# A Connectionist Thematic Grid Predictor for Pre-parsed Natural Language Sentences

João Luís Garcia Rosa

Computer Engineering Faculty - Ceatec
Pontifical Catholic University of Campinas - PUC-Campinas
Campinas, São Paulo, Brazil
joaoluis@puc-campinas.edu.br

**Abstract.** Inspired on psycholinguistics and neuroscience, a symbolic-connectionist hybrid system called $\theta$-PRED (*Thematic* PRED*ictor for natural language*) is proposed, designed to reveal the thematic grid assigned to a sentence. Through a symbolic module, which includes anaphor resolution and relative clause processing, a parsing of the input sentence is performed, generating logical formulae based on events and thematic roles for Portuguese language sentences. Previously, a morphological analysis is carried out. The parsing displays, for grammatical sentences, the existing readings and their thematic grids. In order to disambiguate among possible interpretations, there is a connectionist module, comprising, as input, a featural representation of the words (based on verb/noun *WordNet* classification and on classical semantic microfeature representation), and, as output, the thematic grid assigned to the sentence. $\theta$-PRED employs biologically inspired training algorithm and architecture, adopting a psycholinguistic view of thematic theory.

## 1 Introduction

The system $\theta$-PRED (*Thematic* PRED*ictor for natural language*) combines a symbolic approach, through a logical parser based on a Portuguese language grammar fragment, with a connectionist module, which accepts sentences coded in a semantic representation based on *WordNet* classification for verbs and nouns [1, 2] (see figure 1).

The sentences are parsed in the first module, which generates a logical representation based on events and thematic roles, disambiguating meanings through the production of as many formulae as possible readings.

The second module is responsible for the prediction of non-presented sentences in the first module, provided that the connectionist architecture is trained with representative patterns, allowing this way the generalization over the input sentence. The output of a succeeded propagation should be the correct thematic grid assigned to that sentence.

## 2 Thematic Roles

Thematic roles are the semantic relations between a predicate and its arguments [3, 4]. A predicate (usually the verb) assigns a thematic grid to a sentence, the

D. Liu et al. (Eds.): ISNN 2007, Part II, LNCS 4492, pp. 825–834, 2007.

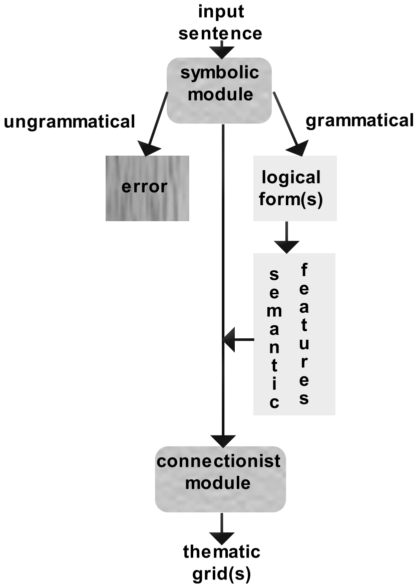

**Fig. 1.** The two modules of $\theta$-PRED system. The words are entered into the symbolic module for parsing. Ungrammatical sentences are discarded (*left*). Grammatical sentences have their logical forms generated (*right*). In addition, a semantic microfeature representation of the grammatical sentence is presented to the connectionist module. The system provides the thematic grids for recognized sentences.

structure containing every single thematic role of that sentence. For instance, the verb *judge*, in the sense *evaluate*, would assign an EXPERIENCER ($i$) and a THEME ($j$), no matter in which sentence it occurs, like in $[I]_i$ *cannot judge* [*some works of modern art*]$_j$. There are verbs, however, which assign different thematic grids to different sentences, for instance the verb *hit* in sentence (1), in the sense *cause to move by striking* and in sentence (2), in the sense *come into sudden contact with*. So, based on an episodic logic [5], a parser based on events ($e$) and thematic roles, can reveal the possible readings for sentences (1) and (2), displaying two different logical forms, one for each sentence.

