

PROCESSAMENTO DE LINGUAGEM NATURAL: ATRIBUIÇÃO DE CASO AOS CONSTITUINTES DAS SENTENÇAS

NATURAL LANGUAGE PROCESSING: CASE ATTRIBUTION TO SENTENCE CONSTITUENTS

Adriano José Imbiriba CARNEIRO*
Andrea Barbosa MENANDRO**
José Eduardo CAMPOS**
Prof. João Luis Garcia ROSA***

ABSTRACT

This article describes the work that has been developed by the Research Group on Artificial Intelligence since 1991 when the group began its studies in the area of natural language processing with the aim of developing a system based on connectionist models for case attribution to sentence constituents of the Portuguese language. The developed system makes a semantic analysis of a stating phrase, that is, it indicates who the agent is, the patient of the action, the instrument made use of to accomplish the action and the modifier (if so). In this article, the functions and the features of the system are dealt with so as to understand how it works.

KEY-WORDS: Natural Language Processing, Connectionism, Neural Network, Semantic Analysis

RESUMO

Este artigo descreve o trabalho desenvolvido pelo Grupo de Pesquisa em Inteligência Artificial desde sua criação em 1991, quando iniciou estudos na área de processamento de linguagem natural, no sentido de desenvolver um sistema baseado em modelos conexionistas para atribuição de caso aos constituintes de sentenças da língua portuguesa. O sistema desenvolvido faz a análise semântica de uma frase declarativa, isto é, indica quem é o agente, o paciente da ação, o instrumento utilizado para realizar a ação e o modificador (se for o caso). Neste artigo, as funções e as características do sistema são abordadas de maneira a se entender o seu funcionamento.

PALAVRAS-CHAVE: Processamento de linguagem natural, Conexionismo, Rede neural, Análise semântica.

INTRODUÇÃO

A especificação de quem faz o quê para quem (atribuição de caso temático aos constituintes da sentença) não resolve todos os problemas, mas reflete um aspecto importante do processo de compreensão da sentença. Este processo envolve a consideração si-

multânea de diferentes fontes de informação, tais como a ordem da palavra, as restrições semânticas e o contexto. No sistema descrito, usou-se uma abordagem baseada em REDE NEURAL, pois ela provê um mecanismo que leva em consideração a ordem da palavra e as restrições semânticas.

(*) Aluno concluinte do Curso de Análise de Sistemas - PUCCAMP

(**) Graduados em Análise de Sistemas pela PUCCAMP

(***) Professor do Curso de Engenharia da Computação - PUCCAMP; Professor do Curso de Análise de Sistemas - PUCCAMP; Professor do Curso de Especialização em Análise de Sistemas - PUCCAMP e Coordenador do Departamento de Eletrônica e Computação - II - PUCCAMP

FONTES DE INFORMAÇÃO

a. ORDEM DA PALAVRA

Por ordem da palavra entende-se a posição ocupada pela palavra analisada na frase.

(1) - O garoto quebrou o vaso.

(2) - O vaso quebrou a vidraça.

Podemos notar claramente que na frase (1) os casos dos constituintes da sentença seriam:

agente => garoto

verbo => quebrou

paciente => vaso

Entretanto, na frase (2) há uma mudança no papel da palavra vaso:

instrumento => vaso

verbo => quebrou

paciente => vidraça

b. RESTRIÇÕES SEMÂNTICAS

Por restrições semânticas entendem-se os casos possíveis de serem assumidos por uma palavra em uma frase.

(3) - O garoto quebrou a vidraça com a pedra.

(4) - O garoto quebrou a vidraça com a cortina.

Na frase (3) os papéis de caso dos constituintes da sentença seriam:

agente => garoto

verbo => quebrou

paciente => vidraça

instrumento => pedra

já na frase (4) os papéis de caso dos constituintes seriam:

agente => garoto

verbo => quebrou

paciente => vidraça

modificador => cortina

CONTEXTO

O sistema leva em consideração a sentença em que a palavra é empregada, uma vez que há casos em que somente a inserção desta em uma dada situação poderá levar à completa e perfeita compreensão da idéia descrita na frase.

