Computational Model of Cognitive Development: Filtering Ambient Speech to Facilitate Word Learning

Gregory A. Caza
Department of Computer Science
University of Otago
PO Box 56
Dunedin 9015, New Zealand
gcaza@cs.otago.ac.nz

ABSTRACT

Infants manage to learn word meanings in very noisy environments. Despite an onslaught of many potentially confusing examples, language acquisition actually becomes more rapid around 18 to 24 months of age. During the same stage of life, a number of cognitive milestones are also reached. A review of developmental psychology and linguistics produces a theory that these parallel progressions may be linked by more than mere coincidence. As part of my master's degree, I have developed a computational model to investigate language learning. It simulates a child who identifies communicative acts and then follows cues from a caregiver to disambiguate a single-word learning situation. Early results from the model are statistically similar to those observed in previous psychological experiments with actual children.

Categories and Subject Descriptors

I.2.0 [**Artificial Intelligence**]: General—*cognitive simulation*; I.2.6 [**Artificial Intelligence**]: Learning—*language acquisition*; J.4 [**Computer Applications**]: Social and Behavioural Sciences—*psychology*

General Terms

Vocabulary Spurt, Joint Attention, multilayer perceptron

Keywords

neural network, language acquisition, infant, cognitive development

1. BACKGROUND

Infants are exposed to an enormous quantity of potentially overwhelming verbal information while acquiring language. This auditory input can bombard them from many sources, including multiple speakers, televisions, and music. In parallel to this verbal stream, there is a separate stream of sensorimotor information from the infant's own actions and the observation of others' actions. The correlation between the two modalities is quite low [10]; a word heard at a given moment will only occasionally correspond to the current sensorimotor representation. A simple illustration is shown as a timeline in Table 1, where the two streams are congruent in only one case (i.e., seeing a cat and hearing the word "cat".)

Table 1: The top row represents visual, sensorimotor information. The bottom row shows examples of ambient speech coinciding in time. The single congruent case is highlighted.

DOG	DOG	MOTHER	**CAT**	FOX	DOG
"run"	"sleep"	"baby"	**"cat"**	"cat"	"jump"

Yet, even under these circumstances, infants display a remarkable ability to acquire language and correctly assign labels. In fact, by the age of 18 to 24 months, words can be learned in a single exposure [9]. Many researchers believe this skill is one of the markers of a **vocabulary spurt** [8]. According to that hypothesis, the rate of word learning shows a sharp increase around 18 months. However, other researchers dispute the existence of a vocabulary spurt–or at least the prevalence of one. They argue for a steady, more gradual increase in word learning, with possibly only 18% of children [5] demonstrating a measurable spurt.

Whatever the case, most researchers would agree that word learning starts slow at around 12 months of age [11] and picks up speed throughout the second year of life. The same developmental span also marks a number of cognitive milestones. Is this a coincidence, or do certain cognitive skills assist the infant in filtering the noise from ambient speech?

Ideally, the two perceptual streams mentioned earlier will be filtered to learn the correct mapping between a sensorimotor representation and a word. Two ideas from developmental linguistics help theorise how an infant might recognise an appropriate correlation and perform the necessary filtering: communicative understanding and intention-reading.

First, the social-pragmatic theory of language acquisition suggests that understanding a communicative situation forms the foundation of word learning [1]. According to this theory, word learning requires an understanding of the intentions behind utterances. 9- to 12-month-olds can share, follow and direct attention but it is not until around the first birthday that these crucial intention-reading skills begin to develop.

NZCSRSC 2008, April 2008, Christchurch, New Zealand.

A second piece of evidence comes from numerous studies (e.g., [2]) showing that older infants, armed with more developed intention-reading skills, will follow **joint attention** (JA). Joint attention between caregiver and child is defined [4] by: both attending to the same object; and, both being aware of the fact that they are attending to the same object. Thus, two people independently looking in the same direction is a necessary–but not sufficient–condition. There is an implicit mutual awareness that adds an interactive dimension. Baldwin [2] demonstrated that 19- to 20-month-olds will give referential cues, such as pointing and gaze, priority when learning object labels. There is also evidence [3] that this mapping will *only* be performed when there is clear referential intent. Thus, word learning is facilitated when an infant recognises the initiation of a communicative act and then correlates his or her own visual attention with what the adult is speaking about.

I developed a computational model to investigate language learning in these interactive, social-communicative situations. The model simulates a child who can identify communicative acts and establish joint attention. The network successfully processes parallel streams of verbal and sensorimotor input that are only sporadically correlated.

