In Proceedings of the 4th Int. Conf. on Cognitive Modeling, Fairfax, VA, pp. 67-72, 2001

A Self-Organizing Neural Network Model of the Acquisition of Word Meaning

Igor Farkaš (ifarkas@richmond.edu)
Ping Li (pli@richmond.edu)
Department of Psychology
University of Richmond
VA 23173, USA

Abstract

In this paper we present a self-organizing connectionist model of the acquisition of word meaning. Our model consists of two neural networks and builds on the basic concepts of Hebbian learning and self-organization. One network learns to approximate word transition probabilities, which are used for lexical representation, and the other network, a self-organizing map, is trained on these representations, projecting them onto a 2D space. The model relies on lexical co-occurrence information to represent word meanings in the lexicon. The results show that our model is able to acquire semantic representations from both artificial data and real corpus of language use. In addition, the model demonstrates the ability to develop rather accurate word representations even with a sparse training set.

Introduction

A central debate in the domain of language acquisition is how children acquire the meanings of words. Although numerous studies have addressed this question in the last few decades, researchers have not yet reached any consensus. One important line of disagreement is whether children can use contextual or structural knowledge from the sentence to bootstrap their learning of the semantic contents of words. Proponents of the *syntactic bootstrapping hypothesis* (e.g., Gleitman, 1990) argue that children can and do make use of the structural information to learn word meaning, whereas advocates of the *semantic bootstrapping hypothesis* (e.g., Pinker, 1994) are suspicious of such an approach that relies on the child's distributional analysis of the input.

In recent years, connectionism and computational analyses of large-scale corpora have revitalized the interest in structural relationships and distributional analyses of language. Independent of proposals like the syntactic bootstrapping hypothesis in child language, research has revealed the power of distributional information in deriving accurate representations of the meaning and function of linguistic components. In particular, it is argued that both grammatical and semantic categories can be acquired by connectionist networks or similar statistical machines through the computation of the statistical regularities inherent in the input data. For example, Elman (1990) showed that categories of nouns and verbs, and subcategories of animates versus inanimates (within nouns), and transitives versus intransitives (within verbs), can emerge from a simple recurrent network's analyses of the lexical co-occurrence properties in the input. Redington, Chater, and Finch (1998) demonstrated that the use of distributional properties in a large-scale speech corpus allows a statistical system to derive grammatical categories. Landauer and Dumais (1997) showed that it is also possible to accurately represent semantic relationships through a high-dimensional word-to-text matrix.

Burgess and Lund (1997, 1999) proposed a high-dimensional space model to represent the meaning of the lexicon. Their model, the Hyperspace Analogue to Language (HAL), attempts to capture meaning by reference to global lexical co-occurrences – how many words co-occur with the target word, and how often, in a large moving window that runs through the text. A co-occurrence matrix for any number of words in a given window is derived, and weighted by the frequency of co-occurrence among the words. The columns and rows in this matrix represent the co-occurrence values for words that follow and precede the target, respectively. The target word is then represented by concatenating the column and row values. Burgess and Lund claim that this method captures the global lexical constraints for words, and the constraints reflect the total contextual history of a word in a high-dimensional space of language use. In an attempt to apply this method to the acquisition of word meaning, Li, Burgess, and Lund (2000) analyzed the 3.8 million word tokens of parental speech in the CHILDES English database (MacWhinney, 2000) and found that it is possible to derive accurate semantic representations given a reasonable size of corpus such as the CHILDES adult speech (rather than a very large corpus such as the Usenet data for the original HAL model). The implication is that young children can acquire word meanings if they exploit the considerable amount of contextual information in the linguistic input. However, this study, like HAL, does not qualify as a true developmental model, because no learning was involved in arriving at the representations – only statistical analyses of the data were involved (e.g., window size, corpus size, and the constraint dimensions were manipulated at each stage). In short, HAL is a representation model and not a processing or learning model, as Burgess and Lund (1999) pointed out.