*The man hit the ball.* (1)

Logical form: $\exists(x)$: man$(x) \wedge \exists(y)$ : ball$(y) \wedge \exists(e, \mathit{simple_past})$: hit$(e) \wedge$ agent$(e{,}x) \wedge$ patient$(e{,}y)$

*The car hit a tree.* (2)

Logical form: $\exists(x)$: car$(x) \wedge \exists(y)$ : tree$(y) \wedge \exists(e, \mathit{simple_past})$: hit$(e) \wedge$ cause$(e{,}x) \wedge$ patient$(e{,}y)$

To the sentences (1) and (2), although the same verb is employed, are assigned different thematic grids. In one possible reading of sentence (1), the thematic grid assigned is [AGENT, PATIENT] and in sentence (2), [CAUSE, PATIENT]. The reason is that *the man*, in the intended reading of sentence (1), is supposed to have the *control of action*, that is, the intention of hitting. The same does not occur in sentence (2). *The car* is not willing to hit anything. Verbs that assign different thematic grids to different sentences are called here *thematically ambiguous*.

The thematic role notion employed here is what some researchers call *abstract* thematic roles [6].

## 3 The Symbolic Module

Departing from an episodic logic, based on events, it is proposed here a Portuguese language Montagovian grammar fragment [7] considering different classes of adverbs. According to Ilari *et al.* [8], Portuguese adverbs for the spoken language may be classified into many types, including predicative and non-predicative. Predicative adverbs modify the meaning of the verb or adjective, implying a higher order predication, because the adverb predicates a property of the quality or action attributed to the subject. When the adverb does not alter the meaning of the verb or adjective, it is called non-predicative.

### 3.1 The Grammar

$\theta$-PRED's lexicon contains several adverbs, according to Ilari *et al.* [8], including qualitative and intensifier predicative adverbs, sentential (modal and aspectual), and non-predicative (negation). Since the analysis presents logical forms based on events, the adverb that comes with the verb, the noun, or the adjective is called *adjunct*.

Predicative adverbs correspond to second order predication, and the parser is implemented in a first order predicate logic, based on events and thematic roles, that do not support higher order predication. Only non-predicative adverbs should be treated as first order arguments or logical operators.

The sentences are formed according to a phrasal grammar, considering adverbs as adjuncts, prepositional phrases, adjectives, relative clauses, anaphora resolution, and phrases connected by the conjunction *and*.

The grammar includes sentence conjunction, allowing anaphora employment (personal pronouns) in the second sentence of the conjunction. It includes also prepositional phrases, through the so-called *with-NPs*, that is, a noun phrase beginning with the word *with*. This allows the analyzer process the ambiguous sentence (3). In this case, two logical forms are obtained: the first, where *binoculars* are the instrument of the verb *see* and the second, where *the girl* owns them.

$$\textit{The man saw the girl with the binoculars.} \tag{3}$$

$$\text{Logical form 1: } \exists(x)\text{: } (\text{man}(x) \wedge \exists(z)\text{: binoculars}(z) \wedge \exists(y)\text{: girl}(y) \wedge \\ \exists(e, \textit{simple\_past})\text{: see}(e) \wedge \text{experiencer}(e{,}x) \wedge \text{theme}(e{,}y) \wedge \text{instrument}(e{,}z)$$

$$\text{Logical form 2: } \exists(x)\text{: } (\text{man}(x) \wedge \exists(z)\text{: binoculars}(z) \wedge \exists(y)\text{: girl}(y) \wedge \\ \exists(e, \textit{simple\_past})\text{: see}(e) \wedge \text{experiencer}(e{,}x) \wedge \text{theme}(e{,}y) \wedge \text{own}(y{,}z)$$

If different sentences contain the same thematically ambiguous verb, like *hit* in sentences (1) and (2), they can be assigned different thematic grids. But, for one ambiguous sentence, like sentence (3), different thematic grids are assigned also, one for each possible interpretation.