2. METHODS

2.1 Simple Word Learning Model

van Oijen [12] created an artificial neural network model of single-word learning. Although his ultimate goals were different than the current research, the basic foundation upon which he built his other networks is useful to the current topic. His simplest feed-forward network is illustrated in Figure 1. The architecture used, known as a **multilayer perceptron**, is very common in the field of neural networks. An input vector is applied to the nodes in the first layer (the input layer.) Each node in the second layer (the hidden layer) receives a weighted sum of those inputs and translates it into one value, based on some activation function. The values from that layer are projected to the final layer (the output layer), where a weighted sum is again passed through an activation function. The outputs of the final nodes form an output vector that is used as the overall response of the network. The connections between the layers start with random weights and are strengthened or weakened during training. Weights are adjusted according to the standard backpropagation algorithm [6]. Backpropagation is a form of supervised learning, wherein the output at each step is compared to the expected result (i.e., the correct answer) and any difference becomes an error signal which is used to assign credit or blame.

van Oijen's network took a binary-encoded sensorimotor (SM) item (e.g., 'DOG') as its input. For example, if the encoding for 'DOG' was 101, the inputs to the first and third nodes would be turned on. Activation would propagate forward and the on-off status of the output nodes would be translated into a binary representation of the word label. Training continued until an acceptable minimum error was reached for all tested inputs. After training, the connections feeding forward between the layers would produce an output which encoded the correct word label (e.g., "dog") for each given input.

The model worked very well for learning isolated words. It also satisfied a difficult constraint of language-learning systems: linguistic bootstrapping. That is, how are the very first words learned when the learner has no background upon which to depend? van Oijen's network actually started word learning with no prior knowledge. However, initial word learning is not the focus of my research, and van Oijen's network cannot accomplish all of my goals. First of all, the network was trained on artificially-sanitised input sets that were not cluttered with the noise of the ambient speech discussed earlier. Furthermore, his network did not model rapid word learning of the sort that the literature anticipates in 18- to 24-month-olds. Thus, how to improve upon van Oijen's work became the open question.

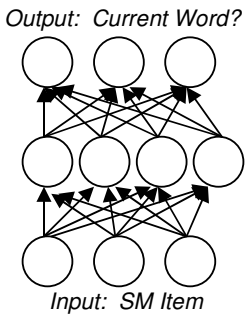

Figure 1: Representation of the simple word learning model, implemented as a backpropagation neural network with one hidden layer.

2.2 Modelling Joint Attention

Consistent with the developmental psychology literature, modelling joint attention was proposed as a method for filtering noise from the ambient speech. With less noise, an adapted version of the single-word learning model could theoretically achieve the task at hand.

Of course, basic joint attention is not the only requirement. Even if a child consistently follows the referential intent (e.g., gaze) of the mother[1], there will not always be a word-learning situation because she might not actually be speaking at the time. What is crucial is that she also produces a word; only then can her gaze be followed to find the appropriate SM input. This idea is illustrated in the timelines of Table 2. The highlighted steps identify the target case. In general, it is assumed that the child is exploring his environment by looking at salient stimuli. When the mother begins talking (as in the first highlighted step), a communicative situation is recognised. At the next time step (the second one highlighted), joint attention is initiated and the child's attention follows the mother's gaze to see what she is looking at (in the first two rows.) The word heard (in the last row) is associated with the visual information observed at the same time step.

Thus, the noisy simulation of ambient speech can be filtered

[1]For the remainder of this paper, I will assume that the caregiver is female and that the infant is male. No gender bias is presumed or intended.

Table 2: Illustration of the joint attention model. The top two rows display visual information. The bottom row shows examples of coinciding ambient speech. The crucial trigger, 'MOTHER-TALK', and the subsequent learning opportunity are highlighted.

