A Biologically Inspired Connectionist System for Natural Language Processing

João Luís Garcia Rosa
Mestrado em Sistemas de Computação - PUC-Campinas
Mestrado em Informática - UniSantos
Rodovia D. Pedro I, km. 136 – Caixa Postal 317 – CEP 13012-970 – Campinas, SP, Brasil
joaol@ii.puc-campinas.br – Fax: +55-19-3756-7195

Abstract

Nowadays artificial neural network models often lack many physiological properties of the nervous cell. Feedforward multilayer perceptron architectures, and even simple recurrent networks, still in vogue, are far from those encountered in cerebral cortex. Current learning algorithms are more oriented to computational performance than to biological credibility. The aim of this paper is to propose an artificial neural network system, called Bio-θR, including architecture and algorithm, to take care of a natural language processing problem, the thematic relationship, in a biologically inspired connectionist approach. Instead of feedforward or simple recurrent network, it is presented a bi-directional architecture. Instead of the well-known biologically implausible backpropagation algorithm, a neurophysiologically motivated one is employed to account for linguistic thematic role assignment in natural language sentences. In addition, several features concerning biological plausibility are also included.

1. Introduction

Successful connectionist natural language processing (NLP) systems often employ recurrent architectures instead of feedforward networks [4] [5] [17]. These systems with "reentrancy" are supposed to be more adequate to deal with the temporal extension of natural language sentences, and, at the same time, they seem to be more physiologically realistic [3]. The search for models based on neuroscience is about to be considered the next generation of artificial neural networks [12], since nowadays models are far from biology, mainly for mathematical simplicity reasons [16] [18]. Another item considered fundamental in a biologically based model is the representation adopted. It is required to be distributed, in a sense that one concept is represented along many units of the connectionist architecture [9] [14], while localist representations lack semantic distinctiveness [5] [17]. Natural language processing systems that use distributed representations have shown good performance [11] [21] [22] [19].

Here, a connectionist NLP system called Bio-θR is presented to account for thematic role relationships in natural language sentences. The architecture employed is a bi-directional (recurrent) artificial neural network. The processors are perceptron-like units and the connectionist learning algorithm uses a simple reinforcement rule, based only on available information of local synapses [13]. The words are presented sequentially to the network and represented by means of distributed semantic microfeature arrays [11] [24]. Twenty three-valued logic semantic microfeature units account for each noun and verb. The schema on table 1 displays the semantic features for verbs. Table 2 shows the microfeatures for nouns.

control of action	no control of action
direct process triggering	indirect triggering
direction to source	direction to goal
impacting process	no impacting process
change of state	no change of state
psychological state	no psychological state
objective	no objective
effective action	no effective action
high intensity of action	low intensity
interest on process	no interest on process

Table 1. The ten semantic microfeature dimensions for verbs. These features were chosen based on a thematic frame [22]

2. Thematic roles

Linguistic theory [6] refers to the roles words usually have in relation to the predicate (often the verb) as *thematic roles*, so that one can say that the verb *frighten*, in one possible reading of sentence (1), assigns the

Proceedings of the VII Brazilian Symposium on Neural Networks (SBRN'02)
0-7695-1709-9/02 $17.00 © 2002 IEEE

thematic structure (grid) [AGENT, EXPERIENCER], because *man* is deliberately responsible for the action of frightening (the "agent"), and *girl* is the "experiencer" affected by the action.

<table>
<tr><td colspan="4">human</td><td colspan="4">non-human</td></tr>
<tr><td colspan="4">soft</td><td colspan="4">hard</td></tr>
<tr><td colspan="2">small</td><td colspan="4">medium</td><td colspan="2">large</td></tr>
<tr><td colspan="2">1-D/compact</td><td colspan="4">2-D</td><td colspan="2">3-D</td></tr>
<tr><td colspan="4">pointed</td><td colspan="4">rounded</td></tr>
<tr><td colspan="4">fragile/breakable</td><td colspan="4">unbreakable</td></tr>
<tr><td>value</td><td colspan="2">furniture</td><td>food</td><td>toy</td><td colspan="2">tool/ utensil</td><td>animate</td></tr>
</table>

Table 2. The seven semantic microfeature dimensions for nouns, separated in rows. Only one value in each dimension is on for each unambiguous noun [22] (adapted from [11])

The man frightened the girl **(1)**

But linguistic theory also assumes that thematic structures may vary for a specific verb. So, in sentence (2), there is a different thematic grid ([CAUSE, EXPERIENCER]) assigned by the same verb *frighten*, since *ball* causes the frightening, but in an involuntary way.

