

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

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

All rights reserved

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

1	Introduction	11
1.1	Motivation	11
1.2	Goals and Contributions	12
1.3	Related Work	12
1.4	Dissertation Structure	14
2	Background	15
2.1	The Semantic Web	15
2.1.1	An Architecture for the Semantic Web	15
2.1.2	The Coding Layer	17
2.1.3	The Structure Layer	18
2.1.3.1	XML + XML Schema + XML Namespaces	18
2.1.3.2	RDF + RDF Schema	23
2.1.4	The Inference Layer	26
2.1.4.1	Ontology Description Languages	26
2.1.4.2	Rule Inference	29
2.1.5	The Linked Data Principles	31
2.1.6	The Linked Open Data Project	32
2.2	Natural Language Processing	33
2.2.1	Sentence Boundary Disambiguation	33
2.2.2	Part-of-Speech Tagging	35
2.3	Classification Methods	37
2.3.1	Overview	37
2.3.2	Logistic Regression	38

2.3.3. Multi-class Perceptron	40
2.4 Summary	42
3 The Relation Extraction Problem	42
3.1 Approach	42
3.2 Extraction Task Definition	43
3.3 Features	44
3.3.1 Natural Language Processing Based Features	44
3.3.2 Ontology Class Hierarchy Based Feature	46
3.4 Summary	48
4 Experiments	49
4.1 Corpus	50
4.2 Experimental Setup	52
4.3 Held-out Evaluation	54
4.4 Human Evaluation	57
4.5 Summary	59
5 Conclusion	60
5.1 Contributions	60
5.2 Limitations and Future Work	60
6 Bibliography	62

List of Figures

Figure 1: The three architectural bases of the Web	16
Figure 2: An example of the referencing mechanism using URI	17
Figure 3: An example of XML for representing books	18
Figure 4: Two stretches of XML files. At the top an HTML table and at the bottom values about a table (furniture)	19
Figure 5: Use of namespaces in an XML document with name conflicts	20
Figure 6: Example of an XML document representing a shipping order	21
Figure 7: Example of an XML Schema that is valid for the XML Document in Figure 6	22
Figure 8: Example of an RDF resource in N3 format	23
Figure 9: Example of an RDF resource in RDF/XML format	23
Figure 10: The graph model described in the RDF examples of Figure 8 and 9	24
Figure 11: Example of the use of constructs of RDF Schema to define a class hierarchy	25
Figure 12: An example of taxonomy of the kingdoms of life	26
Figure 13: An example of OWL file describing an transport ontology	29
Figure 14: The LOD Cloud Diagram at September 2011	32
Figure 15: An example of sigmoid function	37
Figure 16: A class hierarchy sub-tree from DBpedia	45
Figure 17: A class hierarchy sub-tree from DBpedia with cut $h = 2$	46
Figure 18: Histogram for the number of occurrences for each relation	50

List of Tables

Table 1: Lexical features and examples	43
Table 2: Features of s based on Ontology Class Based Hierarchy	
Features	45
Table 3: Top 20 relations in number of examples in the dataset	51
Table 4: Relations for a classifier trained with lexical features only	52
Table 5: Top 20 relations for a classifier trained with semantic features	
only	53
Table 6: Top 20 relations for a classifier trained with lexical and	
semantic features	54
Table 7: Number of classes with at least 70% of F-measure by set of	
features	54
Table 8: Average accuracy for the top 10 relations in examples in our	
dataset for human evaluation of a sample of 100 predictions	55