In this study, we present a self-organizing neural network model that can learn semantics from linguistic input. The basic idea is similar to HAL, but there are two major distinctive features to our model: (1) it is based on unsupervised neural networks that learn on line, (2) it incorporates a mechanism that leads to accurate word representations (and consequently meaningful lexical maps) even when the training data are sparse. Our model builds on the basic concepts of self-organization and Hebbian learning (Kohonen, 1990; Miikkulainen, 1993), and incorporates ideas from semantic categorization in feature maps (Ritter & Kohonen, 1989; Li, 1999, 2000). [1] Preliminary results show that our model is able to acquire rather accurate semantic representations from both artificial data and real corpus of language use.

Method

Our model consists of two neural networks that functionally interact with each other. Fig. 1 presents a diagrammatic sketch of the model. The lower part is a special recurrent neural network, the word co-occurrence detector (WCD), whose modifiable connections are trained to approximate word transitional probabilities. The upper part is a self-organizing map (SOM; Kohonen, 1990), which reads the words distributively represented in the modifiable connections and creates a two-dimensional layout of the lexicon.

An initial assumption of the model is that we have a pool of N localist word representations corresponding to the lexicon. Whenever the word w_i is read, the corresponding unit in layer A becomes activated, creating localist representation $\mathbf{o} = [o_1, ..., o_N]$. At the same time step, layer B holds the previous word w_j (context) represented by vector $\mathbf{c} = [c_1, ..., c_N]$, which was copied over there from layer A in previous time step.

The algorithm

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 learning with implicit normalization to become probabilities. Therefore, two co-occurrence matrices instead of one are used in this model. Assume that at time t, the current word is w_i, and is preceded by word w_j. At every time step, both $\mathbf{l}$ and $\mathbf{r}$ links are modified. Specifically,

[1]Lowe (1997) presented a similar model to simulate semantic priming. Developed independently of his research, our model differs from his in implementation details, and focuses on different theoretical issues.

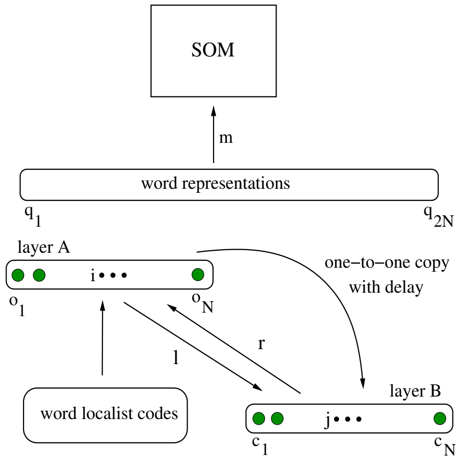

Figure 1: The architecture of the model. Layers A and B of the WCD (see the text) have full connectivity via modifiable $\mathbf{l}$ and $\mathbf{r}$ links of the WCD. Other, one-to-one links serve to feed the unit activity from A to B with (discrete single time-step) delay. The SOM is a self-organizing neural network trained on distributed word representations extracted from modifiable links.

the link l_{ij} is updated to approximate $P(w_j^{t-1}|w_i^t)$, i.e. the probability that the word w_j preceded the word w_i. At the same time, the link r_{ji} is updated to approximate $P(w_i^t|w_j^{t-1})$, i.e. the probability that the word w_i follows the word w_j. Each word is then characterized by a concatenation of vectors

$$\begin{aligned} \mathbf{l}_i &= [l_{i1}, l_{i2}, ..., l_{iN}] \\ \mathbf{r}_i &= [r_{1i}, r_{2i}, ..., r_{Ni}] \end{aligned}$$

where $\mathbf{l}_i$ approximates the probability distribution of words preceding w_i (left context), and $\mathbf{r}_i$ the probability distribution of words following w_i (right context). Each word is thus represented by a real-valued vector $\mathbf{q}_i = [\mathbf{l}_i, \mathbf{r}_i]$ of dimensionality $2N$.