FUNÇÕES DO SISTEMA

O sistema é capaz de aprender a fazer a atribuição de caso (baseado em experiência com sentenças e suas representações de caso), generalizar o que aprendeu para novas sentenças, selecionar contextualmente leituras apropriadas de palavras ambíguas e selecionar apropriadamente o caso dos constituintes de acordo com suas características semânticas.

DESCRIÇÃO DO SISTEMA

O sistema consiste de uma rede neural com três camadas (onde a primeira camada representa a sentença a ser analisada e a terceira sua estrutura de caso) e um dicionário com palavras usadas pelo sistema (e suas respectivas microcaracterísticas semânticas). O algoritmo conexionista empregado foi o "backpropagation".

Cada palavra é representada como um conjunto de microcaracterísticas semânticas (20 no total), que neste caso foram escolhidas para capturar o que se considerou dimensões importantes nos significados das palavras, assumindo valores de acordo com a seguinte convenção [MCC86]:

0 = microcaracterística ausente;

0,5 = pode ser que sim/pode ser que não (o sistema decide);

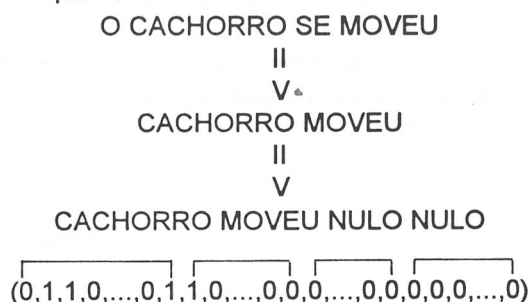
1 = microcaracterística presente.

Por exemplo, na definição das microcaracterísticas do conceito MENINA tem-se:

1	0	1	0	0	1	0	1	0	0	0	1	0	1	0	1	0	0	0	1
h	n	m	d	m	f	p	m	g	c	2	3	p	r	f	i	a	b	f	s
u	ã	o	u	a	e	e	é	r	o	D	D	o	e	r	n	l	r	e	e
m	o	l	r	s	m	q	d	a	m			n	d	á	q	i	r	r	
a	e	o	c	i	u	i	n	p				t	o	g	u	m	n	r	
n	h		u	n	e	o	d	a				i	n	i	e	e	q	a	v
o	u		l	i	n		e	c				a	d	i	b	n	u	m	i
m			i	n	o		t					g	o	r	t	e	e	v	
a			n	o								u			á	o	d	n	o
n			o									d			v	o	t		
o												o			e		a		

O conjunto que representa a sentença a ser analisada é formado pela junção de todas as microcaracterísticas semânticas das palavras formadoras da sentença. Como as sentenças tratadas por este sistema têm no máximo quatro palavras, o conjunto tem 80(oitenta) microcaracterísticas no total. Para sentenças com menos de quatro palavras o sistema insere um conjunto nulo de microcaracterísticas {0,0,0,0,...,0} para completar a entrada.

Por exemplo:



DESCRIÇÃO DO DICIONÁRIO

Usando as microcaracterísticas anteriores, o dicionário da rede fica como descrito na Figura 01:

