Pontifícia Universidade Católica
do Rio de Janeiro

Maria Cláudia de Freitas

Elaboração automática de ontologias de domínio: discussão e resultados

Tese de Doutorado

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Letras da PUC-Rio.

Orientador: Violeta de San Tiago Dantas Barbosa Quental

Rio de Janeiro, janeiro de 2007

Pontifícia Universidade Católica
do Rio de Janeiro

Maria Cláudia de Freitas

Elaboração automática de ontologias de domínio: discussão e resultados

Tese apresentada como requisito parcial para obtenção do título de Doutor pelo Programa de Pós-Graduação em Letras da PUC-Rio. Aprovada pela Comissão Examinadora abaixo assinada.

Profa. Violeta de San Tiago Dantas Barbosa Quental
Orientador
Departamento de Letras – PUC-RIO

Profa. Margarida Maria de Paula Basilio
Departamento de Letras – PUC-RIO

Profa. Helena Franco Martins
Departamento de Letras – PUC-RIO

Profa. Vera Lucia Strube de Lima
Departamento de Fundamentos da Computação – PUC-RS

Prof. Geraldo Bonorino Xexéo
Instituto Alberto Luiz Coimbra de Pós-Graduação
e Pesquisa de Engenharia – UFRJ

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

Rio de Janeiro, _____ de _____ de _____.

Maria Cláudia de Freitas

Graduou-se em letras (Português-Literatura) pela PUC-Rio em 1997. Obteve o título de Mestre em Letras pela PUC-Rio em 2000 e concluiu, em 2007, Doutorado em Letras (área de concentração: Estudos da Linguagem) na mesma instituição. Leciona na PUC-Rio desde 2002, ministrando cursos na área de Comunicação e Expressão, Lingüística e Língua Portuguesa. Participa, como pesquisadora, de projetos na área de Lingüística Computacional, desenvolvidos no CLIC - Centro de Lingüística Computacional da PUC-Rio.

Ficha Catalográfica

Freitas, Maria Cláudia de

Elaboração automática de ontologias de domínio : discussão e resultados / Maria Cláudia de Freitas ; orientadora: Violeta de San Tiago Dantas Barbosa Quental. – 2007.

142 f. ; 30 cm

Dissertação (Mestrado em Letras)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007.

Inclui bibliografia

1. Letras – Teses. 2. Ontologia. 3. Taxonomia. 4. Hierarquia lexical. 5. Extração de informação. 6. Relações semânticas. 7. Léxico. 8. Nomes próprios. I. Quental, Violeta de San Tiago Dantas Barbosa. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Letras. III. Título.

CDD: 400

Agradecimentos

À Violeta Quental – pelo apoio, incentivo, amizade, generosidade, disponibilidade e, sobretudo, pelo bom humor e leveza com que trata o mundo acadêmico.

À Claudia Oliveira - pela generosidade, pela amizade, pelas discussões e por ter, em grande parte, viabilizado a interdisciplinaridade deste trabalho.

À Helena Martins – pela apresentação de "um outro ponto de vista" sobre a linguagem e pela preciosa – e luxuosa – assessoria teórica.

À Erica Rodrigues – pelo "SOS gramática" sempre disponível, pela amizade.

Ao Renato Paes Leme – pela paciência e prontidão com que transformava meus pedidos em um programa de computador.

Ao Cícero Nogueira dos Santos – pelo "suporte 24hs" nos sintagmas nominais, etiquetagens e afins, pelas dicas computacionais e pela paciência.

Ao Marcelo, à Ana e ao Raul – pelas muitas horas que passaram discutindo e avaliando listas de "X é um Y".

Ao CLIC – pela troca valiosa de idéias.

À Chiquinha e à Dy – pela presteza, pela paciência e pelo sorriso.