The ball frightened the girl **(2)**

The Government and Binding linguistic theory [6] states that thematic roles are in the lexicon, so a specific verb assigns a single thematic grid. This is a "slot and filler" lexicalist view. For instance, the verb *kill* would assign an AGENT (*i*) and a PATIENT (*j*), no matter in which sentence it occurs, like in *Michael*$_i$ *killed Peter*$_j$ [22]. But there are verbs, which assign different thematic grids in different sentences, like *frighten* in (1) and (2). This is a problem for this symbolic linguistic theory.

In a componential perspective, it is possible to have a representation for verbs independently of the sentence in which they occur. Considering sentences (1) and (2) again, it seems that the nouns employed as subjects make the distinction between AGENT and CAUSE. In sentence (1), since the subject (*man*) is an animate noun, it can be the "agent" of an action. In sentence (2), the subject (*ball*) is inanimate, so it can not be agent of anything. In other words, thematic roles must be elements with semantic content [2].

In the system Bio-θR, the connectionist architecture processes the sentences in a componential way, allowing that this kind of semantic relationship be approached.

The representation of the verb is strongly based on a non-lexicalist representation; that is, the thematic role assignment compositionally depends on the whole sentence [21].

3. Biologically plausible connectionist models

According to O'Reilly [14], biologically plausible connectionist models should have, as main characteristics, distributed representation, inhibitory competition, bi-directional activation propagation, and error-driven task learning.

3.1. Distributed representation

Several are the advantages of the distributed representation concerning connectionism. According to Hinton and others [9], the connections between a set of units are capable of supporting a large number of different patterns, thus implying in a considerable reduction of the network size. And, regarding cognition, the strengths and weaknesses give rise to some powerful and unexpected emergent properties, like generalization. This way, systems that employ distributed representations are more neurophysiologically realistic. Distributed representations make possible to create new concepts without allocating new hardware. This means that, for NLP purposes, new words can be added to the lexicon of systems that use distributed representations, without modifying the architecture previously employed and trained. Besides this, in a neuroscience standpoint, distributed representations seem to be predominant in the cerebral cortex [15].

3.2. Inhibitory competition

The inhibitory competition feature present in the cerebral cortex is due mainly to inhibitory interneurons. During a lateral inhibition, a neuron excites an inhibitory interneuron that makes a feedback connection onto the first neuron, which is often called self-regulation [10]. As a matter of fact, 20% of the neurons in the cortex are inhibitory interneurons [14]. In Bio-θR, this happens at the output layer, where there is a competition between output units, in a kind of *winner-takes-all* strategy.

3.3. Bi-directional activation propagation

The bi-directionality of the architecture is necessary to simulate a biological electrical synapse, which may be bi-directional [10]. In Bio-θR, this is done by recurrence of the connectionist architecture. The hidden units receive stimuli from both the input and output layers.

3.4. Error-driven task learning

Proceedings of the VII Brazilian Symposium on Neural Networks (SBRN'02)
0-7695-1709-9/02 $17.00 © 2002 IEEE

The most used supervised connectionist algorithm backpropagation [23] requires the propagation of error signals in a manner inconsistent with known neurobiological properties [14] [1]. The error-driven task is important, but not in the way it happens in backpropagation [14]. So, in Bio-θR, a neurophysiologically connectionist learning algorithm is employed [13].

4. The system Bio-θR

The system Bio-θR consists of a bi-directional (recurrent) connectionist architecture, with three layers (A units in input layer, B units in hidden layer, and C units in output layer) and lateral inhibition occurring at the output level (figure 1). At input, words are presented in terms of their semantic microfeatures, one at a time, at its specific slot, until the whole sentence is completely entered. This way, besides semantics, included as part of the distributed representation employed, syntactic constraints are also considered. At output, thematic roles are highlighted as soon as they are assigned. For instance, when the subject of a sentence is presented, no thematic role shows up, because it is unknown which will be the main verb, the predicate that assigns such roles. When the verb appears, immediately the network displays the thematic role assigned to the subject presented previously. For the other input words, the correspondent thematic roles are displayed at the output, one at a time, for every input word.

