J.Basrah Researches (Sciences) Vol.31. Part.3.21-27 (2005)

Self- Organizing Map for Arabic Language Understanding

Abdulkareem Y. Abdalla

Computer Science Department
College of ScienceUniversity of Basrah
Basrah- Iraq.

Received 22/3/2004, Accepted 2/10/2004

Abstract

In recent years, there has been a resurgence in research on neural netwoks in natural language processing. These research employ learning techniques to automatically extract linguistic knowledge from natural language rather then require the system developer to manually encode the requisite knowledge.It is known that the biological brain contains various topographically ordered 'maps', such that different neural cells respond optimally to different signal qualities.In this paper, self-organizing map are used for linguistic representation. The performance of the system is shown for small arabic vocabulary. It constructs an optimal topographical mapping of the words in the arabic sentences which gives high accuracy in language understanding.

Introduction

Sentence understanding belongs to the broader category of natural language processing for which, during the past tens of years, many methods have been tried.Since the operation of the brain at the higher levels relies heavily on abstract concepts, symbolism, and language. It is an old notion that the deepest semantic elements of any language should also be physiologically represented in the neural realms [1]. The "internal representation" of the brain is obtained from the fact that the cortex of the biological brain is essentially a two dimensional sheet, and many of its areas are specialized to different sensory modalities. In these areas, the various cells seem to respond to many abstract qualities of the sensory stimuli in an orderly fashion. For instance, in the semantic areas, there is a scale for different semantic elements, and in the visual areas, one can find a colour map, maps for orientation of line systems, etc.In [1], a neural mapping principle, called self-organizing map, is able to extract automatically a few (usually two) of the most important feature dimension of a multi-dimensional signal space and to display the input vectors in a two dimensional map.

In this paper, the self-organizing maps have been applied to linguistic representation of arabic words. They may bear some similarity to the various feature maps that exist in the biological brain.

A brief history of natural language research

One of the biggest challenges in natural language processing is how to provide a computer with the linguistic sophistication necessary for it to successfully perform language-based tasks. The primary goal of Artificial Intelligence (AI) has been the development of computational methods for natural language understanding. Early research in machine translation illustrated the difficulties of this task with sample problems such as translating the word pen appropriately in " The box is in the pen" versus " The pen is in the box" [2]. It was quickly discovered that understanding language required not only lexical and grammatical information, but semantic, pragmatic, and general world knowledge.Nevertheless, during the 1970, AI systems, which demonstrated interesting aspect of language understanding in restricted domain, were developed [3-5]. During the 1980's, there was a continuing progress on developing natural language systems using hand-coded symbolic grammar and knowledge bases [6].

However, developing these systems remained difficult, requiring a great deal of domain-specific knowledge engineering. In addition, the systems were brittle and could not function

adequately outside the restricted tasks for which they were designed. Partially in reaction to these problems, in recent years, there has been a paradigm shift in natural language research. The focus has been shifted from the rationalist methods based on hand-coded rules, derived to a large extent through introspection to the empirical, in which development is much more data driven and is at least partially automated by using statistical or machine-learning methods to train systems on large amounts of real language data[7].

One of the major styles of empirical methods is neural network, when neural nets were originally popular in the 1950's, most of the research concerned with visual pattern recognition, and language learning were not well represented. However, with the revival of neural nets in the 1980's, applications to language were visible [8-13].

Artificial neural networks

Artificial neural networks can be considered as a massively parallel distributed model that has a natural property of storing experimental knowledge and making it available for use. They represent mathematical models of brain-like systems where knowledge is received through a learning process.The basic processing unit of a neural network is the neuron. As illustrated in figure (1), a neuron i consists of a set of N connecting links, the synapses, which characterized by weights wij. Each input signal xj applied to the synapse j is multiplied by its corresponding weight wij and transmitted to neuron j. All these synapse products are

$$U_i = \sum_{j=1}^{n} w_{ij} x_j \qquad \ldots \quad (1)$$

accumulated by the adder as expressed by the equation:-
An activation function g( ) provides to the output y_i of the unit as :-

$$y_i = g(u_i) \qquad \ldots \quad (2)$$

The activation function, defines the output of the neuron. The most common type of this function is the sigmoid function.