The learning rules used for updating the connections have the form

$$\begin{aligned} \Delta l_{ij}^t &= \beta o_i^t (c_j^t - l_{ij}^t) \\ \Delta r_{ji}^t &= \beta c_j^t (o_i^t - r_{ji}^t) \end{aligned}$$

where $0 < \beta < 1$ is the learning rate. [2] Simultaneously with $\mathbf{q}_i$ update, the SOM is also trained on $\mathbf{q}_i$ as inputs. Every SOM unit k has an array of connections in the form of a codevector $\mathbf{m}_k = [m_{k1}, ..., m_{k,2N}]$ associated with it, which learns to approximate the inputs in such a manner that every SOM unit tends to become "specialized" for a concrete word $\mathbf{q}_i$, and that neighboring units will become specialists ("winners") to similar words.

[2]Basically, these learning rules work as counters, as in the HAL model, except that normalization is performed simultaneously with counting.

The SOM algorithm is standard (Kohonen, 1990). At every time step, the winner k^* is found to satisfy the condition $k^* = \arg\min_k\{\|\mathbf{q}_i(t) - \mathbf{m}_k(t)\|\}$, where $\|.\|$ denotes the Euclidean distance, and then all codevectors within winner's neighborhood are shifted towards the current input via

$$\Delta\mathbf{m}_k(t) = \alpha(t)\ [\mathbf{q}_i(t) - \mathbf{m}_k(t)]\ ,$$

for all k in the neighborhood of k^*. During learning, both neighborhood radius and learning rate $0 < \alpha < 1$ decrease in time.

It is reasonable to start training the SOM codevectors at a relatively later stage of the WCD network. The reason is that the SOM sees statistically accurate word representations only towards the end of training, so delayed update of codevectors facilitates their ordering and convergence.

Generalization

To extend the basic algorithm, we would like the model to have some generalization property for novel word transitions. So far the links would be updated only between adjacent words seen in the training data. For example, if the model has been trained on word strings like `John sees` and `Mary loves` we would like the model to have non-zero prediction probabilities (stored in SOM connections) when processing the word strings `John loves` or `Mary sees`, which have never occurred in the training data. It is well known that human learners are able to make this type of generalization (Fodor & Pylyshyn, 1988; Elman, 1998).

Our hypothesis is that words that occur in the same contexts will tend to have similar vectors, as our model represents words by distributed vectors $\mathbf{q}_i$ that incorporate context information. We can exploit these vector similarities to obtain generalization properties for novel word transitions.

We applied the following mechanism in the later phases of training when the SOM units are expected to have already established some global order. At each time step, the winner for the current word is found in the SOM. A few units j among its neighbors are also identified, which have the status of being the winner for any word. This requires a unit labeling procedure running on line (based on majority voting, so that every unit could have only one word label associated with it). Next, for each SOM unit k, the corresponding output units o_j in the WCD network are set to one, thus enabling the update of their connections. This strategy enables an extended update of connections from *previous-word neighborhood* to *current-word neighborhood*, instead of simply from previous word to current word. As a consequence, more accurate $\mathbf{q}_i$'s can be obtained even when the WCD network sees only a fraction of word transitions in the training data. [3]

Normalized negative log-likelihood To evaluate the model's generalization ability, we measure how the model can generalize to previously unseen word strings. We use the normalized negative log-likelihood (NNL), a commonly used method for performance measure in symbol prediction tasks (Ron, Singer, & Tishby, 1996). This is done as follows. The parameters of the model are fixed after training and for every word in the vocabulary there is a corresponding winner among the SOM units. In a sequence of words $W = w_1w_2...w_s$, every time the model sees the word $w(t)$, we find its winner k^* in the SOM, and the estimate of its next correct-word probability $\hat{P}(w^{t+1}|w^t) = m_{k^*,j+N}$ is read out (from the right-context part of connections' array), where j is an index of w^{t+1}. Hence, for every model $\mathcal{M}$ and the test sequence W_{test} we evaluate