5. The learning procedure

The learning procedure, also employed in Bio-Pred system [20], is inspired by the Recirculation [8] and GeneRec algorithms [13], and uses the two phases notion (minus and plus phases). First of all, the inputs x_i are presented to the input layer. In the minus phase, there is a propagation of these stimuli to the output through the hidden layer (bottom-up propagation). There is also a propagation of the previous actual output o_k back to the hidden layer (top-down propagation). Then, the hidden minus activation h_j^- is generated (sum of the bottom-up and top-down propagations – through the sigmoid activation function, represented by σ in equation 3). Finally, the current real output o_k is generated through the propagation of the hidden minus activation to the output layer (equation 4). The indexes i, j, and k refer to input, hidden, and output units, respectively.

$$h_j^- = \sigma\left(\sum_{i=0}^{A} w_{ij}.x_i + \sum_{k=1}^{C} w_{jk}.o_k\right) \quad \textbf{(3)}$$

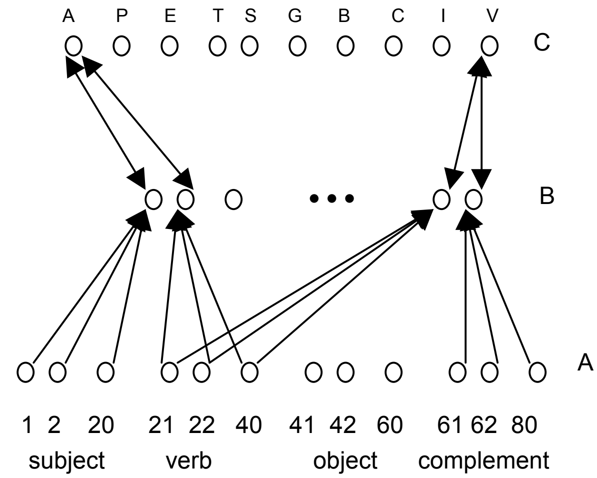

Figure 1. The three-layer bi-directional connectionist architecture of Bio-θR. The labels of the output units (in layer C) represent the ten thematic roles (AGENT, PATIENT, EXPERIENCER, THEME, SOURCE, GOAL, BENEFICIARY, CAUSE, INSTRUMENT, and VALUE). To the input layer A the words, represented by their distributed microfeature arrays, are entered sequentially at their specific slot according to their syntactic category (subject, verb, object, or complement)

$$o_k = \sigma\left(\sum_{j=1}^{B} w_{jk}.h_j^-\right) \quad \textbf{(4)}$$

In the plus phase, there is a propagation from the input x_i to the hidden layer (bottom-up). After this, there is the propagation of the desired output y_k to the hidden layer (top-down). Then the hidden plus activation h_j^+ is generated, summing these two propagations (equation 5). The synaptic weights w are updated in the way represented in equations 6 and 7. Notice the presence of the learning rate (η), considered an important variable during the experiments [20].

$$h_j^+ = \sigma\left(\sum_{i=0}^{A} w_{ij}.x_i + \sum_{k=1}^{C} w_{jk}.y_k\right) \quad \textbf{(5)}$$

$$\Delta w_{jk} = \eta.(y_k - o_k).h_j^- \quad \textbf{(6)}$$

$$\Delta w_{ij} = \eta.(h_j^+ - h_j^-).x_i \quad \textbf{(7)}$$

5.1. Different architectures

The experiments were accomplished by several different neural network architectures, in relation to the hidden layer. The input and output layers have always the same number of units, 80 and 10, respectively. The hidden layer size changed during the simulations. From 10 to 70, many sizes were experimented. The performance can be checked in figure 2 for three hidden layer sizes, for different learning rates. The 50-unit hidden layer was chosen, considering satisfactory training time and best learning performance for a learning rate $\eta = 0.25$.

