Proceedings of the IASTED International Conference ARTIFICIAL
INTELLIGENCE AND SOFT COMPUTING, held July 27-31, 1997,
Banff, Canada

# A THEMATIC CONNECTIONIST APPROACH TO PORTUGUESE LANGUAGE PROCESSING

JOÃO LUÍS GARCIA ROSA
Depto. Eletrônica e Computação - Instituto de Informática - Pontifícia Universidade Católica de Campinas – PUC-Campinas, &
Laboratório de Fonética Acústica e Psicolingüística Experimental - Departamento de Lingüística - IEL - Unicamp
*Address*: Rodovia D. Pedro I, km. 136. Caixa Postal 317 - CEP 13020-904 - Campinas - SP - Brasil
*Telephone*: +55-19-754-7162 - *Fax*: +55-19-754-7195 - *E-mail*: joaol@ii.puc-campinas.br

***ABSTRACT***: In the symbolic approach to Natural Language Processing (NLP), a system can only parse grammatically well constructed sentences. Within such a context, several linguistic phenomena, e.g. the thematic pattern relationships between the sentence constituents, can be accounted for (these pattern relationships are explained by a rule-based linguistic theory called thematic theory [1]). An alternative approach to NLP is a parallel distributed processing model, which has the benefits of learning and generalization. McClelland and Kawamoto [2] and Waltz and Pollack [3] proposed connectionist NLP models in which semantic microfeature representations of the words are used, in order to account for the relationship patterns between words in a sentence. In this paper, an adaptation of these connectionist systems is presented. Here, the microfeatures are based on thematic relationships between the words in the Portuguese sentences. However, will the output of such connectionist system account for the relationship patterns between the words in a sentence in a way which is comparable to that provided by the rule-based approach? The aim of this system is to investigate whether and how connectionist modeling can handle thematic relationships in sentences. An additional aim is to bring the Portuguese language to the scene, for comparative purposes.

***KEYWORDS***: Computational Linguistics; Neural Networks; Architectures for Natural Language Processing; Artificial Intelligence.

## INTRODUCTION

The linguistic approaches that consider thematic roles as elements with semantic content interest to the study of word meaning [4]. A linguistic theory called government and binding (GB) [1] embodies a thematic role system, which associates a predicate meaning with its arguments in semantic interpretation. One could imagine that the words which can fill each one of the slots for a given thematic grid have something in common in semantic terms. Now, one could try to capture such regularity (a) by describing each word in terms of its semantic features, and (b) by generalizing over all such descriptions for each thematic slot.

A decade ago, McClelland and Kawamoto [2] proposed a system, which is actually a treatment of relationship patterns, whose patterns are the words of a sentence. Their system handles relationship patterns between words in a sentence, in order to assign the correct case role to its constituents. In this paper, a connectionist architecture based on an adaptation of this model is proposed. The arrays of microfeatures are arranged on the basis of thematic relationships between the verb and the other words of a sentence, i.e. the model intends to map thematic roles to semantic features. This system is called CPPro, which stands for Connectionist Portuguese Language Processor.

CPPro was trained in affirmative sentences made up of up to four components: a subject, a verb, an object, and a complement. Within each component there are, on one hand, nouns, and on the other, determiners, adjectives, and so on; the latter are discarded after a previous processing. For instance, the sentence

A MENINA BONITA QUEBROU O VASO FRÁGIL
COM UMA PEDRA
(The pretty girl broke the fragile vase with a stone)

would end up as

MENINA-QUEBRAR-VASO-PEDRA
(girl-break-vase-stone)

in which MENINA is the agent (subject), QUEBRAR is the verb, VASO is the theme (object), and PEDRA is the instrument (complement).

Most published papers take only the English language into consideration. One contribution of CPPro is the transposition of procedures and ideas to the Portuguese language, that is, lexical ambiguities and syntactic constructions particular to this language. Take the case of subject/object ellipsis. Unlike the English language, Portuguese sentences may lack a superficial subject or object. For instance, in the sentence

QUEBROU A VIDRAÇA
(broke the window)

there is no subject, while in the sentence

O MENINO VIU
(the boy saw)

there is no object (in certain contexts).

Another goal of the model is to show how it can get the appropriate meaning in the case of an ambiguous word. CPPro does not aim to solve the problem of ambiguity, but contributes with ideas to make it less difficult. The connectionist approach had already proved to be efficient in treating a small set of Portuguese lexical constructions [5].