Resumo

Freitas, Maria Claudia de; Quental, Violeta de San Tiago Dantas Barbosa. **Elaboração automática de ontologias de domínio: discussão e resultados.** Rio de Janeiro, 2007. 142p. Tese de Doutorado - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

O objetivo deste trabalho é apresentar subsídios para a elaboração automática, a partir de corpus, de ontologias específicas quanto ao domínio. Para tanto, assumo que determinadas relações semânticas, como a hiperonímia, podem estar sistematicamente expressas em textos por meio de determinados padrões léxico-sintáticos. Tomando como ponto de partida alguns desses padrões, descritos originalmente em Hearst (1992, 1998), (i) identifico novos padrões para a expressão da relação de hiperonímia; (ii) adapto e refino três padrões já existentes (Hearst, 1992), tendo em vista especificidades da língua portuguesa; (iii) faço um cruzamento entre as informações extraídas com os padrões, a fim de gerar inferências. A perspectiva teórica subjacente é inspirada por reflexões wittgensteinianas sobre o significado, e se mostrou produtiva na medida em que legitima os dados vindos do corpus e as relações de significado que nele aparecem. O modelo de ontologia proposto caracteriza-se principalmente por: (i) não conter categorias pré-definidas, já que categorias são construtos humanos, abstrações que refletem uma perspectiva particular do mundo. A idéia de sustentar a ontologia em corpus busca deslocar o espaço de discussão sobre quais seriam as categorias relevantes de um domínio: as categorias que emergem do corpus refletiriam o conhecimento implícito do domínio em questão; (ii) não conter definições criadas a priori, sendo o significado de cada item decorrente das relações entre as palavras. A metodologia – extração das relações por meio de regras e posterior cruzamento para a realização de inferências – foi aplicada em um corpus do domínio saúde e um corpus genérico. Os resultados positivos indicam que sua utilização pode ser uma importante aliada na elaboração de ontologias e, também, uma ferramenta de auxílio a lexicógrafos e a sistemas de classificação semântica de nomes próprios. Em termos gerais, a metodologia apresenta como principais vantagens (i) a facilidade na automação do processo, minimizando a intervenção humana; (ii) facilidade na categorização de domínios

especializados; (iii) maior dinamicidade, pois o fato de o corpus poder ser constantemente atualizado faz com que esteja menos sujeito a falhas.

Palavras-chave

ontologia; taxonomia; hierarquia lexical; extração de informação; relações semânticas; léxico; nomes próprios

Abstract

Freitas, Maria Claudia de; Quental, Violeta de San Tiago Dantas Barbosa. **Elaboração automática de ontologias de domínio: discussão e resultados.** Rio de Janeiro, 2007. 142p. PhD - Departamento de Letras, Pontifícia Universidade Católica do Rio de Janeiro.

The main goal of this work is to present an automated method for building domain-specific corpus-based ontologies. The assumption is that semantic relationships, such as hypernym, can be systematically expressed through lexical-syntactic patterns. Starting with some of these patterns, originally described in Hearst (1992), I (i) identify new patterns that express hypernym; (ii) adapt three other patterns (Hearst, 1992), considering specificities of the Portuguese language; and (iii) intersect these results, in order to produce inferences. The theoretical approach is inspired by the wittgensteinian ideas about meaning. The resulting ontology's most prominent features are: (i) the fact that it does not have a priori categories, since categories are human constructs, abstractions that reflect a particular world view. Instead of discussing what should be the main categories in a domain, sustaining the ontology on corpora assumes that the corpus reflects the implicit knowledge of a given domain; and (ii) the fact that it does not have a priori definitions: the meaning of a word is derived from its relations with other words. The method – automatic extraction of semantic relations through rules, and the intersection of this information in order to produce inferences – was applied to two corpora: a health domain corpus and a generic corpus. The positive results show that the method can be very useful in ontology building and it can also be a valuable tool for lexicographers and named entity recognition systems. The main advantages of the method are (i) the simplicity of automating the process of ontology building; (ii) the ease of categorizing specialized domains, and (iii) its dynamicity, since the possibility of constantly updating the corpus makes it less subject to errors.

Palavras-chave

ontology; taxonomy; lexical hierarchy; lexicon; proper nouns

Sumário

Lista de figuras

Lista de tabelas

Lista de quadros