(0,1,1,0,1,0,1,0,0,0,5,0,5,0,5,0,1,1,0,1,0,0,0),	{alimento}	(1,0,1,0,0,1,0,0,1,0,0,0,0,1,0,1,0,0,0,0),	{mulher}
(0,1,1,0,0,1,1,0,0,1,0,0,0,1,1,0,1,0,0,0,0),	{batata}	(0,1,1,0,0,1,0,1,0,0,0,0,0,1,1,0,0,0,0,0,0),	{ovelha}
(0,1,1,0,0,1,1,0,0,1,0,0,0,1,0,1,0,1,0,0,0),	{bola}	(0,1,0,1,0,1,1,0,0,0,0,0,1,0,0,1,0,0,0,0,0),	{pedra}
(0,1,1,0,0,1,1,0,0,0,0,0,0,1,1,0,0,1,0,0,0),	{boneca}	(0,1,0,1,1,0,1,0,0,0,0,1,0,1,1,0,0,0,0,0,0),	{prato}
(0,1,1,0,1,0,0,1,0,0,0,0,0,1,0,1,0,0,0,0,0),	{cachorro}	(0,1,1,0,1,0,1,0,0,0,0,1,0,1,1,0,1,0,0,0,0),	{queijo}
(0,1,0,1,0,1,1,0,0,0,1,0,1,0,1,0,1,0,0,0,0),	{cenoura}	(0,1,0,1,1,0,1,0,0,0,1,0,0,1,1,0,0,0,0,0,0),	{vaso}
(0,1,0,1,0,1,1,0,0,0,1,0,0,1,0,1,0,0,0,0,1),	{colher}	(0,1,0,1,0,1,0,1,0,0,0,1,1,0,1,0,0,0,0,0,0),	{vidraça}
(0,1,1,0,0,1,0,1,0,0,0,1,0,1,1,0,0,0,0,0,0),	{cortina}	(0,0),	{<nulo>}
(0,1,0,1,0,1,0,0,1,0,0,0,1,0,0,1,0,0,0,0,0),	{escrivania}	(1,0,0,0,0,1,0,0,0,0,0,1,0,1,0,0,0,1,0,1,1),	{bateu}
(0,1,1,0,0,1,1,0,0,0,5,0,0,0,1,0,5,0,5,0,5,0,0,0),	{galinha}	(1,0,1,0,0,0,1,0,0,0,1,0,0,1,0,1,0,0,1,0,1,0),	{comeu}
(0,1,0,1,1,0,1,0,0,0,1,0,1,0,0,1,0,0,0,0,1),	{garfo}	(1,0,1,0,1,0,0,0,0,0,0,1,1,0,0,1,0,0,1,0,1,0),	{moveu}
(1,0,1,0,1,0,0,1,0,0,0,0,0,1,0,1,0,0,0,0,0),	{garoto}	(1,0,1,0,0,1,0,0,1,0,0,0,0,1,0,0,0,1,0,1,1),	{quebrou}
(1,0,1,0,1,0,0,0,1,0,0,0,0,1,0,1,0,0,0,0,0),	{homem}	(1,0),	{avf}
(0,1,1,0,1,0,0,0,1,0,0,0,1,0,0,1,0,0,0,0,0),	{leão}	(0,1,0),	{avp}
(0,1,1,0,1,0,0,1,0,0,0,0,1,0,0,1,0,0,0,0,0),	{lobo}	(0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),	{avpm}
(0,1,0,5,0,5,1,0,1,0,0,0,0,0,1,0,0,1,0,0,0,5,0),	{macaco}	(0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),	{avpi}
(0,1,1,0,1,0,1,0,0,0,1,0,0,1,1,0,1,0,0,0,0),	{macarrão}	(0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),	{ivp}
(0,1,0,1,1,1,0,1,0,0,0,1,0,0,1,0,1,0,0,1,0),	{martelo}	(0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),	{pv}
(1,0,1,0,0,1,0,1,0,0,0,0,0,1,0,1,0,0,0,0,0),	{menina}	(0,0,0,0,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0),	{avs}

Figura 1

FUNCIONAMENTO DO SISTEMA

O sistema recebe como entrada uma representação canônica da sentença (o conjunto de microcaracterísticas

dos elementos da sentença), sem os artigos, conjunções e quantificadores, fornecida pelo analisador sintático. A rede gera na saída uma única unidade ativa que representa a estrutura de caso da sentença (Figura 02).

