Pontifícia Universidade Católica
do Rio de Janeiro

Guilherme Carlos De Napoli Ferreira

A Machine Learning Approach for Portuguese Text Chunking

Dissertação de Mestrado

Dissertation presented to the Postgraduate Program in Informatics 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
June 2011

Pontifícia Universidade Católica
do Rio de Janeiro

Guilherme Carlos De Napoli Ferreira

A Machine Learning Approach for Portuguese Text Chunking

Dissertation presented to the Postgraduate Program in Informatics of the Departamento de Informática, PUC–Rio as partial fulfillment of the requirements for the degree of Mestre em Informática. Approved by the following commission:

Prof. Ruy Luiz Milidiú
Advisor
Departamento de Informática — PUC–Rio

Prof. Daniel Schwabe
Departamento de Informática — PUC–Rio

Prof. Violeta de San Tiago Dantas Barbosa Quental
Departamento de Letras — PUC–Rio

Prof. José Eugenio Leal
Coordinator of the Centro Técnico Científico — PUC–Rio

Rio de Janeiro — June 10, 2011

Guilherme Carlos De Napoli Ferreira

Graduated in 2008 from the Instituto Militar de Engenharia (IME) in Computer Engineering. Joined the LEARN lab at the Pontifícia Universidade Católica do Rio de Janeiro in 2009, focusing his research on Machine Learning and Natural Language Processing.

Bibliographic data

Ferreira, Guilherme Carlos De Napoli

A machine learning approach for Portuguese text chunking / Guilherme Carlos De Napoli Ferreira ; advisor: Ruy Luiz Milidiú. — 2011.
63 f. : il. ; 30 cm

Dissertação (mestrado)-Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, 2011.
Inclui bibliografia

1. Informática – Teses. 2. Aprendizado de máquina. 3. Processamento de linguagem natural. 4. Segmentação textual. 5. Análise sintática superficial. 6. Aprendizado de transformações guiado por entropia. I. Milidiú, Ruy Luiz. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To Guilherme, Gisela and Gabriella:
the only reason I am here is because you are here.

Acknowledgments

To my family, whose love and encouragement never faltered.

To my girlfriend Letícia Sampaio, for her deep affection and for providing me the rational and emotional support I needed in countless situations.

To my lifelong friends from all different environments, who helped me become who I am now and achieve this victory.

To my colleagues from IME, for being part of my intellectual development.

To Caio Valentim, Daniel Fleischman, Eduardo Cardoso, Paulo Gomide and all my other programming contest fellows, for being the fundamental part of one of the most exciting academic adventures of my life.

To my advisor Ruy Milidiú, for guiding me through my research and supporting me on my career decisions.

To Carlos Crestana, Eduardo Motta, Eraldo Fernandes and all my LEARN colleagues, for helping me with this work and being exceptional coworkers.

To PUC-Rio, for providing an outstanding academic environment, and CNPq, for the financial support.

Abstract

Ferreira, Guilherme Carlos De Napoli; Milidiú, Ruy Luiz. **A Machine Learning Approach for Portuguese Text Chunking**. Rio de Janeiro, 2011. 63p. M.Sc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Text chunking is a very relevant Natural Language Processing task, and consists in dividing a sentence into disjoint sequences of syntactically correlated words. One of the factors that highly contribute to its importance is that its results are used as a significant input to more complex linguistic problems. Among those problems we have full parsing, clause identification, dependency parsing, semantic role labeling and machine translation. In particular, Machine Learning approaches to these tasks greatly benefit from the use of a chunk feature. A respectable number of effective chunk extraction strategies for the English language has been presented during the last few years. However, as far as we know, no comprehensive study has been done on text chunking for Portuguese, showing its benefits. The scope of this work is the Portuguese language, and its objective is twofold. First, we analyze the impact of different chunk definitions, using a heuristic to generate chunks that relies on previous full parsing annotation. Then, we propose Machine Learning models for chunk extraction based on the Entropy Guided Transformation Learning technique. We employ the *Bosque* corpus, from the *Floresta Sintá(c)tica* project, for our experiments. Using golden values determined by our heuristic, a chunk feature improves the $F_{\beta=1}$ score of a clause identification system for Portuguese by 6.85 and the accuracy of a dependency parsing system by 1.54. Moreover, our best chunk extractor achieves a $F_{\beta=1}$ of 87.95 when automatic part-of-speech tags are applied. The empirical findings indicate that, indeed, chunk information derived by our heuristic is relevant to more elaborate tasks targeted on Portuguese. Furthermore, the effectiveness of our extractors is comparable to the state-of-the-art similars for English, taking into account that our proposed models are reasonably simple.

Keywords

Machine learning; natural language processing; text chunking; shallow parsing; entropy guided transformation learning.

Resumo

Ferreira, Guilherme Carlos De Napoli; Milidiú, Ruy Luiz. **Uma Abordagem de Aprendizado de Máquina para Segmentação Textual no Português**. Rio de Janeiro, 2011. 63p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

A segmentação textual é uma tarefa de Processamento de Linguagem Natural muito relevante, e consiste na divisão de uma sentença em sequências disjuntas de palavras sintaticamente relacionadas. Um dos fatores que contribuem fortemente para sua importância é que seus resultados são usados como significativos dados de entrada para problemas linguísticos mais complexos. Dentre esses problemas estão a análise sintática completa, a identificação de orações, a análise sintática de dependência, a identificação de papéis semânticos e a tradução automática. Em particular, abordagens de Aprendizado de Máquina para estas tarefas beneficiam-se intensamente com o uso de um atributo de segmentos textuais. Um número respeitável de eficazes estratégias de extração de segmentos para o inglês foi apresentado ao longo dos últimos anos. No entanto, até onde podemos determinar, nenhum estudo abrangente foi feito sobre a segmentação textual para o português, de modo a demonstrar seus benefícios. O escopo deste trabalho é a língua portuguesa, e seus objetivos são dois. Primeiramente, analisamos o impacto de diferentes definições de segmentação, utilizando uma heurística para gerar segmentos que depende de uma análise sintática completa previamente anotada. Em seguida, propomos modelos de Aprendizado de Máquina para a extração de segmentos textuais baseados na técnica Aprendizado de Transformações Guiado por Entropia. Fazemos uso do *corpus* Bosque, do projeto Floresta Sintá(c)tica, nos nossos experimentos. Utilizando os valores determinados diretamente por nossa heurística, um atributo de segmentos textuais aumenta a métrica $F_{\beta=1}$ de um sistema de identificação de orações para o português em 6.85 e a acurácia de um sistema de análise sintática de dependência em 1.54. Ademais, nosso melhor extrator de segmentos apresenta um $F_{\beta=1}$ de 87.95 usando anotações automáticas de categoria gramatical. As descobertas indicam que, de fato, a informação de segmentação textual derivada por nossa heurística é relevante para tarefas mais elaboradas cujo foco é o português. Além disso, a eficácia de nossos extratores é comparável à dos similares do estado-da-arte para o inglês, tendo em vista que os modelos propostos são razoavelmente simples.

Palavras–chave

Aprendizado de máquina; processamento de linguagem natural; segmentação textual; análise sintática superficial; aprendizado de transformações guiado por entropia.

Contents

List of Figures

List of Tables