Besides ordinary verbs and thematically ambiguous verbs, in $\theta$-PRED lexicon there are two-sense verbs with only one thematic grid (for instance, *love*: according to *WordNet*, there are four senses for verb *love* (here two of them are employed: *enjoy* (sentence 4), and *be in love with* (sentence 5); for both the thematic grid is [EXPERIENCER, THEME])).

$$\textit{I love western movies.} \tag{4}$$

$$\textit{Mary loves her husband.} \tag{5}$$

### 3.2 Computational Implementation of the Symbolic Parser

The computational implementation of a context free grammar fragment with adverbs, based on events and thematic roles, is performed through the logical programming language Prolog, where language statements are transposition of first order predicate logical formulae. A semantic analyzer supplies all possible logical forms of Portuguese declarative sentences, analyzing the determiner employed and giving the adequate quantifier.

The first version of the parser includes also a morphological analysis, which classifies each regular verb, in tense, number, and person, and each noun, adjective, etc., in gender and number[1]. Some irregular verbs are included, like *ser/estar* (to be).

A small lexicon is implemented, where only singular forms of nouns and infinitive forms of verbs are considered (the morphological analysis would discover the number, in case of nouns, and the tense, number, and person, in case of verbs). This analysis is based on a phrasal grammar [9]. If the sentence is ungrammatical, the parser rejects it.

## 4 The Biologically Plausible Connectionist Module

In this section, it is presented the second module of $\theta$-PRED system: the way the words are represented, the connectionist architecture of the system, and the employment of a biologically plausible supervised learning algorithm with simulation experiments.

### 4.1 Word Representation

In order to classify verbs and nouns, $\theta$-PRED employs a representation based on classical semantic microfeature distributed representation [10] and on *WordNet*[2].

[1] In Portuguese, verbs have, besides tense and number, person too, that is, there are different forms for verbs with different persons, no matter which is the tense. Portuguese adjectives agree with the noun they describe, so they feature gender and number. The morphological analysis gives the correct form of the word.

[2] *WordNet* version 2.1: http://wordnet.princeton.edu/obtain.

*WordNet* is a lexical data base (an ontology based on semantics [11]) of the English language [1, 2] which contains around 120,000 synonym sets (synsets) of nouns, verbs, adjectives, and adverbs, each one representing a lexicalized concept. The verbs chosen from *WordNet* represent all kinds of semantic relationships the system intends to treat. Twenty five dimensions with two binary units each account for each verb (see table 1) and thirty dimensions for each noun (see table 2).

**Table 1.** The semantic microfeature dimensions for verbs according to *WordNet* and to a thematic frame [10]

| | | | |
|---|---|---|---|
| *body* | *change* | *cognition* | *communication* |
| *competition* | *consumption* | *contact* | *creation* |
| *emotion* | *motion* | *perception* | *possession* |
| *social* | *stative* | *weather* | *control of action* |
| *process triggering* | *direction of action* | *impacting process* | *change of state* |
| *psychological state* | *objective action* | *effective action* | *intensity of action* |
| *interest on process* | | | |

**Table 2.** The semantic microfeature dimensions for nouns, based mainly on *WordNet*

| | | | | |
|---|---|---|---|---|
| *action* | *life* | *element* | *property* | *corporeal* |
| *social* | *nature* | *miscellaneous* | *size* | *consistency* |
| *form* | *fragility* | *instrument* | *adulthood* | *gender* |
| *body* | *change* | *cognition* | *communication* | *competition* |
| *consumption* | *contact* | *creation* | *emotion* | *motion* |
| *perception* | *possession* | *social* | *stative* | *weather* |

Since the aim of the presented system is to deal with thematic relationships between words in a sentence, the microfeatures chosen for verbs attempt to contemplate the semantic issues considered relevant in a thematic frame. The microfeatures outside this context are meaningless [12].

### 4.2 The Connectionist Architecture

$\theta$-Pred employs a bi-directional three-layer connectionist architecture with a hundred input units, fourteen hidden units, and seven output units, one for each of the thematic roles: AGENT, PATIENT, EXPERIENCER, THEME, LOCATION, CAUSE, and VALUE (see figure 2). In this case, according to Sun [13], the architecture can be classified as *single-module* employing distributed representation. For each sentence, the words are presented sequentially to their specific slot (verb or noun) in input layer.