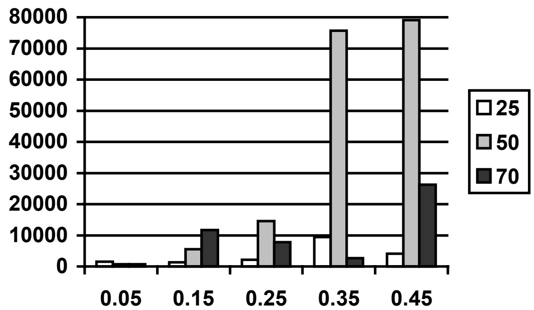

Figure 2. The diagram shows the influence of the learning rate η (represented in the horizontal X-axis) in three different neural networks architectures (hidden layer with 25, 50, and 70 units). The vertical Y-axis shows the number of training cycles needed to reach the output error of 10^{-2}

The average output error is the difference between "actual" output o_k and "desired" output y_k, and it is obtained from the *average squared error energy* formula (equation 8) [7].

$$Error = \frac{1}{2}\left(\sum_{k=1}^{C}(y_k - o_k)^2\right)/C \qquad (\mathbf{8})$$

In order to obtain the final output error, enough to make the system learn the correct sentence-thematic grid relations (which, empirically, was assumed to be 0.01), Bio-θR computes the average of 52 cycles, in order to compose a sample including all possible sentences.

5.2. A kind of competition

At the output layer there is a kind of lateral inhibition, since the unit in the output layer which is most active (the winner), makes the others to be inactive ("the winner takes all"). This is a kind of competitive learning procedure.

The output layer is composed of ten units, representing ten thematic roles (AGENT, PATIENT, EXPERIENCER, THEME, SOURCE, GOAL, BENEFICIARY, CAUSE, INSTRUMENT, and VALUE). The aim of Bio-θR is to learn the correct thematic role assignments for complete semantically sound sentences presented sequentially (one word at a time).

For each output in the training step, the most activated unit will be considered active and the other units will be inactive. Active and inactive units have values next to 1.0 and 0.0, respectively. This is done after the (back) propagation of the output to the hidden layer in the plus phase. That is, the effect of the competition will happen only in the minus phase of the next cycle.

5.3. The training set

In the bi-directional architecture of Bio-θR, the supervised learning procedure employs pairs of semantically sound sentences and the corresponding thematic grid. A sentence-thematic grid pair generator supplies the word and the respective thematic role for each sentence presented sequentially, one word at a time (see tables 3 and 4). For example, for sentence (9), the pairs *boy*-null, *delivered*-AGENT, *curtain*-THEME, and *woman*-GOAL are generated. Notice that for the first word (the subject), there is no thematic role prediction yet, because to the subject may be assigned any thematic role. The verb presentation is crucial for the decision in favor of AGENT.

The boy delivered the curtain to the woman (**9**)

In this version of Bio-θR, a small lexicon is employed, with only 13 verbs (including alternative readings, in case of thematic ambiguity) and 30 nouns. This is enough to validate the componential semantic relationships between the words of a sentence and their thematic roles. This way, it is possible to propose a more biologically realistic system than other systems that account for a similar linguistic problem, like HTRP [21] [22].

6. Distributed semantic microfeature representations

As already well known, distributed representation has several advantages concerning connectionism [9]. It is essential to notice that the verb microfeatures are selected in order to cover the semantic issues considered pertinent in a thematic framework. The microfeatures outside this thematic context are not meaningful [21] [22] [19]. They only make sense in a system like Bio-θR, where the specification of semantic relationships between the words in a sentence plays a leading role.

frame for verb love1	*thematic grid*
the human loves the object	[EXPERIENCER, THEME]
frame for verb love2	*thematic grid*
the human loves the human	[EXPERIENCER, THEME]
frame for verb frighten1	*thematic grid*
the object frightened the human	[CAUSE, EXPERIENCER]
frame for verb frighten2	*thematic grid*
the human frightened the human	[AGENT, EXPERIENCER]
frame for verb hit1	*thematic grid*
the object hit the thing	[CAUSE, PATIENT]
frame for verb hit2	*thematic grid*
the human hit the thing with the hitter	[AGENT, PATIENT, INSTRUMENT]
frame for verb buy1	*thematic grid*
the human bought the object from the human	[AGENT, THEME, SOURCE]
frame for verb buy2	*thematic grid*
the human bought the object by value	[AGENT, THEME, VALUE]
frame for verb give	*thematic grid*
the human gave the object to the human	[AGENT, THEME, BENEFICIARY]
frame for verb deliver	*thematic grid*
the human delivered the object to the human	[AGENT, THEME, GOAL]
frame for verb break1	*thematic grid*
the object broke the fragile object	[CAUSE, PATIENT]
frame for verb break2	*thematic grid*
the human broke the fragile object with the breaker	[AGENT, PATIENT, INSTRUMENT]
the human broke the fragile object	[AGENT, PATIENT]
frame for verb fear	*thematic grid*
the human fears the human	[EXPERIENCER, THEME]

