

William Paulo Ducca Fernandes

Quotation Extraction for Portuguese

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
Programa de Pós-Graduação em Informática

Rio de Janeiro
April 2012

William Paulo Ducca Fernandes

Quotation Extraction for Portuguese

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. Ruy Luiz Milidiú

Rio de Janeiro
April 2012



William Paulo Ducca Fernandes

Quotation Extraction for Portuguese

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 em Informática

Prof. Ruy Luiz Milidiú

Advisor

Departamento de Informática — PUC–Rio

Prof. Daniel Schwabe

Departamento de Informática — PUC–Rio

Prof. Marco Antonio Casanova

Departamento de Informática — PUC–Rio

Prof. José Eugenio Leal

Coordinator of the Centro Técnico Científico — PUC–Rio

Rio de Janeiro, April 9, 2012

All rights reserved.

William Paulo Ducca Fernandes

Graduated in 2008 from the Universidade Federal de Juiz de Fora (UFJF) in Computer Science. Joined the LEARN lab at the Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio) in 2010, focusing his research on Machine Learning and Natural Language Processing.

Bibliographic data

Fernandes, William Paulo Ducca

Quotation Extraction for Portuguese / William Paulo Ducca Fernandes ; advisor: Ruy Luiz Milidiú. — 2012.

59 f. : il. ; 30 cm

Dissertação (Mestrado em Informática)-Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2012.

Inclui bibliografia

1. Informática – Dissertação. 2. Aprendizado de Máquina. 3. Processamento de Linguagem Natural. 4. Extração de Informação. 5. Extração de Citações. 6. Aprendizado de Transformações Guiado por Entropia. 7. Perceptron Estruturado. 8. Agendamento de Tarefas Ponderado. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

Acknowledgments

Firstly, I would like to thank God for guiding me in this journey full of challenges. By His Grace, I am here taking one more step of my life. I can honestly say Ebenezer, “Hitherto hath the Lord helped me.” I would also like to thank my parents, Ricardo and Delza, for being such a strong presence in my life. Thank you for the love, affection and dedication which made of me the man I am today. I dedicate this victory to you both.

I thank you, André, my brother, with whom I have shared a lot of experiences; Jesana, my sister, Tiago, my brother-in-law, and Isabella, my niece, for making my life happier. I want to say thanks to Leyla, my girlfriend, who I met toward the end of this walk, for being such a lovely companion. I would like to thank Ruy, my advisor, for counselling me throughout this work, answering my questions and teaching me many work and life lessons. I would like to thank Eraldo, with whom I interacted quite a bit and learned a lot from, for really helping me with Machine Learning issues and for being a role model for me. Last and not least, I would like to thank Eduardo for his active participation in the annotation process, Carlos, for his joy and good mood, Leandro, for his funny comments, and CNPq, for the scholarship. For you all, my sincere thank you!

Resumo

Fernandes, William Paulo Ducca; Milidiú, Ruy Luiz. **Extração de Citações para o Português**. Rio de Janeiro, 2012. 59p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A Extração de Citações consiste na identificação de citações de um texto e na associação destas com seus autores. Neste trabalho, apresentamos um sistema de Extração de Citações para Português. A tarefa de Extração de Citações já foi abordada usando diversas técnicas e para diversas línguas. Nossa proposta é diferente dos trabalhos anteriores, pois usamos Aprendizado de Máquina para construir automaticamente regras especializadas ao invés de regras criadas por humanos. Modelos de Aprendizado de Máquina geralmente apresentam forte capacidade de generalização comparados a modelos feitos por humanos. Além disso, nós podemos facilmente adaptar nosso modelo para outras línguas, precisando apenas de uma lista de verbos de citação para uma dada língua. Os sistemas propostos anteriormente provavelmente precisariam de uma adaptação no conjunto de regras de forma a classificar corretamente as citações, o que consumiria tempo. Nós atacamos a tarefa de Extração de Citações usando um modelo para o algoritmo de *Aprendizado de Transformações Guiado por Entropia* e um modelo para o algoritmo do *Perceptron Estruturado*. Com o objetivo de treinar e avaliar o sistema, nós construímos o corpus GLOBOQUOTES com notícias extraídas do portal GLOBO.COM. Adicionamos etiquetas morfossintáticas ao corpus, utilizando um anotador estado da arte. O Perceptron Estruturado baseado no agendamento de tarefas ponderado tem desempenho $F_{\beta=1}$ igual a 76,80%.

Palavras-chave

Aprendizado de Máquina. Processamento de Linguagem Natural. Extração de Informação. Extração de Citações. Aprendizado de Transformações Guiado por Entropia. Perceptron Estruturado. Agendamento de Tarefas Ponderado.

