Pontifícia Universidade Católica
do Rio de Janeiro

Carolina Monte Ferreira Gonçalves

As bases lingüísticas para a busca orientada a idéia

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós-graduação em Letras da PUC-Rio como requisito parcial para a obtenção do título de Mestre em Letras.

Orientadora: Violeta de San Tiago Dantas Barbosa Quental

Rio de Janeiro, março de 2006

PUC-Rio - Certificação Digital Nº 0410508/CA

PONTIFÍCIA UNIVERSIDADE CATÓLICA
DO RIO DE JANEIRO

Carolina Monte Ferreira Gonçalves

As bases lingüísticas para a busca orientada a idéia

Dissertação apresentada como requisito parcial para a obtenção de grau de Mestre pelo Programa de Pós-graduação em Letras do Departamento de Lestras do Centro de Teologia e Ciências Humanas da PUC-Rio. Aprovada pela Comissão examinadora abaixo assinada.

_____
Profa. Violeta de San Tiago Dantas Barbosa Quental
Orientadora
Departamento de Letras – PUC-Rio

_____
Profa. Maria Carmelita Padua Dias
Departamento de Letras – PUC-Rio

_____
Profa. Rove Luiza de Oliveira Chishman
UNISINOS

_____
Prof. Paulo Fernando Carneiro de Andrade
Coordenador Setorial do Centro de Teologia e
Ciências Humanas – PUC-Rio

Rio de Janeiro, ____ de _____de _____.

PUC-Rio - Certificação Digital Nº 0410508/CA

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da autora, da orientadora e da universidade.

Carolina Monte Ferreira Gonçalves

Graduou-se em Letras Português Literaturas na Universidade Estadual do Rio de Janeiro em 2003.

Ficha Catalográfica

Gonçalves, Carolina Monte Ferreira

As bases lingüísticas para a busca orientada a idéia / Carolina Monte Ferreira Gonçalves ; orientador: Violeta de San Tiago Dantas Barbosa Quental. – Rio de Janeiro : PUC, Departamento de Letras, 2006.

132 f. : il. ; 30 cm

Dissertação (mestrado) – Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Letras.

Inclui referências bibliográficas.

1. Letras – Teses. 2. Lingüística computacional. 3. Recuperação de Informação. 4. Teoria sintática. 5. Teoria Semântica. I. Quental, Violeta de San Tiago Dantas Barbosa. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

PUC-Rio - Certificação Digital Nº 0410508/CA

PUC-Rio - Certificação Digital Nº 0410508/CA

Para Leda

Agradecimentos

PUC-Rio - Certificação Digital Nº 0410508/CA

Ao velho Omolu. *Valei-me, meu pai, atotô, Obaluaê*.

À minha mãe, Kátia, ao Pedro e ao Rafa, e aos demais familiares – em especial ao Dani pela tradução do resumo.

Ao meu vô Monte (em memória), cuja pequena biblioteca me iniciou nos estudos lingüísticos.

À Ledinha pelo companheirismo, carinho e paciência.

Aos Justas Juliana, Marcelo e ... Edmar pelos projetos megalomaníacos.

Aos amigos Fernando e Carlos pelas conversas e diversões.

À minha orientadora Violeta Quental pelas importantes contribuições.

Aos colegas e professores da Puc-Rio.

Aos funcionários do Departamento de Letras, em especial à sempre prestativa Chiquinha.

Ao CNPq e à Puc-Rio pelos auxílios concedidos.

Resumo

Gonçalves, Carolina Monte Ferreira. **As bases lingüísticas para a busca orientada a idéia**. Rio de Janeiro, 2006. 132p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

A busca orientada a idéia é um novo paradigma para mecanismos de busca em acervos compostos por arquivos de texto. Esse paradigma visa resolver um problema comum aos mecanismos de busca: exigir que o usuário preveja as palavras contidas nos documentos que possam conter a informação que procura, impossibilitando-o, assim, de se concentrar diretamente na informação desejada. Buscando solucionar esse problema, são propostas as bases lingüísticas para o desenvolvimento de um modelo teórico preliminar que acrescente dados semânticos aos arquivos de texto. Nesse modelo, a informação semântica de um texto é representada através do que se chamou de estrutura de conceitos. O principal intuito das estruturas de conceitos é representar de uma mesma maneira frases que expressem o mesmo significado, ou seja, as paráfrases apresentam a mesma estrutura de conceitos. Serão expostos nesta dissertação os primeiros elementos do modelo em suas partes semântica, sintática e textual, além da integração entre as mesmas. A dissertação apresenta ainda um estudo de caso a fim de exemplificar o desenvolvimento de uma aplicação para busca de arquivos de texto em que essa tecnologia seria usada.

PUC-Rio - Certificação Digital Nº 0410508/CA

Palavras-chaves

Letras; lingüística computacional; recuperação de informação; teoria sintática; teoria semântica.

Abstract

Gonçalves, Carolina Monte Ferreira. **Linguistic basis for idea-oriented search**. Rio de Janeiro, 2006. 132p. Dissertação de Mestrado – Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

The idea-oriented search is a new pattern for search engines whose databases are composed by text files. This pattern sets out to solve a usual problem for search engines: demanding that users foresee which words are contained in the desired document, keeping them from focusing on the information they are indeed seeking. To solve this problem, the linguistic bases for the development of a theoretical model that can add semantical data to the text files are laid down. In this model, the semantical information of a text is represented by what has been referred to as structure of concepts. The main goal of the structure of concepts is to give one single representation to sentences that express the same meaning. Thus, paraphrases present the same structure of concepts. In this dissertation, the first elements of this model are exposed in its semantical, syntactic and textual parts. Also present are the integration of these elements. A small case study is presented as well, with the intention of illustrating the development of an application for text files databases search engines in which this technology is used.