$$NNL_{\mathcal{M}}(W_{test}) = \frac{-1}{s-1}\sum_{t=1}^{s-1}\log_N \hat{P}(w^{t+1}|w^t)\ ,$$

where the base of the logarithm equals the number of words in the lexicon. The higher the next correct-symbol probabilities are, the lower the *NNL* is, and vice versa. If $NNL = 0$, prediction accuracy is 100%; if $NNL = 1$, the distribution of probabilities is uniform.

Results

Artificial corpus

We tested our model on data created by a simple language generator (SLG, Rohde 1999). Compared to the well-known Elman 29-word data set (Elman, 1990), our data set was slightly more complex (with 45 words). We added plurals, optional adjectives and determiners to SV(O) sentences, which allowed us to generate more complex sentences such as `the hungry lion chases boys`, `girls sit-in a bus`, and `a dog barks`.

Out of the hundreds of sentences generated by SLG we used 435 unique sentences so that none of the sentences was repeated. All sentences had the end-of-sentence mark (EOS) as an additional symbol. Sentences were presented to the model in random order, one word at a time for each sentence. Learning started in the WCD network ($\beta = 0.005$), and during the second half of the training, learning also took place in the SOM network. Figure 2 shows the SOM for all 45 words. Clearly, word representations based on $\mathbf{q}$'s from our model provide a considerable amount of information, sufficient for the data to be correctly mapped onto a 2D topology preserving space according to syntactic as well as semantic categories.

[3]This method may be thought of as a kind of smoothing the bigram probability estimates based on words with similar statistics, as used in statistical NLP (e.g., Manning & Schütze, 1999).

Figure 2: The SOM trained on artificial data. The network identified various grammatical as well as some semantic categories.

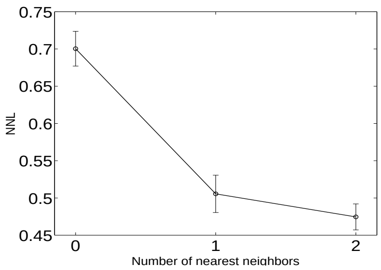

Figure 3: Prediction accuracy of the model during test on artificial data. Inclusion of neighbors in connections' update helps to derive more accurate word representations even with a small amount of training data.

Next, we investigated the generalization property of the model. We wanted to model the fact that human learners (children in particular) process a fraction of all possible (meaningful) word combinations and are able to produce and understand novel word combinations. To simulate this ability, the model should have non-zero next-word prediction probabilities so that new word associations can occur.

The model was trained on a very small portion of the data (81 sentences) and tested on the remaining data (354 sentences). The effect of neighborhood update becomes even more visible in this case. Random splitting was performed 10 times and the results were averaged. As shown in Fig. 3, *NNL* was smaller, if we included at least one nearest neighbor in adaptation. However, higher numbers of nearest neighbors did not improve accuracy, since they already tended to induce non-existing word transitions in word representations. Improvement in accuracy could also be observed in the corresponding maps.

Realistic corpus

To test whether our model can scale up to realistic data, we examined our model's performance on the parental/caregivers' speech in the CHILDES database (see Li, Burgess, & Lund, 2000 for a description of how the data were extracted). We took parental speech from the Wells corpus (Wells, 1981) and used the 300 most frequent words (roughly 150,000 word tokens) from the data. All other words were treated as a single unknown word w_x in the lexicon (hence $N = 301$). However, because of the relatively high frequency of the unknown words (33%) in the data, treating them all as one would induce a bias toward w_x. To correct this word imbalance, we imposed a probability restriction on w_x: whenever the word w_x was read, the update of connections occurred only in 1% of cases; in other cases reading was skipped to the next word (this might simulate the process of treating unknown words as noise).

