictable is human conversation? Imagine a typical exchange with an acquaintance you have not seen in awhile.

Person 1: "Hello, how have you been?"
Person 2: "I'm fine. How have you been doing?"
Person 1: "Pretty good."
Person 2: "That's good."

Predictable exchanges like this demonstrate the systematicity inherent in conversation. This systematicity can be studied at either a syntactic level, i.e., the order of speech act categories, or at a deeper semantic level that takes into account the world knowledge, beliefs, and goals of the speakers. We concentrated exclusively on the systematicity inherent in the ordering of the speech act categories. That is, to what extent is the next speech act category in a conversation predictable?

This research was funded by Office of Naval Research grants N00014-90-J-1492 and N00014-92-J-1826.

There have been extensive debates on what speech act categories are necessary to comprehensively cover the range of human speech acts (D'Andrade & Wish, 1985). A useful set of categories would not only be theoretically grounded, but also empirically adequate in the sense that judges can agree on the assignment of categories. D'Andrade and Wish's (1985) system satisfies these two important constraints. Our speech act categories included many of D'Andrade and Wish's categories and a few of our own. Table 1 lists and defines each of the speech act categories. Table 1 also presents the *a posteriori* likelihood that each speech act category occurred in the naturalistic conversations in our corpus.

The primary goal of this research is to develop and test models that capture the predictability in sequences of speech act categories. Each model predicts the category of the next speech act (n+1), given the sequence of previous speech act categories (1 to n). We investigated different connectionist architectures in capturing this systematicity. We adopted a connectionist model because these models are able to induce the structure and systematicity that exists in the data.

There has been some previous research on sequential predictions. For example, Schegloff and Sacks (1973) analyzed adjacency pairs in which all the systematicity can be explained by the previous speech act. One common adjacency pair would be the [Question then Reply-to-question] sequence. Other researchers have specified longer sequences. Mehan (1979) identified a three-part sequence: [Question then Reply-to-question then Evaluation]. Clark and Schaefer (1989) identified even longer sequences: [Question, then Counter-question, then Reply-to-counter-question, then Reply-to-original-question]. Although these sequences and structures have been identified, researchers have not assessed the extent to which this predictability is captured in normal conversation.

Table 1: Speech Act Categories

Speech Act	Definition	Example	A Posteriori Distribution
Question (Q)	Interrogative or information seeking expression that is not an indirect request.	How does this piece fit here?	.207
Reply to Question (RQ)	Response that specifically answers a previous question.	Yes, I'm fine.	.141
Directive (D)	Request signaled by imperative form.	Give me the blue piece.	.040
Indirect Directive (ID)	Request in non-imperative form.	Can you open the door?	.015
Assertion (A)	Report about some state of affairs that could be true or false.	The risograph machine is broken.	.403
Evaluation (E)	An expression of sentiment.	That stinks!	.031
Verbal Response (R)	Spoken acknowledgment of the previous speech act.	Uh-huh.	.069
Nonverbal Response (N)	Unspoken acknowledgment of the previous speech act.	(Head nod)	.026
Juncture (J)	Lengthy pause in conversation.	(Pause)	.069

Connectionist Models Of Predicting Speech Act Categories

We investigated three connectionist architectures: (1) Elman's recurrent network (Elman, 1990), (2) a single-entry backpropagation network that uses one prior speech act in making its prediction, and (3) a double-entry backpropagation network that uses two prior speech acts. The details of each model are covered in this section.

Elman's Recurrent Network

The first speech act network consists of the recurrent architecture developed by Elman (1990). Elman's recurrent network keeps an encoding of all previous input and is able to use this information to induce the structure underlying temporal sequences. The use of prior context makes the Elman network quite suitable for the task of discovering the systematicity in a temporal ordering of speech acts.

The Elman model consists of four layers of nodes, as shown in Figure 1. The four layers include the input layer, the hidden layer, the context layer, and the output layer. The input layer uses local representation, i.e., there is a node for each possible speech act category. There are two participants in the conversation, each with eight possible speech act categories, so there are 16 total speech act categories. These 16 possible speech acts, plus the juncture, yield 17 speech act categories that the input layer can encode. Since local representation is used, a single input (speech act n) is sent into the network simply by activating the corresponding input node. For example, suppose person 1 asked a question. This would be represented by activating the Q1 node on the input layer of the network.

The output layer contains the network's predictions for speech act n+1. Output items have activation in direct proportion to the degree to which the network predicts them.