The data used in experiments are realistic in the way they reflect situations found “in the wild.” The method used for generating sentences for training and test (i.e. by filling out the slots of sentence frames) creates a compelling set of

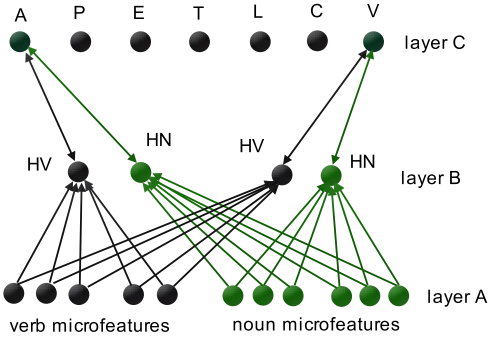

**Fig. 2.** The connectionist architecture of $\theta$-PRED. The sentence is presented to the input layer A and its thematic grid is revealed at output layer C. Notice that there are different slots for verbs and nouns. In the hidden layer B there are the conjunction of verb inputs in *HV* and the conjunction of noun inputs in *HN*. These two units are connected to one unit, regarding a specific thematic role, in the output layer C. Notice the bi-directional links between hidden (B) and output (C) layers, while there are unidirectional links from input (A) to hidden (B) layer. Legend for the output layer C (thematic roles): A = AGENT, P = PATIENT, E = EXPERIENCER, T = THEME, L = LOCATION, C = CAUSE, and V = VALUE.

training or test instances, because the chosen frames are representative for the kinds of sentences $\theta$-PRED intends to deal with.

### 4.3 Biologically Plausible Supervised Learning

In each sentence presentation an output is computed, based on an input pattern and on current values of net weights. The actual output can be quite different from the "expected" output, i.e. the values that it should have in the correct reading of the sentence, that is, the correct thematic grid assigned to the input sentence. During training, each output 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 system output correspond, as close as possible, to the master input.

The learning algorithm used in $\theta$-PRED is inspired by Recirculation [14] and GeneRec algorithms [15] . This algorithm is considered biologically more plausible since it supports bidirectional propagation, among other items [16].

The algorithm consists of two phases: *minus* and *plus* (figure 3). In the *minus* phase, the semantic microfeature representation of the first word of a sentence is presented to the input layer $A$. Then, there is a propagation of these stimuli to the output through the hidden layer $B$ (*bottom-up propagation*). There is also a propagation of the previous actual output, which is initially empty, from output layer $C$ back to the hidden layer $B$ (*top-down propagation*). Then, a hidden minus activation is generated (sum of the bottom-up and top-down propagations),

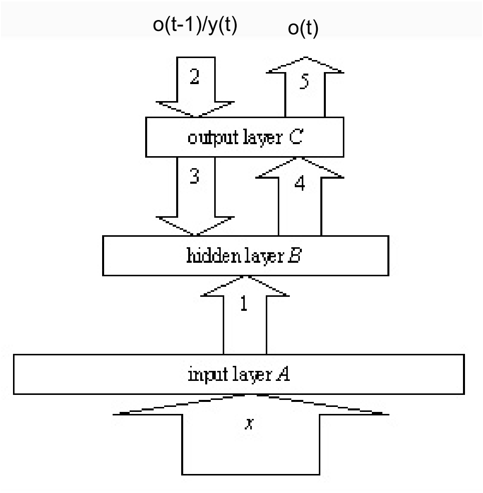

**Fig. 3.** The two phases of GeneRec algorithm. In the *minus* phase, when input $x$ is presented to input layer $A$, there is propagation of these stimuli to the hidden layer $B$ (1). Then, a hidden minus signal is generated based on input and previous output stimuli $o(t-1)$ (2 and 3). Then, these hidden signals propagate to the output layer $C$ (4), and an actual output $o(t)$ is obtained (5). In the *plus* phase, input $x$ is presented to layer $A$ again; there is propagation to hidden layer (1). After this, expected output $y$ (2) is presented to the output layer and propagated back to the hidden layer $B$ (3), and a hidden plus signal is generated, based on input and on expected output. Recall that the architecture is bi-directional, so it is possible for the stimuli to propagate either forwardly or backwardly.

through the sigmoid logistic activation function $\sigma$ (equation 6). Finally, the current actual output is generated through the propagation of the hidden minus activation to the output layer (equation 7) [17].