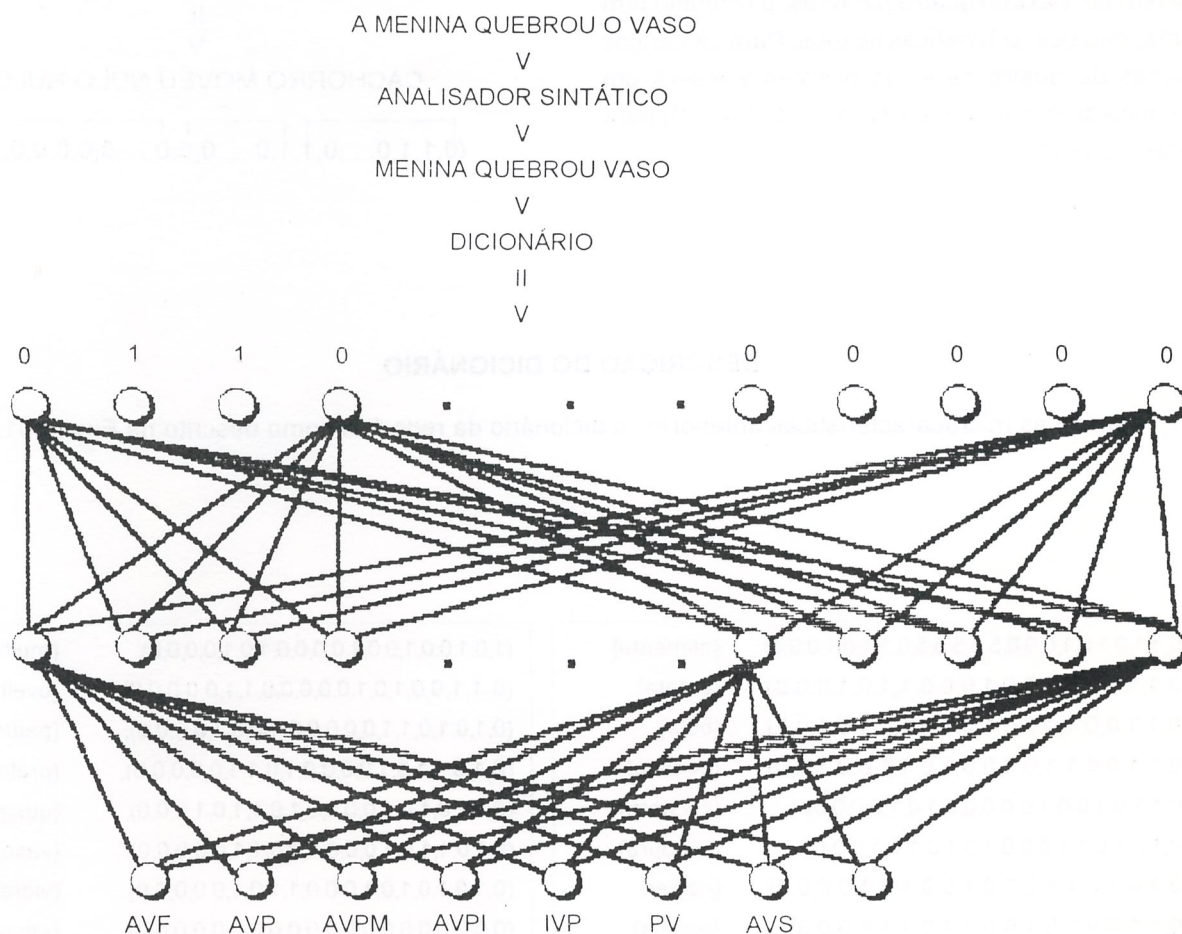


Figura 2

As estruturas de caso que o sistema pode atribuir à sentença são:

- * AVF = Agente Verbo Paciente-Implicito
- * AVPI = Agente Verbo Paciente Instrumento
- * AVP = Agente Verbo Paciente
- * IVP = Instrumento Verbo Paciente
- * AVPM = Agente Verbo Paciente Modificador
- * PV = Paciente Verbo
- * AVS = Agente Verbo Paciente = Agente