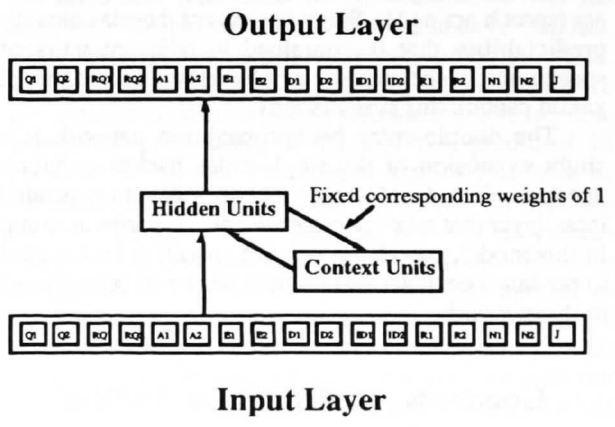

Figure 1: Elman's Recurrent Network

For example, if Q1 were used as input, the RQ2 node would probably be highly activated because it is person 2's answer to the question.

The Elman network differs from the normal backpropagation networks because it has a context layer that keeps track of previous inputs. This context layer allows the Elman network to induce temporal sequences (Elman, 1990; Cleeremans, Servan-Schreiber, & McClelland, 1989). The context layer stores the activations of the hidden layer from the previous time step. Since the activations of the hidden layer depend on the previous input (speech act n) along with the activation of the context layer, the hidden layer is receiving information about the present input along with information about past inputs. This combination of the input with the context layer is subsequently copied to the context layer, allowing the network to store its encoding of history in the context layer.

There was a total of 440 connections that were allowed to vary in the weight space. There were 170 connections between the input layer and the hidden layer, given that there were 17 input nodes and 10 hidden unit nodes. Similarly there were 170 connections from the hidden layer to the output layer. The final 100 connections linked the 10-node context layer to the 10-node hidden layer. It should be noted that there are 10 connections from the hidden layer to the context layer that are fixed at 1.0.

Single and Double Backpropagation Networks

The single-entry backpropagation network is a simplification of the Elman network. It has only three clusters of nodes: an input layer, a hidden layer, and an output layer. There is no context layer in the single-entry backpropagation network to encode previous information. There were 340 connections in this network. This model uses only the previous speech act (speech act n) to predict the category of the next speech act (speech act n+1). Some speech act theories identify predictability that is contained in adjacent pairs of speech acts. The single-entry backpropagation model would capture this systematicity.

The double-entry backpropagation network is a slight expansion of the single-entry backpropagation network. The double-entry network has an expanded input layer that takes two speech act categories as input. In this model, speech act n-1 and speech act n are used to predict speech act n+1. There were 510 connections in this network.

Goodness of Prediction Indices

Goodness of prediction (GOP) indices measure the extent to which the model is correctly predicting the category of the next speech act (n+1). There are two different GOP indices: maximal activation and above-threshold.

The *maximal activation* GOP index considers only the output node with the highest activation to be the network's prediction for speech act n+1. The maximal activation GOP index measures the probability that the predicted speech act actually occurs in the data, over and above the base rate of the speech act category. The GOP index is computed as shown in formula 1.

Maximal Activation GOP index =

(Hit Rate - Base Rate) / (1 - Base rate) (1)

In the above calculation, hit rate was the probability (over many observations) that the theoretically predicted category matches the actual category. The base rate likelihood was the probability that the speech act would be predicted on the basis of its *a posteriori* distribution (see Table 1).

The *above-threshold* GOP index assesses the network's performance when allowed multiple predictions. All output nodes that exceed a threshold activation level are considered theoretical predictions for n+1. The GOP index is computed in the same manner as formula 1. However, in this case, hit rate is probability that the actual output is among the set of theoretically predicted categories, and the base rate is the sum of the *a posteriori* probabilities of all the predicted speech acts.

Simulation Details

Data Set

The data used in the simulation was taken from videotaped conversations of second and sixth graders (Sell, et al., 1991). Dyads of second and sixth graders were observed in three different tasks: a question answering task (akin to 20 questions), freeplay, and a puzzle task. The children were further segregated according to how well they knew each other. Relations included mutual friends, acquaintances, and unilateral friends (where child A considers B a friend, but not vice versa). There were 18 different subject groups and 5 subjects per group: 3 (type of relation) x 2 (age) x 3 (task). Therefore, there were 90 10-minute dyadic conversations altogether. These conversations were segmented into sequences of the 17 speech act categories (see Table 1). The juncture category was a special category used to reset context when there was a new conversation, when there was a long pause in a conversation (5 or more seconds), or when the speech was incomprehensible. The agreement between a pair of judges in assigning speech acts to categories was moderate (overall kappas of .82, .76, and .74 for the question task, puzzle task, and free time task, respectively).

