In Proc. of the 6th Joint Conf. on Information Scineces, Research Triangle Park, NC, pp. 553-556, 2002

Modeling the Development of Lexicon with a Growing Self-Organizing Map

Igor Farkaš [1] and **Ping Li**
ifarkas,pli@richmond.edu
Department of Psychology
University of Richmond, VA 23173, USA

Abstract

We present a self-organizing neural network model that can acquire an incremental lexicon. The model allows the acquisition of new words without disrupting learned structure. The model consists of three major components. First, the word co-occurrence detector computes word transition probabilities and represents word meanings in terms of context vectors. Second, word representations are projected to a lower, constant dimension. Third, the growing lexical map (GLM) self-organizes on the dimension-reduced word representations. The model is initialized with a subset of units in GLM and a subset of the lexicon, which enables it to capture the regularities of the input space and decrease chances of catastrophic interference. During growth, new nodes are inserted in order to reduce the map quantization error, and the insertion occurs only to yet unoccupied grid positions, thus preserving the 2D map topology. We have tested GLM on a portion of parental speech extracted from the CHILDES database, with an initial 200 words scattered among 800 nodes. The model demonstrates the ability to highly preserve learned lexical structure when 100 new words are gradually added. Implications of the model are discussed with respect to language acquisition by children.

Introduction

Contexts in which a word occurs provide a considerable amount of information for the representation of word meaning. This position has been the core for various computational models of language (such as HAL [1] or LSA [2]), in which every word is represented by its context vector in a high-dimensional space. In the connectionist literature, context-based word representations can be derived analogically by clustering analyses of hidden unit activations of a recurrent network that has been trained on a next-word prediction task [3]. It has also been shown how various syntactic and semantic categories can emerge in the self-organizing map (SOM) when it is trained on context vectors [4]. Li, Burgess and Lund [5] have studied the effect of various model parameters (window size, corpus size and word dimensionality) on the quality of lexical maps when generated by SOM, trained on HAL vectors.

All the above mentioned approaches assume a lexicon of constant size. However, human speakers acquire a lexicon that develops incrementally over time. In this paper, we extend our previous work [6] to model the developmental process of lexical acquisition. We propose a model of a growing lexical map (GLM) whose size grows over time.

A number of growing self-organizing neural network models have been previously proposed to cope with the case of variable input space. However, most of these models have arbitrary dimensionality and connectivity (see, e.g., [7] and references therein) which makes them difficult to visualize in two dimensions. The IGG model [8] overcomes this difficulty in that it preserves a strictly 2D topology. Similarly to IGG, the insertion of new nodes in our GLM model is restricted to the grid positions. Unlike IGG, however, new nodes in GLM are inserted in between existing nodes instead of the perimeter of the grid.

In a psychologically plausible model of lexical development, it is necessary that (1) learning new words does not override the existing knowledge, and (2) the model remains well settled at all stages of development, but is plastic enough to learn new words. These characteristics require that GLM overcome catastrophic interference and the stability-plasticity dilemma. Various solutions have been proposed in the literature for overcoming catastrophic interference in connectionist networks [9]. In our approach we pretrain GLM on a subset of most frequent words that are expected to capture the regularities of the input space, which enable the addition of new words into the existing structure. The stability-plasticity trade-off is modulated in GLM by controlled ("see-saw" type) learning rate.

The model

Our model consists of three components (Fig.1). The first component is a special recurrent neural network, the word co-occurrence detector (WCD). The second component, non-trained two-layer network, reduces the data dimensions to a constant lower dimension. The third component is a growing neural network that shares some features with SOM [10]. It reads word representations and forms a 2D layout of the growing lexicon. The attempt to project high-dimensional word vectors onto a 2D space is motivated by known properties of the human cortex [11].

An initial assumption of the model is that we have available a pool of N (maximum lexicon size) localist word representations that correspond to the entire lex-