Training on this large-scale data set was rather time-consuming. To speed up learning, we first trained the WCD network to develop word representations ($\beta = 0.01$) and then trained the SOM off-line on converged word representations.

Fig. 4 presents the SOM that was trained on the CHILDES Wells corpus for two epochs. Upon closer examination, one can observe various grammatical categories that were clustered in the map: proper nouns, verbs, auxiliary verbs, adjectives, pronouns, etc. Semantic similarities also emerged within the categories.

Discussion

An important premise behind the syntactic bootstrapping hypothesis is that children can acquire word meanings through their distributional analyses of the linguistic input. The distribution-analysis approach has a long tradition in linguistics and psycholinguistics. In fact, much of the pre-Chomsky linguistics is the so-called structuralism that examines the structural relationships between linguistic units (Bloomfield, 1935). Saussure (1916) proposed that the function of a given linguistic "entity" (e.g., a word) is defined entirely by reference to the relationships that hold between this entity and other entities, much like that the role of a chess piece is determined by its relationship with other pieces on the chessboard. For example, words of the same class tend to occupy the same slot in a sentence (*paradigmatically* similar) and have the same co-occurrence constraints with other words (*syntagmatically* similar). [4] Chomsky (1957), however, treating structuralism as purely associationist, threw out this approach in linguistics and replaced it with an emphasis on higher-order hierarchical relationships for linguistic structures. Thus, few linguists today believe that associationism or structuralism could work for language.

[4] In this context, we can claim that our model places paradigmatically similar words close to each other in the map, whereas syntagmatically similar words can be far from each other. However, the mutual associations between the latter are captured in the SOM codevectors in the form of prediction probabilities.

Figure 4: The SOM (40×40 units) of 300 most frequent words in a portion of the CHILDES parental data. Various grammatical categories are distributed across the map, showing considerable structure. For example, one cluster of nouns can be located in the top left part of the map, whereas a group of proper nouns can be found in the center. Adjectives are located in top right corner. Pronouns are mainly placed in top middle part. Verbs are spread mainly over the bottom half of the map space and within this category, one can identify auxiliary verbs in the bottom right part, etc. Also, semantically similar words are often positioned to neighboring units in the map. For example, `mummy's,daddy's` in top middle part, `now,then` and `there,here` bottom right, `up,down` bottom left, and so on.

Connectionist approaches represent a new way of looking at associations and structural relationships in language. Our model, along with many other similar statistical models of NLP, appears to fare well in capturing the higher-order relationships in language without recourse to strictly symbolic rules and hierarchies. The model we presented here shows that lexical semantics, traditionally considered a difficult part of language, can indeed emerge from the learning in connectionist systems, in particular, self-organizing neural networks (see also Li, 1999, 2000, in press). Some researchers may also view our model as a connectionist implementation of statistical NLP, because we examine many of the same issues there.

We have demonstrated in this paper that it is possible to learn meanings of words through the learning of co-occurrence statistics, in line with the syntactic bootstrapping hypothesis and the HAL model. Our model learns very simple co-occurrence constraints, and acquires syntactic and semantic categories, in both artificial and realistic linguistic corpora. Our model is also able to display generalization characteristics, that is, to learn novel combinations of word strings that are absent in the training data. The simulation results also indicate that the model does not require a very large amount of data to arrive at accurate semantic representations, contrary to what is commonly expected of statistical learning models.

Self-organizing maps provide an important tool for us to capture lexical relationships, as demonstrated in this study. There is a lot of information in the SOM connections (both lateral and input) which quantitatively describe how the words are associated with each other. Although our model does not directly incorporate lateral connections, these connections can replace the grid topology in the SOM and can be used to represent more complex semantic relationships. One future research direction is to implement lateral connections in our model.