Training the Networks

The data taken from the children's conversations formed temporal sequences of speech acts. The 90 individual conversations provided an overall sequence of 16657 speech acts. A 17-bit binary string was used to locally represent the particular speech act. Once the data was transformed into the appropriate local encoding, each network was trained on 12 passes through the data set. Each connectionist architecture then underwent eight different training runs, with different random sets of initial weights.

Results and Discussion

Which Model Best Explains the Data?

Table 2 segregates the maximal activation and the above-threshold analyses. Within each of these, there is the GOP index score, the hit rate, and the base rate.

When considering the maximal activation GOP index there were significant differences among the three networks, $F(2,23)=257.94$, $p<.05$. Follow up tests revealed that both the recurrent and the double-entry backpropagation networks perform significantly better than the single-entry backpropagation networks. According to the data, a single speech act of context is not enough to capture all of the systematicity in the data.

When analyzing the above-threshold GOP index, where multiple predictions are allowed, again there were significant differences among the networks $F(2,23)=227.56$, $p<.05$. The recurrent network performed significantly better than both backpropagation networks. The two GOP indices together indicate that the recurrent network is best able to accommodate the systematicity inherent in the sequence of speech act categories.

TABLE 2
Goodness of Prediction (GOP) Scores For Three Network Architectures

	Recurrent Network	Backpropagation Single-Entry	Backpropagation Double-Entry
Maximal Activation Analysis			
GOP Index	.289	.268	.292
Hit Rate	.375	.357	.379
Base Rate	.122	.122	.122
Above Threshold Analysis			
GOP Index	.376	.322	.367
Hit Rate	.565	.523	.554
Base Rate	.304	.295	.296

How Much Context is Needed?

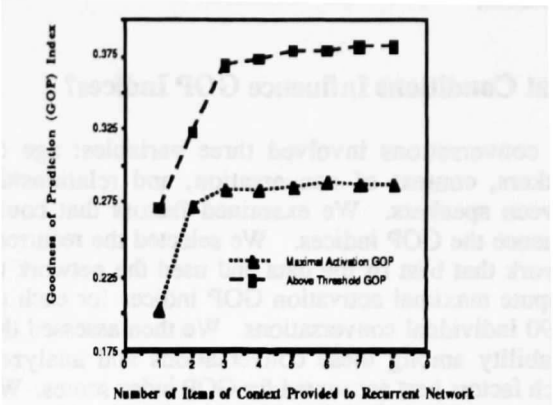

Figure 2: Goodness of Prediction Index Based on Number of items of Context

We examined how much of the context is made use of in the recurrent network. We varied the number of items of context that was fed into the recurrent network. Figure 2 plots the GOP indices as a function of the number of context items considered. The results indicate that only 2 to 3 items of context are made use of by the recurrent network. Given that the recurrent network is expected to unpack most or all regularities in the data, we conclude that only three speech acts of context capture any systematicity; any information further back has a very small impact on the prediction of speech act categories.

What is an Ideal Threshold?

An important parameter in the simulation was the threshold we selected for the above-threshold analysis. The findings rely on the value picked for the threshold. We wanted a threshold value that would maximize goodness-of-prediction, while also making a reasonable number of predictions. A network that predicted 16 possible categories for speech act n+1 would be undiscriminating and therefore not useful. In deciding on the threshold, we tested the network at various thresholds and recorded the GOP indices. We then divided these GOP indices by the average number of output categories that were above threshold (assigning a minimum value of 1 for number of categories). This adjusted above-threshold GOP index matched both our criteria. We wanted to maximize the GOP indices while also keeping the number of predictions at a low number. At thresholds of .05, .10, .15, .20, and .25 we found adjusted GOP indices of .122, .184, .214, .221, and .177. We ended up choosing a threshold of .18

which had a maximum adjusted GOP index of .221. This threshold also limited the network to predicting an average of 1 or 2 speech act categories, which seemed reasonable.

What Conditions Influence GOP Indices?