[1] also with IMS SAS, Bratislava, Slovakia

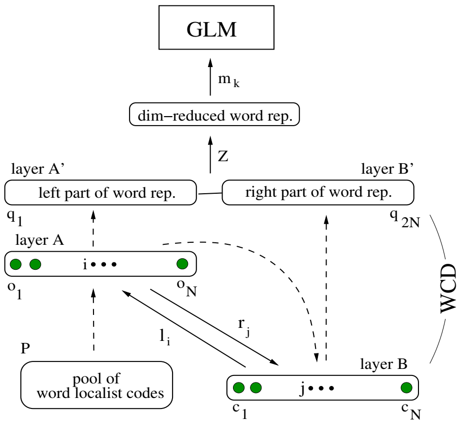

Figure 1: A diagrammatic sketch of the model. The bottom half represents WCD; in the upper part, $\mathbf{Z}$ is the random, dimension-reducing matrix of fixed connections, and GLM is a SOM-like neural network (with codevectors $\mathbf{m}_k$). The solid links between layers of units represent activity propagation via full connectivity weights, and dashed lines stand for pattern transport (via one-to-one links).

icon considered. We also assume that at earlier stages of learning, the known lexicon contains $n < N$ words, leaving thus the remaining units in WCD unutilized. As new words become acquired at later stages, n will grow. The whole learning process is split into two phases: (1) initialization, during which GLM is initialized (pre-trained) with a set of words (N_0), and (2) growth, during which new words are added to the existing lexicon. The growth process stops when size N is reached.

Word co-occurrence detector. Learning in WCD proceeds as follows (see [6] for an earlier version). Assume that at time t the current word is i ($i = 1, .., n$), and is represented by a localist vector $\mathbf{o} = [o_1, ..., o_N]$ in layer A. Previous word j is represented by vector $\mathbf{c} = [c_1, ..., c_N]$ in layer B. The adaptable connections between layers A and B serve to approximate the transitional probabilities between successive words, and as such, they are trained by Hebbian-like learning that normalizes them. Specifically, the link l_{ij} is updated to approximate $P(j^{t-1}|i^t)$, i.e., the probability that the word j precedes the word i. At the same time, the link r_{ji} is updated to approximate $P(i^t|j^{t-1})$, i.e., the probability that i follows j. Word i is then characterized by a concatenation of vectors

$$\mathbf{l}_i = [l_{i1}, l_{i2}, ..., l_{iN}], \quad \mathbf{r}_i = [r_{1i}, r_{2i}, ..., r_{Ni}], \qquad (1)$$

where $\mathbf{l}_i$ approximates the probability distribution of words preceding i (left context), and $\mathbf{r}_i$ the probability distribution of words following i (right context). The learning rules (with $0 < \beta < 1$) used for updating the connections have the form

$$\Delta l_{ij}^t = \beta o_i^t (c_j^t - l_{ij}^t), \quad \Delta r_{ji}^t = \beta c_j^t (o_i^t - r_{ji}^t). \qquad (2)$$

At each iteration (started by randomly picking a word from the pool), the distributed representation of the word appears at layers A' and B', ready to be processed further. [2]

Random mapping. With a growing lexicon the number of non-zero elements in $\mathbf{q}$'s also grows, which represents a difficulty for learning in GLM. Therefore, we make the dimensionality of $\mathbf{q}$'s constant by linearly mapping them using a random matrix $\mathbf{Z}$ with normalized Euclidean length of columns. Hence,

$$\tilde{\mathbf{q}}_i = \mathbf{Z}\mathbf{q}_i, \qquad i = 1, 2, ..., n. \qquad (3)$$

It has been recently shown that a random linear transformation of (sufficiently) high-dimensional data can preserve enough structure of the original data, if the output dimension is not too small [12]. Since $\mathbf{Z}$ (type $D \times 2N$) can be constant, its coefficients are not subject to adaptation and can thus be set *a priori*. In our simulations, we chose $D = 100$. [3]

Growing lexical map. During initialization, GLM contains a subset of nodes (e.g., 40% of all available nodes that fit into the underlying grid) and is first trained on a portion of the lexicon. All units are restricted to have grid positions and the map topology is induced by a triangulation procedure which says: "connect every pair of units in GLM, for which there exists a point in the map being closest to these two units in the map space." As in SOM, every GLM unit has an adaptable codevector $\mathbf{m}_k$ associated with it. Like in some other growing net models (referred to in [7]), every GLM unit k updates its error value

$$E_k(t) = E_k(t-1) + \|\tilde{\mathbf{q}}_i - \mathbf{m}_k\|^2, \qquad (4)$$

whenever it becomes the winner for input $\tilde{\mathbf{q}}_i$. A high E_k indicates that too many inputs map onto the node k. Once a predefined number of inputs has been presented, a new unit is added to one of randomly chosen, unoccupied grid spots around the unit with the largest E_k. [4] The new node becomes connected with surrounding units using the triangulation procedure (applied in the neighborhood of the new node), and its codevector is initialized as the average of the neighboring codevec-