DOG	DOG	**MOTHER**	**CAT**	FOX	DOG
	JUMP	**TALK**			RUN
"run"	"sleep"		**"cat"**	"cat"	"jump"

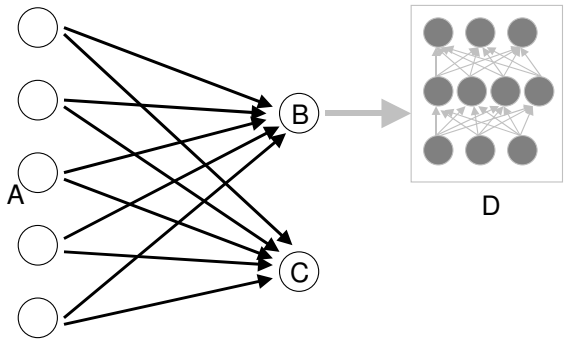

Figure 2: Example of a more advanced word learning model. The original neural network is gated by a subnetwork that implements the joint attention filter.

by gating the input to single-word learning and only turning it on at appropriate times. According to this strategy, language learning is only enabled when the crucial trigger is recognised. When the infant recognises a "MOTHER-TALK" situation, he follows her gaze to see what she is talking about. At this point in the model, language learning is enabled and the network is trained to map the spoken label to the attended-to SM information. (E.g., (a) Mother says "cat" and indicates a cat; (b) baby looks at the cat and hears "cat".)

A network with the joint attention filter included is represented in Figure 2. SM pairs are presented at the inputs (area *A* of the figure.) Each input is fully-connected to the two nodes in the next layer, unidirectionally. One node in this middle layer (labelled with a *C* in the figure) represents a simple state of observation. The other node in the middle layer (node *B* in the figure) is used to gate the word learning and should be turned on in the language acquisition context. Activating this node allows the SM information to pass through to the word learning subnetwork (shown in the figure as shaded area *D*.) During training, the SM pair is coupled with the auditory cue to learn the correct word mapping.

The network was presented with a training set of 20 nouns and 9 verbs. The auditory and visual streams were simulated by a series of SM-word pairs (e.g, "ELEPHANT-eat; DOG-eat; CAT-cat; ELEPHANT-elephant; etc.") The pairs were randomly generated and a correct correlation (e.g., "DOG-dog") between the streams was enforced at a chance level of 15%. The goal was to learn the correct mapping between the noun SM items and their corresponding words. The set was first used in "No JA" training runs without a filter, effectively reducing the network structure to that of van Oijen's. For the second type of "JA" training runs, the cognitive development of joint attention was simulated by enabling the filter after 5 nouns had been successfully learned. Once enabled, the gate would be fully functional for the remainder of the training period.

3. RESULTS

The best performance was observed with 50 nodes in the hidden layer, a learning constant of 0.4, and a target error of 0.01. Typical results for the word learning network are summarised in Figure 3. Success is measured by the number of words correctly learned after a given number of training epochs.

Without joint attention filtering (e.g., **No-JA** in the figure), the network will learn about nine words and then effectively flatline; it progresses very slowly and cannot solve the problem space. If the filter is added after five words are learned, the network is able to learn the full set of 20 nouns before the end of 15 000 training runs. Such successful learning may display either a spurt-like progression (e.g., **JA-1**) or be more gradual (e.g., **JA-2**).

A sharp increase (e.g., **JA-1**) occurred in approximately 20% of the test runs, which is very close to the 18% found experimentally [5]. Also, averaging over a number of examples produces a much more gradual growth (**JA Mean**), with no discernible spurt or steep acceleration in slope.

4. DISCUSSION AND FUTURE WORK

It is too early in my research to make definitive conclusions. However, it is very encouraging that the model displays behaviour that is superficially similar to what has been found in experiments with actual children. As of yet, I have performed no formal analysis of *why* the network learns better for certain test runs. The logical next step is to analyse the root cause of the variability.

There are also improvements that can be made to increase the neurobiological plausibility of the model. First, it must work with a larger vocabulary. Experiments should also be performed with different critical periods, in order to simulate individual differences. Finally, the dynamic nature of the developing infant can be approximated by dynamically growing the network as the vocabulary increases.

The perfectly-attuned joint attention filter used is highly simplified, suggesting an idealised representation of a word learner. The ability to recognise and take advantage of a communicative situation is not innate. A more interesting question is how the child learns *when* to learn. How does the baby identify "MOTHER-TALK" as a tool and use it to facilitate future word learning? Recent findings about the role of dopamine in word learning [7] suggest a reward-dependent learning scheme, which will be investigated in a future step of my research.