## THE IMPLEMENTATION

McClelland and Kawamoto's [2] and Waltz and Pollack's [3] systems deal with the word as a set of semantic microfeatures. According to them, every word is described by an array of bits in which each subset has an associated meaning, like "human-non human", "soft-hard", "male-female", and so on (see table 1). The aim of CPPro is to employ the idea of microfeature representation in order to build an architecture able to analyze and to learn the correct thematic relationship attributions of the words in a sentence.

The network is fed with a canonical representation of the word, that is, its set of semantic microfeatures. A verb has up to three networks: one for the agent, one for the theme, and one for the complement (which may be an instrument or a beneficiary). For instance, for the structure menina-quebrar-vaso-pedra (girl-break-vase-stone), the agent network is activated for a structure, which is the conjunction of microfeatures of MENINA with themselves, called input sentence structure (ISS). The network output format is called output thematic structure (OTS), which is the conjunction of microfeatures of the noun MENINA with microfeatures of the verb QUEBRAR. The process is repeated for the other networks.

Since the networks (one for each verb) were trained for many sentences, they are supposed to be able to verify whether a new sentence, not belonging to the learning set, displays the appropriate thematic role attributions. The algorithm used to implement these nets is the supervised backpropagation algorithm [6]. It proceeds as follows. First, it attributes random weights to the connections between the net nodes. When the user enters an ISS, the net output is compared with the intended output. The connections are modified in order to approximate the actual output to the intended one. This process is repeated tens of times, until the net converges, that is, until the net "learns" that structure.

## THE GOALS OF CPPRO

The main goal of CPPro is to provide a mechanism that deals with the role of semantic constraints on thematic role attribution. The model has to be able to learn to do this based on experience with sentences and their thematic representations, and has to be able to generalize to new sentences [2].

As it has already been said, model inputs are not raw sentences but semantic microfeature representations of the constituent structures of sentences. For nouns (table 1) and verbs (table 2), the features are grouped in several dimensions. Each dimension consists of a mutually exclusive value set. Each word is represented by a 20-bit array in which one, and only one value in each dimension is "on" for the word and all the other values are "off". Values that are "on" are represented in the feature arrays as "1"s. Values that are "off" are represented as "0"s (see tables 3 and 4).

| NOUNS - DIMENSIONS | FEATURE VALUES |
|---|---|
| HUMAN (2 bits) | human, non human |
| SOFTNESS (2 bits) | soft, hard |
| GENDER (2 bits) | male, female |
| VOLUME (3 bits) | small, medium, large |
| FORM (3 bits) | 1-D/compact, 2-D, 3-D |
| POINTINESS (2 bits) | pointed, rounded |
| BREAKABILITY (2 bits) | breakable, unbreakable |
| OBJECT TYPE (4 bits) | food, toy, tool/utensil, animate |

TABLE 1. FEATURE VALUES OF NOUNS (ADAPTATION FROM MCCLELLAND AND KAWAMOTO [2]).

| VERBS - DIMENSIONS | THEMATIC VALUES |
|---|---|
| AGENT (4 bits) | animate, inanimate, experiencer, none |
| PATIENT (4 bits) | animate, inanimate, theme, none |
| INSTRUMENT (2 bits) | has, does not have |
| TOUCH (4 bits) | agent, theme, both, none |
| BENEFACTION (2 bits) | yes, no |
| LOCALITY (4 bits) | source, goal, location, none |

TABLE 2. THEMATIC ROLES ASSIGNED BY VERBS (BASED ON HAEGEMAN'S [1] AND DOWTY'S [4] THEMATIC ROLE CLASSIFICATION).

The model can also handle the problem of ambiguity. For ambiguous words (for instance MACACO, which may be "monkey" or "mechanical jack") the input pattern is the average of each feature pattern of word readings. It means that in cases in which the two readings agree with the values of an input dimension, this dimension has the agreed value in the input representation. In cases in which the two readings disagree, the feature has the value 0.5 in the input representation. The goal is to verify whether the model can come up with the correct values for such unspecified slots or positions in the input array.

