Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Henrique Ribeiro de Assis

Distant Supervision for Relation Extraction using Ontology Class Hierarchy-Based Features

DISSERTAÇÃO DE MESTRADO

Dissertation presented to the Programa de Pós-Graduação em Informática of the Departamento de Informática, PUC-Rio as partial fulfillment of the requirements for the degree of Mestre em Informática

Advisor: Prof. Marco Antonio Casanova

Rio de Janeiro
March 2014

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Henrique Ribeiro de Assis

Distant Supervision for Relation Extraction using Ontology Class Hierarchy-Based Features

Dissertation presented to the Programa de Pós-Graduação em Informática, of the Departamento de Informática do Centro Técnico Científico da PUC-Rio, as partial fulfillment of the requirements for the degree of Mestre.

Prof. Marco Antonio Casanova
Advisor
Departamento de Informática – PUC-Rio

Prof. Alberto Henrique Frade Laender
Departamento de Ciência da Computação – UFMG

Prof. Ruy Luiz Milidiú
Departamento de Informática – PUC-Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico da PUC-Rio

Rio de Janeiro, March 20th, 2014

Pedro Henrique Ribeiro de Assis

Graduated in Computer Science from Universidade Federal de Minas Gerais (UFMG), Minas Gerais - Brazil in 2011. He joined the Master in Informatics at Pontifical Catholic University of Rio de Janeiro (PUC-Rio) in 2012.

Bibliographic data

Ribeiro de Assis, Pedro Henrique

Distant Supervision for Relation Extraction using Ontology Class Hierarchy-Based Features / Pedro Henrique Ribeiro de Assis; advisor: Marco Antonio Casanova. – 2014.

64 f. : il. (color) ; 30 cm

Dissertação (Mestrado em Informática)–Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2014.

Inclui bibliografia

1. Informática – Teses. 2. Extração de relações. 3. Web semântica. 4. Processamento natural de linguagens. 5. Aprendizado de máquina. 6. Supervisão a distância. I. Casanova, Marco Antonio. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

I would like to thank one of the best advisors I have ever had, professor Marco Antonio Casanova. His support, wisdom, experience and patience were always a source of motivation for me and were essential for this work.

To Prof. Ruy Milidiú for his support, attention and help with mathematical models applied in this work.

To PUC-Rio for funding my research and for giving me the opportunity to be a student of the Department of Informatics which is a fantastic and remarkable academic community.

To my parents for unconditional love.

And to my friends for always helping me to stay motivated and focused, except, of course, on sunny weekends.

Abstract

Assis, Pedro; Casanova, Marco Antonio (Advisor). **Distant Supervision for Relation Extraction using Ontology Class Hierarchy-Based Features.** Rio de Janeiro, 2014. 64p. MSc. Dissertation – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Relation extraction is a key step for the problem of rendering a structure from natural language text format. In general, structures are composed by entities and relationships among them. The most successful approaches on relation extraction apply supervised machine learning on hand-labeled corpus for creating highly accurate classifiers. Although good robustness is achieved, hand-labeled corpus are not scalable due to the expensive cost of its creation. In this work we apply an alternative paradigm for creating a considerable number of examples of instances for classification. Such method is called distant supervision. Along with this alternative approach we adopt Semantic Web ontologies to propose and use new features for training classifiers. Those features are based on the structure and semantics described by ontologies where Semantic Web resources are defined. The use of such features has a great impact on the precision and recall of our final classifiers. In this work, we apply our theory on corpus extracted from Wikipedia. We achieve a high precision and recall for a considerable number of relations.

Keywords

Relation Extraction, Distant Supervision, Semantic Web, Machine Learning, Natural Language Processing

Resumo

Assis, Pedro; Casanova, Marco Antonio. **Supervisão à distância em extração de relacionamentos usando características baseadas em hierarquia de classes em ontologias.** Rio de Janeiro, 2014. 64p. Dissertação de Mestrado – Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Extração de relacionamentos é uma etapa chave para o problema de identificação de uma estrutura em um texto em formato de linguagem natural. Em geral, estruturas são compostas por entidades e relacionamentos entre elas. As propostas de solução com maior sucesso aplicam aprendizado de máquina supervisionado a corpus anotados à mão para a criação de classificadores de alta precisão. Embora alcancem boa robustez, corpus criados à mão não são escaláveis por serem uma alternativa de grande custo. Neste trabalho, nós aplicamos um paradigma alternativo para a criação de um número considerável de exemplos de instâncias para classificação. Tal método é chamado de supervisão à distância. Em conjunto com essa alternativa, usamos ontologias da Web semântica para propor e usar novas características para treinar classificadores. Elas são baseadas na estrutura e semântica descrita por ontologias onde recursos da Web semântica são definidos. O uso de tais características tiveram grande impacto na precisão e recall dos nossos classificadores finais. Neste trabalho, aplicamos nossa teoria em um corpus extraído da Wikipedia. Alcançamos uma alta precisão e recall para um número considerável de relacionamentos.

Palavras-chave

Extração de Relacionamentos, Supervisão à Distância, Web Semântica, Aprendizado de Máquina, Processamento Natural de Linguagens

Table of Contents

List of Figures

List of Tables