Several other directions/limitations will also be considered in extending the current model. First, the dimension of the word representations grows with the number of words, which makes it difficult to scale up to a very large lexicon. It is a challenge to overcome the initial localist representation while preserving lexical identities. Second, in the current model, the information for SOM is extracted from WCD connections, thus making the model somewhat unusual in terms of information access and flow. Therefore, it would be useful to transform this information to unit activations which then feed to the SOM. Third, the current model is not designed to learn lexicon incrementally; that is, it is unable to take new words to the existing lexicon during learning. Finally, the model uses only the shortest window of context (immediately before and after the target word) to derive semantic representations. Experiments in the HAL model show that larger window sizes yield more accurate word representations (Li, Burgess, & Lund, 2000). In English, this parameter may appear less significant due to the strict word order; in other languages with relatively flexible word order (e.g., Chinese, Italian), the size of the context window may prove to be very important.

Acknowledgments

This research was supported by an NSF grant (#BCS-9975249) to P.L. We are grateful to Curt Burgess, Brian McWhinney, Risto Miikkulainen, and Peter Tiňo for comments and discussions at various stages of the project, and to three anonymous reviewers for final comments. I.F. is also with the Slovak Academy of Sciences, Bratislava, Slovakia.

References

Bloomfield, L. (1935). *Language*. London: Allen & Unwin.

Burgess, C. & Lund, K. (1997). Modelling parsing constraints with high-dimensional semantic space. *Language and Cognitive Processes*, 12, 1–34.

Burgess, C. & Lund, K. (1999). The dynamics of meaning in memory. In Dietrich, E. and Markman, A. (Eds.), *Cognitive Dynamics: Conceptual and Representational Change in Humans and Machines*, Lawrence Erlbaum, Hillsdale, NJ.

Chomsky, N. (1957). *Syntactic Structures*. The Hague: Mouton.

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

Elman, J. (1998). Generalization, simple recurrent networks, and the emergence of structure. In Gernsbacher, M. & Derry, S. (Eds.), *Proceedings of the 20th Annual Conference of the Cognitive Science Society*. Lawrence Erlbaum.

Fodor, J. & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28, 3–71.

Gleitman, L. (1990). The structural sources of verb meaning. *Language Acquisition*, 1, 3–55.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, 78, 1464–1480.

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

Li, P. (1999). Generalization, representation and recovery in a self-organizing feature map model of language acquisition. In Hahn, M. & Stoness, S. (Eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp. 308–313). Mahwah, NJ: Lawrence Erlbaum.

Li, P. (2000). The acquisition of lexical and grammatical aspect in a self-organizing feature map model. In Gleitman, L. & Jashi, A. (Eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp. 304-309). Mahwah, NJ: Lawrence Erlbaum.

Li, P. (in press). Language acquisition in a self-organizing neural network model. In Quinlan, P. (Ed.), *Connectionist Models of Development*. Psychology Press, Philadelphia and Briton.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In Clark, E. (Ed.), *Proceedings of the 30th Child Language Research Forum.* (pp. 167–178). Stanford, CA: Center for the Study of Language and Information.

Lowe, W. (1997). Semantic representation and priming in a self-organizing lexicon. In *Proceedings of the 4th Neural Computation and Psychology Workshop* (pp. 227–239). Springer-Verlag.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk*. Hillsdale, NJ: Lawrence Erlbaum.

Manning, C. D. & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

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

Pinker, S. (1994). How could a child use a verb syntax to learn verb semantics? *Lingua*, 92, 377–410.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, 22, 425–470.

Ritter, H. & Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61, 241–254.

Rohde, D. (1999). *The simple language generator: Encoding complex languages with simple grammars* (Tech. Rep. CMU-CS-99-123). Pittsburg, PA: Carnegie Mellon University.

Ron, D., Singer, E., & Tishby, N. (1996). The power of amnesia: learning probabilistic automata with variable memory length. *Machine Learning*, 25, 117–149.

Saussure, F. de (1916). *Cours de Linguistique Générale*. Paris: Payot.

Wells, C.G. (1981). *Learning Through Interaction: The Study of Language Development*. Cambridge: Cambridge University Press.