## LEARNING

In sentence presentation an OTS unit is computed, based on ISS pattern and on current values of net weights. The OTS can be quite different from the "intended" output, i.e. the values that it should have in the correct reading of the sentence. During learning, each OTS unit

| NOUN | TRANSLATION | HUMAN | SOFTNESS | GENDER | VOLUME | FORM | POINTINESS | BREAKABILITY | OBJECT TYPE |
|---|---|---|---|---|---|---|---|---|---|
| HOMEM | man | 10 | 10 | 10 | 001 | 001 | 01 | 01 | 0001 |
| MACACO | (ambiguous) | 01 | ?? | 10 | 100 | 001 | ?? | ?? | 00?? |
| MACACO-AN | monkey | 01 | 10 | 10 | 100 | 001 | 01 | 01 | 0001 |
| MACACO-ME | mechanical jack | 01 | 01 | 10 | 100 | 001 | 10 | 10 | 0010 |
| MARTELO | hammer | 01 | 01 | 10 | 100 | 100 | 10 | 10 | 0010 |
| MENINA | girl | 10 | 10 | 01 | 010 | 001 | 01 | 01 | 0001 |
| MENINO | boy | 10 | 10 | 10 | 010 | 001 | 01 | 01 | 0001 |
| PEDRA | stone | 01 | 01 | 01 | 100 | 001 | 10 | 01 | 0010 |
| VASO | vase | 01 | 01 | 10 | 100 | 100 | 01 | 10 | 0010 |
| VIDRAÇA | window | 01 | 01 | 01 | 010 | 010 | 10 | 10 | 0010 |

TABLE 3.SEMANTIC MICROFEATURES ARRAYS OF SOME NOUNS (THE “?” MEANS 0.5 - THERE IS AN AMBIGUITY). SEE TABLE 1.

| VERB | translation | agent | patient | instr. | touch | benef | locality |
|---|---|---|---|---|---|---|---|
| AMAR | love | 0010 | 0010 | 01 | 0001 | 01 | 0001 |
| BATER | hit | 1000 | 0100 | 10 | 0100 | 01 | 0001 |
| COMER | eat | 1000 | 0100 | 10 | 0100 | 01 | 0001 |
| DAR | give | 1000 | 0010 | 01 | 0001 | 10 | 0100 |
| MOVER | move | 1000 | 0001 | 01 | 0001 | 01 | 1000 |
| QUEBRAR | break | 1000 | 0100 | 10 | 0100 | 01 | 0001 |
| VER | see | 0010 | 0010 | 01 | 0001 | 01 | 0010 |

TABLE 4. THEMATIC MICROFEATURES ARRAYS OF SOME VERBS. SEE TABLE 2.

is compared to the correct reading, supplied as a “master input”. This master input should represent what a real language learner would construct from the context in which the sentence occurs. Learning may be described as the process of changing the connection weights to make the model output correspond, as close as possible, to the master input [2].

The learning phase initiates with a random weight set attribution. The network adjusts its weights every time it receives an input-output pair. Each pair requires two steps: a forward step and a backward step. The forward step consists in the presentation of an input sample to the network and the propagation of activation towards the output layer. During the backward step, the actual network output (after the forward step) is compared with the intended output and the error estimations are computed for the output units. The connection weights of the output units can be adjusted to reduce these errors. The model uses the error estimations of the output units to derive the error estimations of the hidden units. Finally, the errors are propagated back to the connections that were originated from the input units. Then a cycle is completed. The training of the backpropagation network usually requires many cycles.

## SIMULATION EXPERIMENTS

The networks used in CPPro have three layers: the input layer, to which the ISS is made available; the hidden layer, which allows the network to develop internal representations; and the output layer, from which the OTS representation is generated by the model.

The sentences presented to the net are generated by filling each category slot of sentence frames. Each frame specifies a verb, a thematic role set and a list of possible fillers of each thematic role. So, the sentence frame O HUMANO DEU O OBJETO PARA O HUMANO (The human gave the object to the human) is a generator for sentences in which HUMANO, the agent and the beneficiary, are replaced by one of the words in the human list, like MENINO (boy) or HOMEM (man), and OBJETO (the theme) is replaced by one of the words in the list of objects, like MACACO-ME (mechanical jack), since DEU (gave) asks for an agent (the one that gives), a theme (the thing that is given), and a beneficiary (the person who receives the thing). Then the sentence O MENINO DEU O MACACO-ME PARA O HOMEM (The boy gave the mechanical jack to the man) could be generated. Each verb has its generator. See table 5, for the verb COMER (eat). Note that in the last two frames there are no objects.

If all possible inputs and outputs are shown to a backpropagation network, the net will find a weight set that approximately maps the inputs to the outputs. For many Artificial Intelligence problems, however, it is impossible to provide all possible inputs. To solve this problem, the backpropagation network uses the generalization mechanism, i.e. the net will interpolate when inputs, which have never been received before, are supplied. In the case of this system, since words are described by microfeatures arrays, there are words with related meanings (like, for instance, HOMEM (man) and MENINO (boy)). These words are expected to have many microfeatures in common, so the distance between their microfeatures arrays is small.