PUC-Rio - Certificação Digital Nº 0410508/CA

Keywords

Computational Linguistics; information retrieval; syntactic theory; semantic theory.

Sumário

1 Motivação 12
1.1 A busca orientada a palavra 13
1.2 A busca orientada a idéia 14
2 Influências 16
2.1 A gramática de valências 16
2.2 O léxico gerativo 20
2.3 Representações de conceitos 24
2.3.1 Traços semânticos 24
2.3.2 Redes semânticas e *frames* 28
2.4 A *web* semântica 29
2.5 A UNL e a dependência conceitual 31
2.6 Conclusão do capítulo 34
3 A proposta 35
3.1 Representação semântica 37
3.1.1 O significado do texto como o significado das palavras que o compõem 38
3.1.2 Considerações gramaticais 41
3.1.3 O significado da palavra definido a partir do significado do texto 47
3.2 Formalizando a proposta 58
3.2.1 Formalizando as entradas lexicais 62
3.2.2 A estrutura de conceitos das palavras 81
3.2.3 O componente textual do modelo 88
4 Estudo de caso 94
4.1 O dicionário eletrônico 96
4.2 Problemas ainda não resolvidos pelo modelo 117
4.2.3Trabalhos futuros 119
5 Conclusão 121
6 Bibliografia 123
7 Anexo – Corpus 127

PUC-Rio - Certificação Digital Nº 0410508/CA

Lista de Figuras

Figura 1 – rede semântica 27
Figura 2 – modelo de *frames* 28
Figura 2 – união dos significados 39
Figura 3 – seqüência dos significados 40
Figura 4 – seqüência dos significados em uma paráfrase 40
Figura 5 – decompondo um significado 48
Figura 6 – relação argumental 49
Figura 7 – significados 49
Figura 8 – redistribuindo os significados 50
Figura 9 – atribuindo setas iguais às paráfrases 50
Figura 10 – nomes de animais 51
Figura 11 – a relação caça 51
Figura 12 – grafo da caça 51
Figura 13 – grafo com outra configuração 52
Figura 14 – novo grafo de caça 52
Figura 15 – grafos congruentes 53
Figura 16 – exemplo de grafos não congruentes 53
Figura 17 – grafo de uma sentença não paráfrase 54
Figura 18 – um grafo contendo outro grafo 56
Figura 19 – subgrafo 56
Figura 20 – habilidades de um grafo 57
Figura 21 – grafo de hipônimos contendo o grafo de hiperônimos 58
Figura 22 – algoritmo 1 59
Figura 23 – algoritmo 2 59
Figura 24 – algoritmo 3 60
Figura 25 – algoritmo 4 61
Figura 26 – algoritmo 5 62
Figura 27 – núcleos do sintagma adjetivo e nominal 67

PUC-Rio - Certificação Digital Nº 0410508/CA

Figura 28 – ignorando o núcleo do sintagma preposicional 68
Figura 29 – hierarquia entre sintagma nominal do sujeito e sintagma verbal 68
Figura 30 – hierarquia entre sintagma nominal do objeto e sintagma verbal 69
Figura 31 – hierarquia entre os argumentos do verbo 69
Figura 32 – hierarquia entre verbo e seu argumento 69
Figura 33 – hierarquia entre o adjetivo e seu argumento 70
Figura 34 – estrutura do sintagma substantivo 71
Figura 35 – sintagma nominal como argumento do verbo 71
Figura 36 – substituição no sintagma nominal 73
Figura 37 73
Figura 38 – estrutura de um sintagma com recursão 73
Figura 39 – recursão no sintagma nominal 74
Figura 40 – sintagma preposicionado como argumento do substantivo 75
Figura 41 – recursão no sintagma verbal 75
Figura 42 75
Figura 43 – representação da estrutura de argumentos 77
Figura 44 – exemplo de estrutura de argumento 77
Figura 45 – seleção semântica 78
Figura 46 – representação da seleção semântica na estrutura de argumentos 78
Figura 47 – exemplo de seleção semântica 78
Figura 48 – grafos congruentes 79
Figura 49 – ordenação topológica 80
Figura 50 – ordenação topológica na estrutura de argumentos 81
Figura 51 – hiperônimo e hipônimo 84
Figura 52 – exemplo de atribuição de conceitos 86
Figura 53 – representação da estrutura de conceitos 87
Figura 54 – estrutura de argumentos selecionando um grafo 87
Figura 55 – exemplo de grafo conexo 89
Figura 56 – exemplo de grafo desconexo 89
Figura 57 – coordenação 1 91
Figura 58 – coordenação 2 91

PUC-Rio - Certificação Digital Nº 0410508/CA

Figura 59 – anáfora 1 91
Figura 60 – anáfora 2 92
Figura 61 98
Figura 62 98
Figura 63 99
Figura 64 99
Figura 65 101
Figura 66 101
Tabela 1 101
Tabela 2 102
Figura 67 103
Tabela 3 104
Figura 68 106
Figura 69 106
Figura 70 107
Tabela 4 108
Figura 71 108
Tabela 5 109
Tabela 6 112
Tabela 7 113
Tabela 8 114
Tabela 9 116

PUC-Rio - Certificação Digital Nº 0410508/CA