[2] All units are linear. At every iteration, the model performs a sequence of operations each of which falls into one of the 3 categories: pattern transport (denoted by an arrow), activity propagation, and weight adaptation. The sequence is as follows: (1) transport previous word w^{t-1}: $A \rightarrow B$, (2) pick up (transport) a new word w^t: $P \rightarrow A$, (3) adapt $\mathbf{L}$ and $\mathbf{R}$ links (eq. 2), (4) $A \rightarrow B$, (5) propagate $\mathbf{o}^t = \mathbf{L}\mathbf{c}^t$ to layer A, (6) $A \rightarrow A'$ yielding $\mathbf{q}_i^L$, (7) pick up w^t again: $P \rightarrow A$, (8) propagate $\mathbf{c}^t = \mathbf{R}\mathbf{o}^t$ to layer B, (9) $B \rightarrow B'$ yielding $\mathbf{q}_i^R$, (10) propagate (and process) $\mathbf{q}_i = [\mathbf{q}_i^L, \mathbf{q}_i^R]$, further up (random mapping, etc.), (11) go to step 1.

[3] According to our PCA analyses of word representations, 100 dimensions cover more than 99% of the total variance. High-dimensional models [1,2] also use a reduced space, typically with 100-300 dimensions.

[4] As we consider a rectangular grid, the maximum number of candidates is 8, but preference is given to one of the 4 nearest neighbors found along the grid lines.

tors:

$$\mathbf{m}_{new} = \frac{1}{|\mathcal{N}_{new}|} \sum_{k \in \mathcal{N}_{new}} \mathbf{m}_k, \tag{5}$$

where $|\mathcal{N}_{new}|$ is the number of nodes linked to the new node. After a unit insertion, all Es are reset to zero.

Node deletion was not implemented in GLM. This is because new nodes are added only to areas that need them, and so each node tends to participate in the representation of one word or a blend of closest words in GLM.

The GLM employs the SOM procedures [10] for winner search and localized codebook update during initialization. However, we use a considerably smaller neighborhood radius $\mathcal{N}_c$, in order to avoid heavy reordering of the GLM codebook during development. For the same reason, we use a smaller learning rate α. Both $\mathcal{N}_c$ and α follow the "see-saw" profile. Specifically, α is always set to increase after a new unit has been inserted, and decays to zero toward the next insertion.

Experiment

We tested the model on the parental corpus of the CHILDES database [13]. We extracted the 300 most frequent words (roughly 500,000 word tokens) from the Belfast dataset [14] (see [5] for a description of the parental CHILDES corpus). All other words in the data were treated as a single unknown word.

Setup of initialization and the growth process. We ran 10 simulations using the off-line batch mode. [5] By "off-line" we mean separate training of WCD and GLM. By "batch" we mean the staircase-like lexicon growth, which we found easier to implement and evaluate. At first, WCD was trained on data corpus with an incremental lexicon from $N_0 = 200$ to $N = 300$, with a step 25 words. For every lexicon sized N_g ($g = 0, 1, ..., 4$), after a single pass through the data (using $\beta = 0.05$), the word representations were collected in an $N_g \times 2N$ matrix $\mathbf{Q}$, resulting in five different data matrices. All five $\mathbf{Q}$s were then transformed to $\tilde{\mathbf{Q}} = \mathbf{ZQ}$. The set of matrices $\tilde{\mathbf{Q}}^{(0)}$ through $\tilde{\mathbf{Q}}^{(4)}$ was used for all simulations. Once the input data was prepared, GLM was initialized with a set of 800 units randomly chosen from the underlying rectangular 45×45 grid (i.e., maximum number of map units $U = 2025$). GLM was pre-trained (like a classical SOM) with $\tilde{\mathbf{Q}}^{(0)}$ during 200,000 iterations. Pretrained GLM served as a common starting point for every single simulation that comprised four growth stages. Words were always randomly picked out from the corresponding data set.

Fig.2 shows an example of GLM at the end of stage $g = 2$, i.e., after 50 words were added. In all simulations, the parameters for the growth phase were as follows: Every growth epoch took 50,000 iterations, and within it, a new node could be inserted every 1,000 iterations (thus allowing the addition of maximum 50 nodes during every epoch). After every node insertion,

[5] This simplified training scenario helps to stabilize and to speed up learning.