Table 3. The sentence-thematic grid pair generator, showing only one sample sentence frame for each verb [22]. Each category will be filled by the words displayed on table 4

category	*words*
human	man, girl, boy, woman
object	ball, mechanical jack, doll, plate
thing	doll, chicken, sleeve, vase
fragile object	window, vase, plate
hitter	mechanical jack, hammer, stone
breaker	ball, hammer, vase, stone
value	ten, hundred, thousand

Table 4. Some words for each filler for sentence-thematic grid pair generator (table 3)

Besides the inclusion of lexically ambiguous nouns (like *bat*), Bio-θR allows thematically ambiguous verbs in its lexicon, as well. Thematic ambiguity means that a same verb can assign different thematic grids in different sentences, like the verb *frighten*. These words have indefinite dimensions, represented by the "?" symbol, in the microfeature array. As an example, the microfeatures of the verb *frighten* and its two alternative readings (*frighten1* and *frighten2*) are shown on table 5.

microfeature	*frighten*	*frighten1*	*frighten2*
control of action	?	no	yes
process triggering	?	indirect	direct
direction	goal	goal	goal
impacting process	yes	yes	yes
change of state	no	no	no
psychological state	yes	yes	yes
objective action	?	no	yes
effective action	no	no	no
intensity of action	low	low	low
interest on process	?	no	yes

Table 5. The thematically ambiguous verb *frighten*, with the default reading and two alternative readings. The "?" sign represents ambiguity

7. Conclusion

Bio-θR is a connectionist natural language processing system that account for the thematic role relationships between words in a sentence. Unlike most systems, Bio-θR adopts a biologically motivated model, including a bi-directional architecture and a physiologically plausible learning procedure.

Several experiments were made to reach an architecture that, in conjunction with an error-driven task learning algorithm that resembles GeneRec [13], is able to learn the thematic roles of sentences presented componentially one word at a time. It is important to notice that the word representation is distributed, in the sense that a set of units is used to represent one word. This is crucial in a system which aims to be neurophysiologically based. In addition, other biologically inspired features were included, like lateral inhibition.

8. References

[1] F. H. C. Crick, "The recent excitement about neural networks", *Nature* 337, 1989, pp. 129-132.

[2] D. Dowty, "On the Semantic Content of the Notion of 'Thematic Role'", in G. Chierchia, B. H. Partee, and R. Turner (Eds.), *Properties, Types, and Meaning II*. Dordrecht, Kluwer, 1989, pp. 69-129.

[3] G. M. Edelman, and G. Tononi, *A Universe of Consciousness – How Matter Becomes Imagination*, Basic Books, 2000.

[4] J. L. Elman, "Finding Structure in Time", *Cognitive Science* 14, 1990, pp. 179-211.

[5] J. L. Elman, "Learning and Development in Neural Networks: the Importance of Starting Small", *Cognition* 48, 1993, pp. 71-99.

[6] L. Haegeman, *Introduction to Government and Binding Theory*. Cambridge: Blackwell, 1991.

[7] S. Haykin, *Neural Networks - A Comprehensive Foundation*, 2nd edition. Prentice Hall, Upper Saddle River, New Jersey, 1999.

[8] G. E. Hinton, and J. L. McClelland, "Learning representations by recirculation", in D. Z. Anderson (Ed.), *Neural Information processing Systems, 1987*. New York: American Institute of Physics, 1988, pp. 358-366.