$$g(x) = \frac{1}{1+\exp(-\alpha x)} \qquad \ldots \quad (3)$$

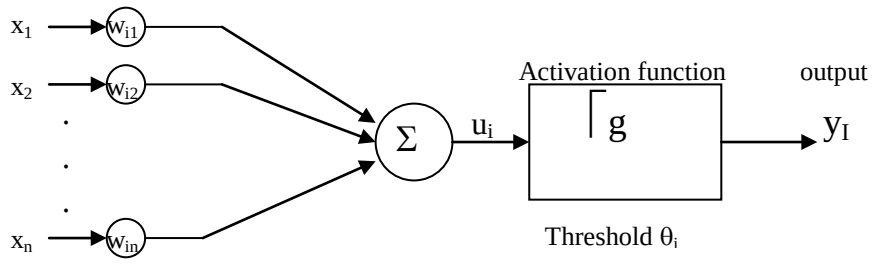

Figure (1): Basic neuron model

A neural network is characterized by its architecture and the learning algorithm used to train it. The network architectures used to model nervous systems can roughly be divided into three categories, each based on a different philosophy.
Feed forward neural network [14] transforms sets of input signals into sets of output signals. The desired input-output transformation is usually determined by supervised adjustment of the system parameters. In feedback neural networks [15], it has one or more feedback loops, in the sense that each neuron feeds its output signal back to the inputs of all the other neurons.

In the third category, neighboring cells in a neural network compete in their activities by means of mutual lateral interactions, and develop adaptively into specific detectors of different signal patterns. In this category, learning is called competitive, unsupervised, or self-organizing [1].The learning process in a neural network implies the adjustment of the weights of the learning network in an attempt to minimize an error function suitable for the type of the network used. Considering that $w_{ij}(t)$ denotes the synaptic weight at time t and that an adjustment $\Delta w_{ij}(t)$ is applied to this weight at the same time, the updated value $w_{ij}(t+1)$ indicates the new weight at the next processing time. The above statement can be formally

$$w_{ij}(t+1) = w_{ij}(t) + \Delta w_{ij}(t) \quad \dots \quad (4)$$

written as :-
There are two basic classes of learning algorithms:
Supervised learning and unsupervised learning algorithms. During supervised learning an external teacher is provided to the network that holds the knowledge of the correct output for each input pattern assigned to the network. The network actual response, $y_i(t)$, and the corresponding desired output value $d_i(t)$ allow the computation of an error $e_i(t)=d_i(t) - y_i(t)$. This error value is used to adjust the weights of the network. The convergence criterion of a supervised learning algorithm is the minimization of some error function for the inputs used for the network training. The back propagation learning algorithm is an example of supervised learning algorithm. In unsupervised learning no knowledge from a teacher is available and the network must perform a kind of self-organized learning. As the network is not aware of the desired output values, it examines the input patterns according to some local measurements of similarity of degrees of quality; a division of the input set into a number of self-tuned groups. The competitive learning rule is a kind of unsupervised learning that performs a competition among the output neurons of a network with the result that only one output neuron is activated (fires) at each time. A special and increasingly popular class of neural networks based on competitive learning is the class of self-organizing feature maps, that are characterized by a topographic map of the input patterns, such that the coordinates of the neurons upon a lattices correspond to features of the input patterns. It must be noted that competitive learning can be also used as supervised learning.

The self-organizing map for arabic sentence understanding

The self-organizing neural networks, developed by kohenon [1], assume a topological structure among the cluster units. This property observed in the brain, but not found in other artificial neural networks. There are m cluster units, arranged in a one or two-dimensional array; the input signals are n-tuples.

During the self-organization process, the cluster unit whose weight vector matches the input patterns most closely chosen as the winner. The winning unit and its neighboring units update their weights.