Our conversations involved three variables: age of speakers, context of conversation, and relationship between speakers. We examined factors that could influence the GOP indices. We selected the recurrent network that best fit the data and used the network to compute maximal activation GOP indices for each of the 90 individual conversations. We then assessed the variability among these conversations and analyzed which factors best accounted for GOP index scores. We found that context was the only significant predictor of GOP indices. The GOP indices were .38, .07, and .18 for the 20 questions task, puzzle task, and freeplay, respectively, $F(2,72)= 50.09$, $p<.05$. The F-scores for Age and Relationship were both non-significant ($F < 1$). We also did not find any significant interactions.

Technical Assumptions

We performed a number of auxiliary analyses that assessed technical assumptions behind our architectures. To assess our choice for the number of hidden units and context units, we varied the number of nodes in the hidden layer. The above-threshold GOP indices were .363, .376, .376, .384, .383 when there were 6, 8, 10, 12, and 14 hidden nodes, respectively. Clearly, the number of hidden units did not affect the GOP index very much.

In an attempt to find the amount of training needed, we trained the network at various number of epochs, up to 35 passes through the corpus of 16657 speech act categories. The network's performance appeared to asymptote within 10 passes through the data.

We used the split-half transfer technique to test generalization. We randomly assigned half of each conversation to a training set and the other half to a transfer set. The network was then trained on the items in the training set. After training, GOP indices were recorded for network performance on both the training and transfer sets. The GOP index scores for the training set were exactly the same as the GOP index scores for the transfer set. The network performed as well on new data (not previously trained) as on the training items. Therefore, the network clearly achieved generalization.

Some Network Regularities

The network's activations and patterns of prediction were examined for any consistent regularities. Two of the most pervasive adjacency pairs discovered were the [Question then Response-to-question] pair and the [Assertion then Assertion] pair. The Response-to-question output node only received activation when preceded by a question from the opposite person. When this occurred, the output node had the highest average activation awarded by the network. The Response-to-question was correctly predicted 89% of the time. The Assertion also received high activation when preceded by an Assertion by either party. Assertions were correctly predicted 59% of the time.

General Discussion

There were a number of findings in this study. First, Elman's recurrent network performed better than both backpropagation networks in accounting for the data. The backpropagation networks in this paper only used a context of two items at most. The context layer of the recurrent network appears to enable the network to utilize context further back than two items. However, the context analysis (Figure 2) suggests that not much information is gained after three items of context. It appears that only 2 or 3 items of context are utilized in predicting speech act categories. These findings fit with current theories that have specified regularities for 3 items of context or less. For example, Schegloff and Sacks' (1973) adjacency pair analysis contains one item of context. Mehan's (1979) Question-Answer-Evaluation triplet contains two items of context, while Clark and Schaefer's Question Counter-question example contains 3 items of context.

We are currently in the process of comparing the recurrent connectionist network's performance to other computational models: an Augmented Transition Network (ATN) designed similarly to Stevens and Rumelhart's (1975) ATN for English sentences, and a production system model (Anderson, 1983; Kieras & Bovair, 1984).

References

Anderson, J.R., (1983). The architecture of cognition. Cambridge, MA: Harvard University Press.

Clark, H. H., & Schaefer, E. F. (1989). Contributing to discourse. Cognitive Science, 13, 259 - 294.

Cleeremans, A., Servan-Schreiber, D. & McClelland, J.L. (1989). Finite state automata and simple recurrent networks. Neural Computation, 1, 372-381.

D'Andrade R.G., & Wish (1985). Speech act theory in quantitative research on interpersonal behavior. Discourse Processes, 8, 229-259.

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

Goffman, E. (1974). Frame analysis. Cambridge, MA: Harvard University Press.

Kieras, D.E., & Bovair, S. (1984). The role of a mental model in learning to operate a device. Cognitive Science, 8(3), 255-273.

Mehan, H. (1979). "What time is it Denise?": Asking known-information questions in classroom discourse. Theory into practice, 18, 285-294.

Schegloff, E. A., & Sacks, H. (1973). Opening up closings. Semiotica, 8, 289-327.

Sell, M.A., Cohen, R., Crain, M., Duncan, M.K., MacDonald, C.D., & Ray, G.E. (1991, April). The context of dyadic interactions: Communication as a function of task demands and social relationships. Society for Research on Child Development, Seattle, WA.

Stevens, A.L., & Rumelhart , D.E. (1975). Errors in reading: Analysis using an augmented transition network model of grammar. In D. Norman and D. Rumelhart (Eds.), Explorations in cognition. San Francisco: Freeman.