$$h_j^- = \sigma(\Sigma_{i=0}^{A} w_{ij}.x_i + \Sigma_{k=1}^{C} w_{jk}.o_k(t-1)), \tag{6}$$

$$o_k(t) = \sigma(\Sigma_{j=1}^{B} w_{jk}.h_j^-). \tag{7}$$

In the *plus* phase, there is a propagation from input layer $A$ to the hidden layer $B$ (bottom-up). After this, there is the propagation of the expected output to the hidden layer (top-down). Then a hidden plus activation is generated, summing these two propagations (equation 8). For the other words, presented one at a time, the same procedure (*minus* phase first, then *plus* phase) is repeated. Recall that since the architecture is bi-directional, it is possible for the stimuli to propagate either forwardly or backwardly [17].

$$h_j^+ = \sigma(\Sigma_{i=0}^{A} w_{ij}.x_i + \Sigma_{k=1}^{C} w_{jk}.y_k). \tag{8}$$

In order to make learning possible the synaptic weights are updated (equations 9 and 10), considering only the local information made available by the synapse. The learning rate $\eta$ used in the algorithm is considered an important variable during the experiments [18].

$$\Delta w_{jk} = \eta.(y_k - o_k(t)).h_j^-, \tag{9}$$

$$\Delta w_{ij} = \eta.(h_j^+ - h_j^-).x_i. \tag{10}$$

### 4.4 Simulation Experiments

The sentences presented to the net are generated by filling each category slot of sentence frames. Each frame specifies a verb, a noun set and a list of possible fillers of each noun. So, the sentence frame *the human buys the thing* is a generator for sentences in which the subject *human* is replaced by one of the words in the human list, like *man*, and *thing* is replaced by one of the words in the list of things, like *car*, since *buy* assigns the following thematic roles: an AGENT (the one who buys) and a THEME (the thing that is bought). Then the sentence *the man bought the car* could be generated. And the output for this sentence would be the assigned thematic grid [AGENT, THEME].

If all possible inputs and outputs are shown to a connectionist network employing a supervised training procedure, 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 training algorithm uses the generalization mechanism, i.e. the network 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, *man* and *boy*). These words are expected to contain many microfeatures in common, so the distance between their microfeatures arrays is small, favoring generalization.

The system is trained to learn the correct thematic grids assigned to input sentences. The training set was chosen in order to contain representative verbs and nouns of each thematic category present in $\theta$-PRED. For the system evaluation, test sentences are generated automatically. These sentences are different from the sentences generated by the training sentence generator, although their thematic frames are basically the same (the difference relies on the choice of the words involved). In this case, only the default readings for thematically ambiguous verbs are generated, simulating a user entering sentences to be analyzed. The user does not need to know which thematic reading is expected for the verb; $\theta$-PRED will decide, based on sentence context, which will be the correct reading and, consequently, arrive at the expected thematic grid for that sentence.

In relation to accuracy, the connectionist module of the system presents recall and precision rates of 94%[3], since only seven words revealed inadequate thematic roles in 120 words belonging to a limited, but sufficient, set of test sentences.

[3] According to Jurafsky and Martin [19], *recall* is defined by the number of correct answers given by the system divided by the total number of possible correct answers in the text, while *precision* is the number of correct answers given by system divided by the number of answers given by the system. Since $\theta$-PRED is fed only by correct sound sentences, in this case *recall* and *precision* coincide.

## 5 Concluding Remarks

The purpose of this paper is to present a symbolic-connectionist hybrid system consisting of two modules: a symbolic parser based on events, employing a grammar which takes into consideration classes of adverbs, according to Ilari *et al.* [8], in addition to transitive and intransitive verbs, and a biologically plausible connectionist thematic grid predictor. Since most of adverbs modify the meaning of a verb or an adjective, they experiment a kind of second order predication. For this reason, a parser based on events is chosen.

In connectionist Natural Language Processing (NLP) systems, the words belonging to a sentence must be represented in such a way as to keep the meaning of the words and, at the same time, to be useful for the network to develop significant internal representations. The representation of semantic features adopted in this system would also easily allow for new words to be entered in order to increase its lexicon, provided that their semantic microfeature arrays are supplied.