1- Network architecture

The network structure for arabic sentences understanding is a self-organizing neural network with a rectangular two-dimensional network of cells as shown in figure (2). Let $x=[x_1, x_2, \dots, x_n] \in R^n$ be the input vector. The number n of the elements of the input depends on the size of the training data set. The weight vector of cell i denoted by $m_i = [m_{i1}, m_{i2}, \dots, m_{in}]^T \in R^n$. The typical and more convenient matching criterion may be used, based on the Euclidean distances between x and m_i. The minimum distance defines the "winner" m_c. Neighborhoods of the units of radii R=2, 1, and 0 are shown in figure (3). Note that each unit has a number of nearest neighbors depends on the shape of the map. Each unit in the rectangular grid has eight nearest neighbors, but only six in the hexagonal grid.

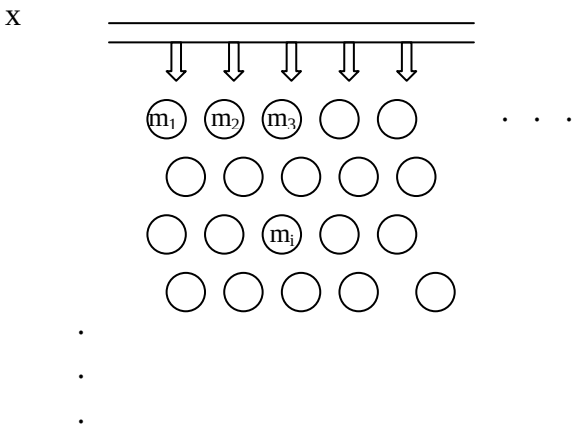

Figure (2): Cell arrangement of two-dimentional network

2- Training algorithm

Assume a two dimensional array of units or neurons which are arranged in a hexagonal lattice (figure (3)). To every node i of it we assign a time-variable weight vector $m_i(t) \in R^n$, t=0, 1, 2, … n.

The initial values $m_i(0)$ can be chosen as random vectors. Assume that an input pattern vector $x(t) \in R^n$ is broadcast to and concurrently compared with all the $m_i(t)$. The following two rules define a process in which the above mapping is formed by self-organization when a sufficient number of statistically distributed input vectors are applied.

1- Find unit c whose weight vector m(t) has the best match with x(t). a sample vector:
unit c is thereby said to respond to x(t). In the simplest case, Euclidean norms are used.

$$\|x(t) - m_c(t)\| = \min_i \{\|x(t) - m_i(t)\|\} \quad \ldots \quad (5)$$

2- Modify the weight vectors of unit c and its topological neighbors:
The topological neighborhood N_c, illustrated in figure (3), refers to the lattice of nodes and is usually a function of time. The adaptation law in this model reads;

$$m_i(t+1) = m_i(t) + \alpha(t)[x(t) - m_i(t)] \qquad \text{for } i \in N_c \quad \ldots \quad (6)$$
$$m_i(t+1) = m_i(t) \qquad \text{for } i \notin N_c$$

The learning rate α is a slowly decreasing function of time. The radius of the neighborhood around a cluster unit also decreases as the clustering process progresses.

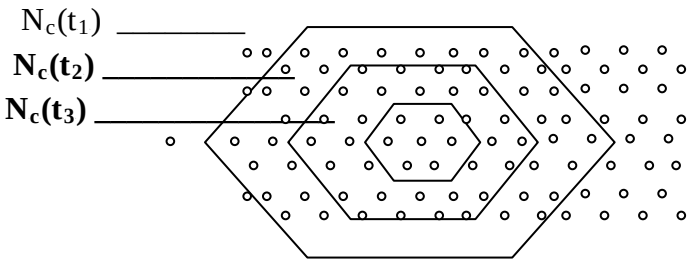

Figure (3): Neighborhoods for hexagonal grid.

3- Training data set

In this experiment, a simple arabic language vocabulary are used, which comprises nouns, verbs, and adverbs as shown in figure (4-a). Each word class has further categorical subdivisions, such as names of persons, animals, and inanimate.

The first difficulty is encountered when trying to find metric distance relations between symbolic items. Thus, during the learning process, the similarities of pairs of items, is presented in context. In linguistic representations context might mean a few adjacent words. Similarity between items would then be reflected through the similarity of the contexts. It is sufficient that the input data present together with a sufficient amount of context.