[9] G. E. Hinton, J. L. McClelland, and D. E. Rumelhart, "Distributed Representations", in D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing – Explorations in the Microstructure of Cognition, Volume 1 – Foundations*. A Bradford Book, MIT Press, 1986, pp. 77-109.

[10] E. R. Kandel, J. H. Schwartz, and T. M. Jessell, *Essentials of Neural Science and Behavior*. Appleton & Lange, Stamford, Connecticut, 1995.

[11] J. L. McClelland, and A. H. Kawamoto, "Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences", in J. L. McClelland and D. E. Rumelhart (Eds.), *Parallel Distributed Processing – Explorations in the Microstructure of Cognition – Volume 2: Psychological and Biological Models*. A Bradford Book, MIT Press, 1986, pp. 272-325.

[12] D. A. Medler, "A Brief History of Connectionism", *Neural Computing Surveys* 1, 1998, pp. 61-101.

[13] R. C. O'Reilly, "Biologically Plausible Error-driven Learning using Local Activation Differences: The Generalized Recirculation Algorithm", *Neural Computation*, 8:5, 1996, pp. 895-938.

[14] R. C. O'Reilly, "Six Principles for Biologically-Based Computational Models of Cortical Cognition", *Trends in Cognitive Science*, 2, 1998, pp. 455-462.

[15] R. C. O'Reilly, and Y. Munakata, *Computational Explorations in Cognitive Neuroscience – Understanding the Mind by Simulating the Brain*. A Bradford Book, The MIT Press, Cambridge, Massachusetts, 2000.

[16] A. F. Rocha, *Neural Nets - A Theory for Brains and Machines*, Berlin, Heidelberg: Springer-Verlag, 1992.

[17] D. L. T. Rohde, and D. C. Plaut, "Language Acquisition in the Absence of Explicit Negative Evidence: How Important is Starting Small?", *Cognition* 72, 1999, pp. 67-109.

[18] J. L. G. Rosa, "An Artificial Neural Network Model Based on Neuroscience: Looking Closely at the Brain", in V. Kůrková, N. C. Steele, R. Neruda, and M. Kárný (Eds.), *Artificial Neural Nets and Genetic Algorithms - Proceedings of the International Conference in Prague, Czech Republic, 2001 – ICANNGA-2001*. April 22-25, Springer-Verlag, pp. 138-141.

[19] J. L. G. Rosa, "HTRP II: Learning thematic relations from semantically sound sentences", in *Proceedings of the 2001 IEEE International Conference on Systems, Man, and Cybernetics - SMC2001*, October 7-10, 2001, Tucson, Arizona, United States of America, pp. 488-493.

[20] J. L. G. Rosa, "A Biologically Motivated Connectionist System for Predicting the Next Word in Natural Language Sentences", in *Proceedings of the 2002 IEEE International Conference on Systems, Man and Cybernetics – SMC'2002*, October 6-9, 2002, Hammamet, Tunisia, *accepted for publication*.

[21] J. L. G. Rosa, and E. Françozo, "Hybrid Thematic Role Processor: Symbolic Linguistic Relations Revised by Connectionist Learning", in *Proceedings of IJCAI'99 – Sixteenth International Joint Conference on Artificial Intelligence*, Stockholm, Sweden, July 31-August 6, Volume 2, Morgan Kaufmann, 1999, pp. 852-857.

[22] J. L. G. Rosa, and E. Françozo, "Linguistic Relations Encoding in a Symbolic-Connectionist Hybrid Natural Language Processor", in M. C. Monard and J. S. Sichman (Eds.), *Lecture Notes in Artificial Intelligence 1952, Advances in Artificial Intelligence, Proceedings of the International Joint Conference 7th. Ibero-American Conference on AI – 15th. Brazilian Symposium on AI – IBERAMIA-SBIA 2000*, Atibaia, Sao Paulo, Brazil, November 19-22, Springer-Verlag Berlin Heidelberg 2000, pp. 259-268.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning Internal Representations by Error Propagation", in D. E. Rumelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing – Explorations in the Microstructure of Cognition, Volume 1 - Foundations*. A Bradford Book, MIT Press, 1986, pp. 318-362.

[24] D. L. Waltz, and J. B. Pollack, "Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretations", *Cognitive Science* 9, 1985, pp. 51-74.