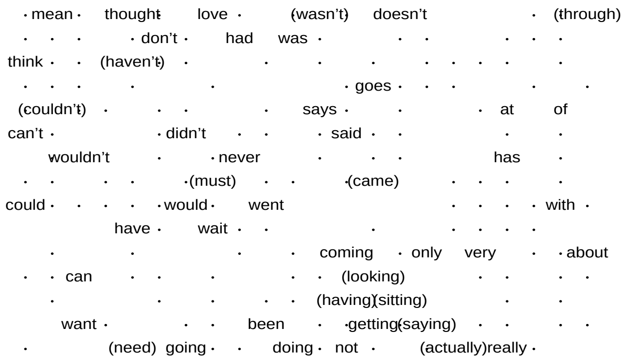

Figure 2: A snapshot of GLM in the middle of the growth process (only part of the map is shown due to space limit). Words added to the map during growth are embedded in parentheses. Either a label or a dot denotes an existing node. Some of the earlier acquired words may have slightly moved in GLM during growth, due to slight changes in their representations. The complete map currently has 856 nodes.

α was raised to 0.05 and then linearly decreased to zero. Similarly, neighborhood radius $\mathcal{N}_c$ was raised to 2 and then gradually decreased to zero until the next node insertion.

Growing map evaluation. In each simulation, the GLM configuration was saved at the end of every growth epoch g. For map comparison, we focused on evaluating how much the two maps $\mathcal{M}^{g1}$ and $\mathcal{M}^{g2}$ differed in their output responses when tested on words common to both maps. The output of GLM for each word i was computed as a global output activity $\mathbf{a} = [a_1, ..., a_U]$ centered around the winning unit c, yielding a coarse-coded representation with components

$$a_k(i) = \begin{cases} \exp(-\|\tilde{\mathbf{q}}_i - \mathbf{m}_k\|/\sigma), & \text{if } k \in \mathcal{N}_c, \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where $\mathcal{N}_c$ denotes the neighborhood (its radius was set to 5 in simulations) around the unit c, σ was set to 1, and all $\mathbf{m}_k$ corresponding to non-existing node positions were kept to zero. [6] For every common word i, both output responses $\mathbf{a}^{g1}(i)$ and $\mathbf{a}^{g2}(i)$ were computed. If the responses of the two maps to this word were closer to each other (in terms of Euclidean distance) than to any other word j in the lexicon, i.e., if

$$\|\mathbf{a}^{g1}(i) - \mathbf{a}^{g2}(i)\| = \min_j \{\|\mathbf{a}^{g1}(i) - \mathbf{a}^{g2}(j)\|\}, \tag{7}$$

then word i was considered to have preserved its representation in the map. Otherwise, a mismatch occurred, which increased the word mismatch (WM) evaluated as WM = #mismatches/#words_checked.

Results. The GLM outputs for a pair of GLMs have been compared for all combinations of $\mathcal{M}^{g1}$ and $\mathcal{M}^{g2}$. Values of all average WM values are shown in Fig.3. As expected, all mismatches that occurred were related to word couples whose labels were close to each other

[6] We also tried slightly different values of $\mathcal{N}_c$ and σ, but the evaluation results were similar.

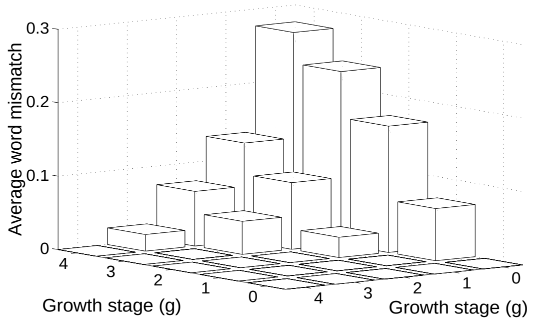

Figure 3: Average WM for combinations of all pairs of GLMs at various growth stages. Each bar corresponds to a pair of GLMs whose outputs were compared on all common words. The average WM increases to the highest at the 4-0 pair. This suggests that word representations gradually change during growth, yielding lower errors for temporarily neighboring maps (such as a 3-2 pair).

in GLM. In most cases, the confused words belonged to the same grammatical category (e.g., *mummy* and *daddy*, or *wouldn't* and *didn't*); in other cases, they were not grammatically related, though placed next to each other (e.g., *one* and *home*, or *the* and *our*). A slight improvement could be obtained if all maps had been given some extra fine-tuning in order to decrease the number of ambiguous units (i.e., those having two labels).