## SYSTEM OPERATION

First, the system shows a menu through which the user enters an option to run CPPro. The first option is the learning step. The user is asked to say if this is the first time the net will be trained. If it is so, the system begins to train the network for some sentences given by the generators. The first verb to be trained is the verb AMAR

| SENTENCE FRAME | TRANSLATION | THEMATIC ROLES |
|---|---|---|
| O humano comeu o alimento com o utensílio. | The human ate the food with the utensil | animate agent - inanimate patient - instrument |
| O humano comeu o alimento. | The human ate the food | animate agent - inanimate patient |
| O humano comeu. | The human ate | animate agent |
| O animal comeu. | The animal ate | animate agent |

TABLE 5. THE GENERATOR FOR SOME SENTENCES WITH THE VERB COMER (EAT). SEE TABLE 6.

| CATEGORY | TRANSLATIO N | SOME FILLERS |
|---|---|---|
| HUMANO | human | homem (man), menina (girl) |
| ALIMENTO | food | batata (potato), frango (chicken), queijo (cheese) |
| UTENSÍLIO | utensil | colher (spoon), garfo (fork) |
| ANIMAL | animal | macaco-an (monkey) |

TABLE 6. SOME NOUN CATEGORIES FOR SOME FILLERS.

(love). Then, the other verbs (see table 4). For the verb AMAR, the first subject generated is HOMEM (man), and so on, until all sentences are given by the generator for AMAR. So a cycle is completed. Every cycle is completed for all nouns expected for the verb considered. When the network has already been trained, the system asks for the next sentence to be entered (this option allows the network to be trained incrementally).

The second option of the main menu is recognition. The system asks the user to enter a sentence. The thematic recognition output, after the inclusion of MENINO-DAR-PEDRA-HOMEM (boy-give-stone-man) as the subject-verb-object-complement structure, consults the disk files, in which the network weights for the verb DAR are stored, and gives the following output, for the subject MENINO: "This is a semantically acceptable word in this position".

Then the system gives the outputs for the other sentence nouns. In order to give these outputs, the system compares the average resultant array of thematic microfeatures given by the net with the intended array. When the dimension difference between both is small, the system considers these values equivalent.

The system performance analysis considered valid sentences, i.e. sentences that are supposed to be accepted by the system, and invalid sentences, which should be rejected. For about 6000 valid sentences, the system rejected only 5, and for about 3000 invalid sentences, the system accepted 26. It can be said that the performance of CPPro is very good, for the types of sentences for which it was trained.

## CONCLUSION

This system brings, as a contribution to NLP, a connectionist approach to the Portuguese language. It deals with thematic relationships between the words in a sentence. It also handles some particularities of the Portuguese language, like the absence of the subject or the object. It provided satisfactory results within the proposed plan. This research is expected to be further developed to incorporate richer vocabulary and structures, so that more complex Portuguese sentences can be implemented. In order to do this, it is necessary to upgrade the system, building larger networks, with larger microfeature arrays to distinguish better the new dimensions given to the words. By using more powerful machines, the number of input/output pairs for the network training could be raised. Thus, a system with a better overall performance could result.

***ACKNOWLEDGMENTS***: I would like to thank professors Eleonora Albano, Plínio de Almeida Barbosa, and Márcio Luiz de Andrade Netto, of Unicamp, and Lucia Alves Costa, of PUC-Campinas, and specially, professor Edson Françozo (IEL-Unicamp), for their extensive comments and inspiration on drafts of this paper.

## REFERENCES

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

[2] J. L. McClelland and A. H. Kawamoto, Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences. In J. L. McClelland, D. E. Rumelhart and the PDP Research Group, *Parallel Distributed Processing - Explorations in the Microstructure of Cognition, Volume 2: Psychological and Biological Models* (Cambridge, Massachusetts - London, England. A Bradford Book, The MIT Press, 1986).

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

[4] D. R. 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* (Dordrecht, Kluver, 1989).

[5] J. L. G. Rosa and M. L. A. Netto, Lógica e Conexionismo em Processamento de Linguagem Natural ("Logic and Connectionism in Natural Language Processing"), SUCESUSP'94 Proceedings - II Jornada USP-SUCESU-SP de Informática e Telecomunicações. June. São Paulo, SP, Brasil, 1994.

[6] R. P. Lippmann, An Introduction to Computing with Neural Nets, *IEEE ASSP Magazine*, April, 1987, 4-22.