DETALHES DO APRENDIZADO

Para realizar o aprendizado o sistema usou o algoritmo "Backpropagation" [RIC91] com fator de

correção de erro $M_i = 0.2$ e função de ativação $F(\text{Alfa}) = 1/(1 + \exp(-\text{Alfa}))$. Para gerar as sentenças usou-se um Gerador Aleatório de Sentenças que escolhe uma das estruturas geradoras de sentenças conhecidas e gera uma frase qualquer com aquela estrutura e sua respectiva estrutura de caso (ver Figura 03), onde:.

humano	[HOMEM, MENINA]
alimento	[CENOURA, GALINHA]
utensílio	[COLHER, GARFO]
animal	[GALINHA, CACHORRO]
predador	[LEÃO, LOBO]
presa	[GALINHA, OVELHA]
objeto-frágil	[VASO, VIDRAÇA]
quebrador	[BOLA, MARTELO]
coisa	[HOMEM, OVELHA]
propriedade	[BOLA, CACHORRO]
batedor	[BOLA, MARTELO]
objeto	[BOLA, MACACO]

Figura 3

ESTRUTURA GERADORA DE SENTENÇAS	ESTRUTURA DE CASO
O humano comeu	AVF
O humano comeu o alimento	AVP
O humano comeu o alimento com o alimento	AVPM
O humano comeu o alimento com o utensílio	AVPI
O animal comeu	AVF
O predador comeu a presa	AVP
O humano quebrou o objeto-frágil	AVP
O humano quebrou o objeto-frágil com o quebrador	AVPI
O quebrador quebrou o objeto-frágil	IVP
O animal quebrou o objeto-frágil	AVP
O objeto-frágil quebrou	PV
O humano bateu na coisa	AVP
O humano bateu no humano com a propriedade	AVPM
O humano bateu na coisa com o batedor	AVPI
O batedor bateu na coisa	IVP
O humano se moveu	AVS
O humano moveu o objeto	AVP
O animal se moveu	AVS
O objeto moveu	PV

CONCLUSÃO

O objetivo inicial do trabalho foi desenvolver um analisador semântico para a língua portuguesa, utilizando-se um sistema que levasse em consideração as particularidades das palavras envolvidas.

Os resultados obtidos, considerando-se simultaneamente tanto a ordem da palavra quanto as restrições semânticas, demonstraram que é possível determinar os casos dos elementos na frase.

Com a implementação do projeto e sua execução pode-se afirmar que seu objetivo foi atingido, possibilitando ainda uma continuidade deste trabalho, pois como é possível determinar o papel de cada palavra na frase, um terceiro módulo poderia ser acrescentado ao sistema, módulo este que seria responsável pelo armazenamento e reagrupamento das frases para consultas futuras.

BIBLIOGRAFIA

- [DAH83] DAHL, Veronica "Logic programming as a representation of knowledge". IEEE Computer, pages 37-41, October, 1983.
- [WAL85] WALTZ, David L.; Pollack, Jordan B. "Massively parallel parsing: A strongly interactive model of natural language interpretation". Cognitive Science 9, pages 51-74 - 1985.
- [MCC86] MC CLELLAND, J. L.; Kawamoto, A. H. "Mechanisms of sentence processing: Assigning roles to constituents of sentences". In 'Parallel Distributed Processing - Explorations in the microstructure of cognition' - Volume 2: Psychological and Biological Models - A Bradford Book, The MIT Press, 1986.
- [CRO87] GROSSBERG, Stephen "Competitive learning: From interactive activation to adaptative resonance". Cognitive Science 11, pages 23-63, 1987.
- [LIP87] LIPPMANN, Richard P. "An introduction to computing with neural nets". IEEE ASSP Magazine, Abril, 1987, pages 4-22.
- [RIC91] RICH, E. & Knight, K. "Artificial Intelligence". 2nd. edition. McGraw-Hill, 1991.
- [ROS93] ROSA, J. L. G. & Netto, M. L. A. "Redes Neurais e Lógica Formal em Processamento de Linguagem Natural". (artigo a ser publicado).