Abstract

Fernandes, William Paulo Ducca; Milidiú, Ruy Luiz (advisor). **Quotation Extraction for Portuguese**. Rio de Janeiro, 2012. 59p. MSc Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Quotation Extraction consists of identifying quotations from a text and associating them to their authors. In this work, we present a Quotation Extraction system for Portuguese. Quotation Extraction has been previously approached using different techniques and for several languages. Our proposal differs from previous work since we use Machine Learning to automatically build specialized rules instead of human-derived rules. Machine Learning models usually present stronger generalization power compared to human-derived models. In addition, we are able to easily adapt our model to other languages, needing only a list of verbs of speech for a given language. The previously proposed systems would probably need a rule set adaptation to correctly classify the quotations, which would be time consuming. We tackle the Quotation Extraction task using one model for the *Entropy Guided Transformation Learning* algorithm and another one for the *Structured Perceptron* algorithm. In order to train and evaluate the system, we have build the GLOBOQUOTES corpus, with news extracted from the GLOBO.COM portal. We add part-of-speech tags to the corpus using a state-of-the-art tagger. The Structured Perceptron based on weighted interval scheduling obtains an $F_{\beta=1}$ score of 76.80%.

Keywords

Machine Learning. Natural Language Processing. Information Extraction. Quotation Extraction. Entropy Guided Transformation Learning. Structured Perceptron. Weighted Interval Scheduling.

Contents

1	Introduction	11
2	Quotation Extraction	14
2.1	The Task	14
2.2	Related Work	18
3	Machine Learning Algorithms	20
3.1	Entropy Guided Transformation Learning	20
3.2	Structured Perceptron	23
4	Models	26
4.1	Entropy Guided Transformation Learning	26
4.2	Structured Perceptron	32
5	Experiments	38
5.1	Corpus	38
5.2	Experimental Setup	40
5.3	Quality Results	41
5.4	Error Analysis	43
6	Conclusions	44
	Bibliography	46
7	Glossary	51
A	Annotation Guidelines	52

List of Figures

2.1	Quotation identification subtask	14
2.2	Author candidates for the quotation association to its author subtask.	14
2.3	Association between quotation and author subtask	15
2.4	Task decomposition diagram	15
2.5	Several examples of quotations	16
2.6	Several examples in which quotation marks are not used to delimit quotations	16
2.7	Several examples of association between quotation and author	17
2.8	Several examples of association between quotation and author in which the author is not the nearest coreference	17
3.1	TBL algorithm	21
3.2	ETL algorithm	22
3.3	Decision tree	22
3.4	Decision tree after the elimination process and the extracted templates.	23
3.5	Structured Perceptron algorithm	24
4.1	First rule exception example	27
4.2	First rule application example for the quote beginning identification subtask.	28
4.3	Second rule application example for the quote beginning identification subtask.	28
4.4	First rule application example for the quote end identification subtask.	29
4.5	Second rule application example for the quote end identification subtask.	29
4.6	Third rule application example for the quote end identification subtask.	30
4.7	First rule application example for the association between quotation and author subtask.	32
4.8	Second rule application example for the association between quotation and author subtask.	32
4.9	Linear time algorithm for WIS	33
4.10	Example to illustrate the input features for Structured Perceptron	34
4.11	Example to illustrate the construction of x	35
5.1	GLOBO.COM corpus distribution of the main news genres	38
5.2	GLOBO.COM corpus distribution of the remaining news genres	39
5.3	Distribution of distances from quotations to their authors in GLOBOQUOTES	40
5.4	Several examples of SP-WIS recall errors	43
A.1	Illustrative example of NE annotation	52
A.2	Application example of the first annotation guideline for NE	52
A.3	Application example of the second annotation guideline for NE	53

A.4	Application example of the third annotation guideline for NE	54
A.5	Application example of the fourth annotation guideline for NE	54
A.6	Application example of the fifth annotation guideline for NE	54
A.7	Application example of the sixth annotation guideline for NE	55
A.8	Application example of the seventh annotation guideline for NE	55
A.9	Application example of the eighth annotation guideline for NE	55
A.10	Application example of the ninth annotation guideline for NE	56
A.11	Application example of the tenth annotation guideline for NE	56
A.12	Application example of the eleventh annotation guideline for NE	56
A.13	Illustrative example of coreference annotation	57
A.14	Application example of the first annotation guideline for coreference	57
A.15	Application example of the second annotation guideline for coreference	57
A.16	Illustrative examples of different types of annotated quotations	58
A.17	Application example of the annotation guideline for association between quotation and coreference	59

List of Tables

1.1	Quotation Extraction performance on the test set	13
2.1	Examples of the several quotation types found in news	18
4.1	Derived feature Bounded Chunk example	27
4.2	Derived feature Verb of Speech Neighbourhood example	27
4.3	Derived feature First Letter Upper Case example	27
4.4	Rule application example for the quote bounds subtask.	30
4.5	Derived feature Coreference Indicator example	30
4.6	Sentence before concatenation	31
4.7	Sentence after concatenation	31
4.8	Sentence before elimination	31
4.9	Sentence after elimination	31
5.1	Annotated corpus excerpt	40
5.2	Annotated corpus set sizes	41
5.3	Performance of the Quotation Extraction task	42
5.4	Performance of the quote beginning identification subtask	42
5.5	Performance of the quote end identification subtask	42
5.6	Performance of the quote bounds subtask	42