A sequence of randomly generated meaning three word arabic sentences was used as the input data to the self-organizing process. Meaningful sentence patterns had therefore first to be constructed on the basis of word categories as shown in figure (4-b). Each explicit sentence was then constructed by randomly substituting the numbers in a randomly selected sentence pattern from figure (4-b) by words with compatible numbering in figure (4-a). A total of 226 different three-word sentences are possible, a few of which are exemplified in figure (4-c).

(a) (b) (c)

Figure (4): Small arabic vocabulary used in this experiment. (a) List of used words. (b) Sentence patterns. (c) Some examples of generated sentences.

4- Simulation study

For the simulation, a rectangular lattice of 10 by 12 cells was used. The initial weight vectors of the cells were chosen randomly. Updating was based on (5) and (6). The learning rate was set to 0.7. After training for 1500 epochs, the responses of the neurons to presentation of the symbol parts alone were tested. In figure (5), the symbolic label is written to that site at which the symbol signal gives the maximum response. Words of the same type, i.e., nouns, verbs, and adverbs, are segregated into separate, large domains. Each of these domains is further organized according to the similarities on the semantic level.

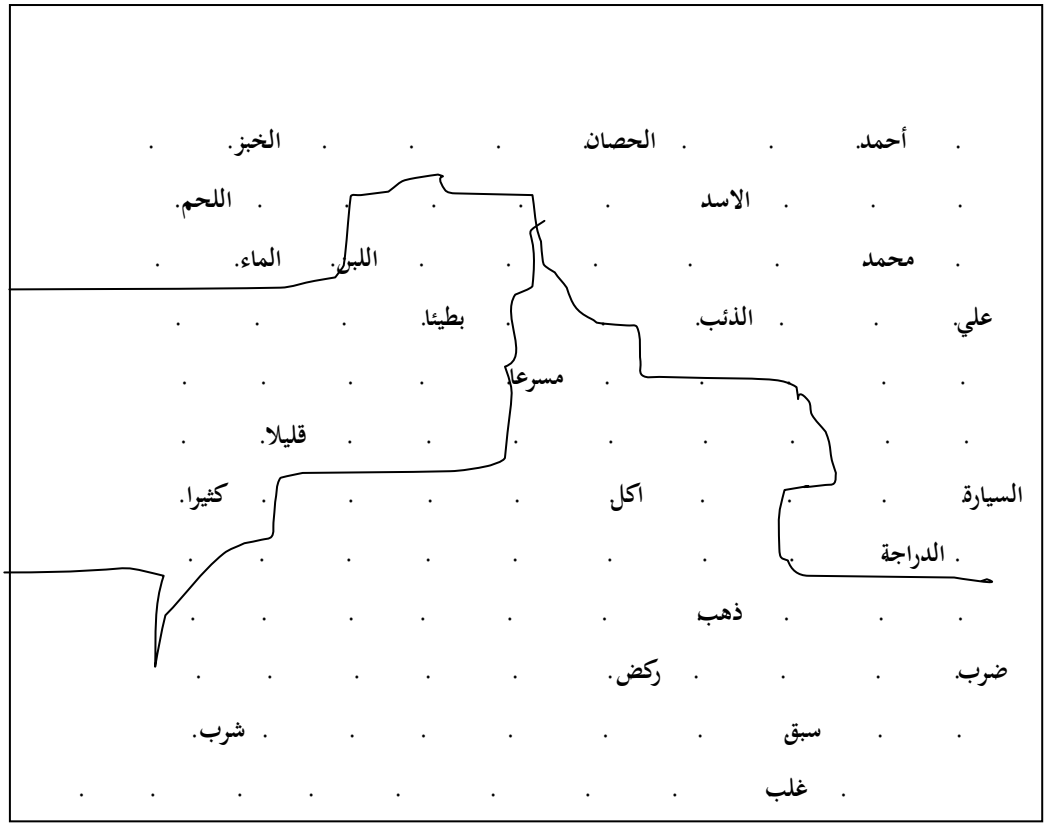

Figure (5): Semantic map obtained on a network of 10x12 cells.