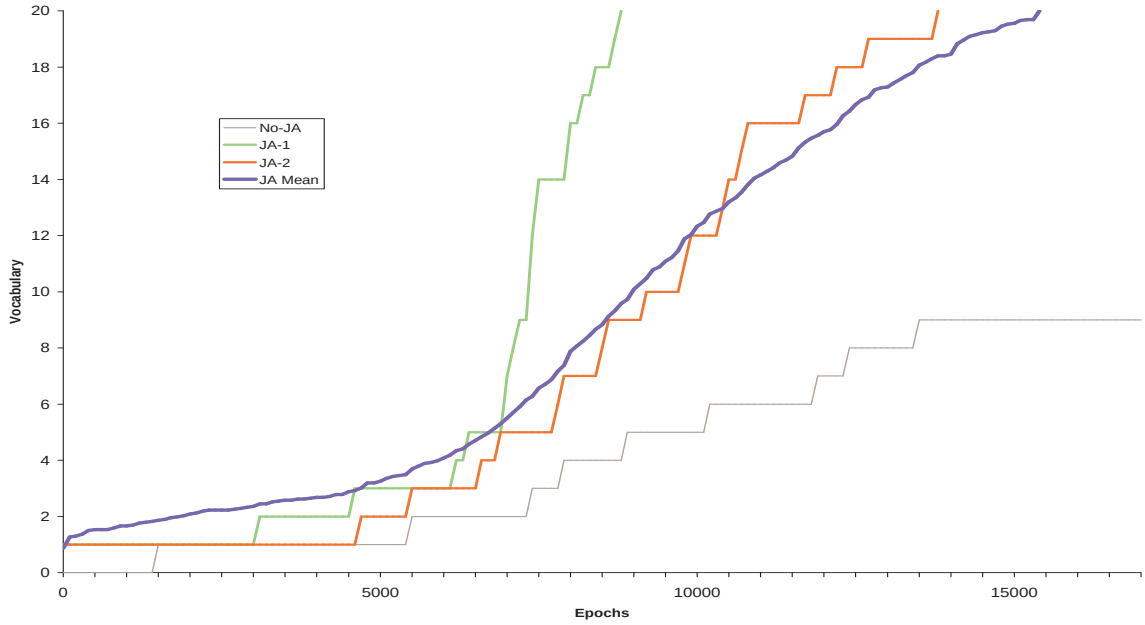

Figure 3: Results from different training regimes, represented as number of words learned over successive epochs. *No JA*: an example without joint attention filtering; *JA-1*: an example of spurt-like behaviour; *JA-2*: an example of gradual acceleration; *JA Mean*: mean results from 30 different filtered networks.

5. REFERENCES

[1] N. Akhtar and M. Tomasello. The Social Nature of Words and Word Learning. In *Becoming a Word Learner: A Debate on Lexical Acquisition*, pages 115–135. Oxford University Press, Oxford, England, 2000.

[2] D. A. Baldwin. Early Referential Understanding: Infants' Ability to Recognize Referential Acts for What They Are. *Developmental Psychology*, 29(5):832–843, 1993.

[3] D. A. Baldwin, E. M. Markman, B. Bill, R. N. Desjardins, J. M. Irwin, and G. Tidball. Infants' Reliance on a Social Criterion for Establishing Word-Object Relations. *Child Development*, 67:3135–3153, 1996.

[4] N. Eilan. Joint Attention, Communication, and Mind. In N. Eilan, C. Hoerl, T. McCormack, and J. Roessler, editors, *Joint Attention: Communication and Other Minds*, pages 1–33. Oxford University Press, Oxford, England, 2005.

[5] J. Ganger and M. R. Brent. Reexamining the Vocabulary Spurt. *Developmental Psychology*, 40(4):621–632, 2004.

[6] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Macmillan, New York, NY, USA, 1999.

[7] S. Knecht, C. Breitenstein, S. Bushuven, S. Wailke, S. Kamping, A. Floel, P. Zwitserlood, and E. B. Ringelstein. Levodopa: Faster and Better Word Learning in Normal Humans. *Annals of Neurology*, 56(1):20–26, 2004.

[8] W. O'Grady. *How Children Learn Language*. Cambridge University Press, Cambridge, UK, 2005.

[9] T. Regier. The Emergence of Words: Attentional Learning in Form and Meaning. *Cognitive Science*, 29:819–865, 2005.

[10] M. A. Sabbagh and D. A. Baldwin. Understanding the Role of Communicative Intentions in Word Learning. In N. Eilan, C. Hoerl, T. McCormack, and J. Roessler, editors, *Joint Attention: Communication and Other Minds*, pages 165–184. Oxford University Press, Oxford, England, 2005.

[11] M. Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, Cambridge, MA, USA, 2003.

[12] J. van Oijen. A model of the relationship between language and sensorimotor cognition. http://www.cs.otago.ac.nz/research/techreports.html, 2006.