Gradual changes of "old" word representations (and consequently, their slight shifts in GLM during growth) may correspond to the real learning situation in humans who also adapt their word semantic representations as they hear the words in newer contexts.

Discussion

In this paper, we present a growing lexical map, a model that addresses the task of incremental acquisition of the lexicon. As we consider word representations based on context words, we solve the problem of increasing word dimension by mapping the words to a constant lower dimension, while preserving their mutual relationships. We tested the model on a portion of the parental CHILDES corpus, and found that the model is able to learn new words while highly preserving the learned structure.

We consider it an important virtue of the model to allow the acquisition of new words without causing catastrophic intereference on learned structure. Human language learners assimilate new words rapidly – according to one estimate, the average child learns some 14,000 words by age six [15]. In our model, new words are added to the existing structure across stages of growth. The model has the ability to avoid catastrophic interference and the stability-plasticity dilemma.

As an initial attempt to model an incremental lexicon, we have used only a small vocabulary in our data set (300 most frequent words). In principle, our model can be used to simulate the development of a lexicon of potentially larger size (with some modifications). This is because the network does not have inherent resource limitations (although the size of the underlying map grid has to be chosen in advance). In previous self-organizing connectionist models of language acquisition, researchers needed to face the resource limitation problem, especially if they started with a limited number of nodes in training [16]. It is empirically debatable whether the growing network architecture as the one in our model has biological plausibility (but see [17]), but it certainly has psychological plausibility – for one thing, we know that working memory capacity gradually develops in young children.

Acknowledgments

Supported by an NSF grant (#BCS-9975249) to P.L.

References

[1] C. Burgess and K. Lund. Modelling parsing constraints with high-dimensional semantic space. *Language and Cogn. Processes*, 12(2/3):177–210, 1997.

[2] T.K. Landauer and S.T. Dumais. A solution to Plato's problem: the latent semantic analysis theory of induction and representation of knowledge. *Psychological Review*, 104:211–240, 1997.

[3] J. Elman. Finding structure in time. *Cognitive Science*, 14:179–211, 1990.

[4] H. Ritter and T. Kohonen. Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254, 1989.

[5] P. Li, C. Burgess, and K. Lund. The acquisition of word meaning through global lexical co-occurrences. In *Proc. of the 30th Child Language Research Forum*, pp. 167–178. Stanford, CA: CSLI, 2000.

[6] I. Farkaš and P. Li. A self-organizing neural network model of the acquisition of word meaning. In *Proc. of the 4th Int. Conf. on Cogn. Modeling*, pp. 67–72, 2001.

[7] F.H. Hamker. Life-long learning cell structures – continuously learning without catastrophic interference. *Neural Networks*, 14(4/5):551–573, 2001.

[8] J. Blackmore and R. Miikkulainen. Visualizing high-dimensional structure with the incremental grid growing neural network. In *Machine Learning: Proc. of the 12th Int. Conf.*, pp. 55–63, 1995.

[9] R.M. French. Catastrophic interference in connectionist networks: Causes, consequences and solutions. *Trends in Cogn. Sciences*, 3(4):128–135, 1999.

[10] T. Kohonen. *Self-Organizing Maps*. Springer Verlag, 1997. Second edition.

[11] E.I. Knudsen, S. du Lac, and S.D. Esterly. Computational maps in the brain. *Annual Review of Neuroscience*, 10:41–65, 1987.

[12] S. Kaski. Dimensionality reduction by random mapping: Fast similarity computation for clustering. In *Proc. of the IJCNN*, pp. 413–418, 1998.

[13] B. MacWhinney. *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum, Hillsdale, NJ, 2000.

[14] A. Henry. *Belfast English and standard English: Dialect variation and parameter setting*. Oxford University Press, New York, 1995.

[15] S. Carey. *Linguistic theory and psychological reality*, chapter The child as word learner, pp. 264–293. Cambridge University Press, Cambridge, 1978.

[16] P. Li. Language acquisition in a self-organizing neural network model. In P. Quinlan (ed.), *Connectionism and Developmental Theory*. Psychology Press, Philadelphia and Briton, in press.

[17] E. Gould, P. Tanapat, N.B. Hastings, and T.J. Shors. Neurogenesis in adulthood: a possible role in learning. *Trends in Cogn. Sciences*, 3:186–191, 1999.