Conclusions

One of the biggest challenges in natural language processing is overcoming the linguistic knowledge-acquisition bottleneck by providing the machine with the linguistic sophistication necessary to perform robust natural language processing.Recently, a number of exciting results have shown the feasibility of learning linguistic knowledge automatically.The effectiveness of this approach has been demonstrated across the spectrum of natural language processing tasks. In this paper, we present a self-organizing map for understanding arabic sentences. The system allows to extract the abstract information from multidimensional primary signals, and to represent it as a location in a two dimentional network. The performance of the system are shown on a small arabic vocabulary. A future development is the use of largest vocabulary consists of different types of sentences.

References

[1] Teuvo Kohonen, " The Self-Organizing Map," Proceedings of the IEEE, Vol. 78, No.9, September 1990.

[2] Y. Bar-Hillel, "Language and Information," Reading, Mass: Addison Wesley,1964.

[3] T. Winograd, " Understanding Natural Language," San Diego, Calif: Academic Press,1972.

[4] W. A. Woods, " Lunar Rocks In Natural English: Explorations in Natural Language Question Answering," In Linguistic Structures Processing, ed. Zampoli, New York: Elsevier, 1977.

[5] D. L. Waltz, " An English language Question-Answering System for a Large Relational Data base," Communications of the Association for Computing Machinery, 21 (7): 526-539, 1978.
[6] J. F. Allen, " Natural Language Understanding," Menlo Park, Calif: Benjamin, Cummings, 1987.
[7] E.Bill and R. J. Mooney, “ An overview of empirical natural languag processing,” AI Magazine, 1990.
[8] D. E. Rumelhart, and J. L. McClelland,eds., " Parallel Distributed Processing: Explorations in the Microstructure of congition, Volume 1 and 2, Cambridges Mass: MIT press,1986.
[9] H. Ritter and T. Kohenon, " Self-organizing semantic maps," Fiol. Cybern., Vol. 61, pp. 241-254, 1989.
[10] R. P. Lippman, " Review of research on neural net for speech," Neural computation 1(1), 1989.
[11] K. Plunkett and V. Marchman, " From rote learning to system building: Acquiring verb morphology in children and connectinist nets," Cognition 48(1): 21-69, 1993.
[12] R. G. Reilly and N. E. Sharkey, eds., " Connectionist approachs to natural language processing," Hillsdale, N. J. : Lawrence Erlbaum, 1992.
[13] T. J. Sejnowski and C. Rosenberg, " Parallel networks that learn to pronounce english text,' Complex systems 1: 195-198, 1987.
[14] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, " Learning internal representations by error propagation," in Parallel of Cognition, Vol. 1: Foundations, D. E. Rumelhart, J. L. McClelland and the PDP research group, Eds., Cambridge,Mass,: MIT Press,pp. 318-362, 1986.
[15] J. J. Hopfield, " Neural networks and physical systems with emergent collective computational abilities," Proc. Natl. Acad. Sci. USA, Vol. 79, pp. 2554-2558, 1982.

خارطة التنظيم الذاتي لفهم اللغة العربية

عبدالكريم يونس عبدالله

فزل عحطك حزائ ة

تكي بطكعكهل /جئ لع بطك ا شنب

طك لكخ ش

في السنوات الاخيرة، زاد الاهتمام باستخدام الشبكات العصبية في معالجة اللغات الطبيعية، وذلك لان الشبكات العصبية تستخدم تقنيات تعليم لاستخلاص المعرفة اوتوماتيكيا من اللغات الطبيعية بدلا من الحاجة الى تطوير انظمة يتم من خلالها الحصول على المعرفة المطلوبة يدويا.من المعروف ان الدماغ البشري يحتوي على مواقع مختلفة ومرتبة، بحيث كل مجموعة من الخلايا العصبية تعطي استجابة مثلى لمجموعة من اشارات الادخال المختلفة.في هذا البحث، تم استخدام خارطة التنظيم الذاتي في التمثيل اللغوي. تم ملاحظة انجازية النظام من خلال استخدام محموعة جمل عربية بسيطة، حيث تم بناء خارطة توزيع مثلى للكلمات الموجودة في الجمل العربية المستخدمة، والتي تعطي دقة عالية في فهم اللغة.