$\theta$-Pred presents as a novelty a more biologically plausible architecture and training procedure based on neuroscience [15], which comprises a bi-directional connectionist architecture, to account for chemical and electrical synapses that occur in the cerebral cortex, and a training procedure that makes use of this architecture.

## References

1. Fellbaum, C.: English Verbs as a Semantic Net. Intl. J. of Lexicography **3** (1990) 278-301
2. Miller, G.A.: Nouns in Wordnet. In Fellbaum, C., ed.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, Massachusetts (1998)
3. Chomsky, N.: Lectures on Government and Binding: the Pisa Lectures. Holland: Foris Pub. (1981)
4. Chomsky, N.: Knowledge of Language: its Nature, Origin, and Use. New York: Praeger Pub. (1986)
5. Schubert, L.K., Hwang, C.H.: Episodic Logic Meets Little Red Riding Hood - a Comprehensive Natural Representation for Language Understanding. In Iwanska, L.M., Shapiro, S.C., eds.: Natural Language Processing and Knowledge Representation - Language for Knowledge and Knowledge for Language. AAAI Press / The MIT Press (2000) 111-174
6. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational Linguistics **28** (2002) 245-288
7. Dowty, D.R., Wall, R.E., Peters, S.: Introduction to Montague Semantics. Reidel Pub. Co. (1981)
8. Ilari, R., de Castilho, A.T., de Castilho, C.M., Franchi, C., de Oliveira, M.A., Elias, M.S., de Moura Neves, M.H., Possenti, S.: Considerações sobre a posição dos advérbios. In: Gramática do Português Falado - Volume I: A Ordem. Editora da Unicamp/Fapesp, Campinas, SP, Brazil (1990) 63-141
9. Pereira, F.C.N., Warren, D.H.D.: Definite Clause Grammars for Language Analysis - a Survey of the Formalism and a Comparison with Augmented Transition Networks. Artificial Intelligence **13** (1980) 231-278

10. McClelland, J.L., Kawamoto, A.H.: Mechanisms of Sentence Processing: Assigning Roles to Constituents of Sentences. In McClelland, J.L., Rumelhart, D.E., eds.: Parallel Distributed Processing, Volume 2 - Psychological and Biological Models. A Bradford Book, MIT Press (1986)
11. O'Hara, T.P.: Empirical Acquisition of Conceptual Distinctions via Dictionary Definitions. PhD thesis, NMSU CS (2004)
12. Rosa, J.L.G., da Silva, A.B.: Thematic Role Assignment through a Biologically Plausible Symbolic-connectionist Hybrid System. In: Proceedings of the Intl. Joint Conf. on Neural Networks - IJCNN 2004, Budapest, Hungary (2004) 1457-1462
13. Sun, R.: Hybrid Connectionist/Symbolic Systems. In Arbib, M.A., ed.: The Handbook of Brain Theory and Neural Networks. 2 edn. A Bradford Book, MIT Press (2003) 543-547
14. Hinton, G.E., McClelland, J.L.: Learning Representations by Recirculation. In Anderson, D.Z., ed.: Neural Information Processing Systems. American Institute of Physics, New York (1988) 358-366
15. O'Reilly, R.C.: Biologically Plausible Error-driven Learning Using Local Activation Differences: the Generalized Recirculation Algorithm. Neural Computation **8** (1996) 895-938
16. O'Reilly, R.C.: Six Principles for Biologically-based Computational Models of Cortical Cognition. Trends in Cognitive Science **2** (1998) 455-462
17. Rosa, J.L.G.: A Biologically Inspired Connectionist System for Natural Language Processing. In: Proceedings of the 2002 VII Brazilian Symposium on Neural Networks - SBRN 2002, Recife, Brazil, IEEE Computer Society Press (2002) 243-248
18. Haykin, S.: Neural Networks - a Comprehensive Foundation. 2 edn. Prentice Hall (1999)
19. Jurafsky, D., Martin, J.H.: Speech and Language Processing - an Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall (2000)