Carlos Eduardo Meger Crestana

A Token Classification Approach to Dependency Parsing

DISSERTAÇÃO DE MESTRADO

DEPARTAMENTO DE INFORMÁTICA
Postgraduate Program in Informatics

Rio de Janeiro
March 2010

Pontifícia Universidade Católica
do Rio de Janeiro

Carlos Eduardo Meger Crestana

A Token Classification Approach to Dependency Parsing

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
March 2010

Pontifícia Universidade Católica
do Rio de Janeiro

Carlos Eduardo Meger Crestana

A Token Classification Approach to Dependency Parsing

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

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

Prof. Clarisse Sieckenius de Souza
Departamento de Informática – PUC-Rio

Prof. Raúl Rentería
Departamento de Informática – PUC-Rio

Prof. José Eugenio Leal
Coordenador Setorial do Centro Técnico Científico —
PUC–Rio

Rio de Janeiro, March 10^{th} 2010

Carlos Eduardo Meger Crestana

Graduated in Computer Engineering by the Pontifícia Universidade Católica do Rio de Janeiro. His research is focused on Machine Learning, Natural Language Processing and Information Extraction.

Bibliographic data

Crestana, Carlos

A Token Classification Approach to Dependency Parsing/ Carlos Eduardo Meger Crestana; advisor: Ruy Luiz Milidiú. — 2010.

66f.: il. (col.) ; 30cm

Dissertação (Mestrado em Informática) — Pontifícia Universidade Católica do Rio de Janeiro, Departamento de Informática, Rio de Janeiro, 2007.

Inclui bibliografia.

1. Informática — Teses. 2. Aprendizado de Máquina. 3.Processamento de Linguagem Natural.I. Milidiú, Ruy. II. Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática. III. Título.

CDD: 004

To my parents, Luís Fernando and Janete.

Acknowledgements

To God, for everything.

To my mother, my father and my family, for their continuous support.

To Thiago Araújo and Ricardo Costa, for their indispensable friendship and help since undergraduate studies.

To my *WhileTrue* friends: Eduardo Cardoso, Iam Jabour, Thuener Silva, Luana Lachtermacher, Felipe Nogueira, David Sotelo, Evelin Amorim and Daniel Fleischman, as well as Thiago Araújo and Ricardo Costa, for their friendship, the long talks, the infinite email threads and for making the beauty of computation even more explicit.

To my Advisor, Ruy Milidiú, for his encouragement, help and support during all my dissertation.

To Cícero dos Santos, Eraldo Fernandes and the LEARN team, for their talks, help and suggestions.

To CAPES and CNPq, for their financial support.

Abstract

Crestana, Carlos; Milidiú, Ruy (Advisor). **A Token Classification Approach to Dependency Parsing**. Rio de Janeiro, 2010. 66p. MSc. Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

One of the most important tasks in Natural Language Processing is *syntactic parsing*, where the structure of a sentence is inferred according to a given grammar. Syntactic parsing, thus, tells us how to determine the meaning of the sentence from the meaning of the words in it. Syntactic parsing based on dependency grammars is called dependency parsing. The Dependency-based syntactic parsing task consists in identifying a head word for each word in an input sentence. Hence, its output is a rooted tree, where the nodes are the words in the sentence. This simple, yet powerful, structure is used in a great variety of applications, like Question Answering, Machine Translation, Information Extraction and Semantic Role Labeling. State-of-the-art dependency parsing systems use transition-based or graph-based models. This dissertation presents a token classification approach to dependency parsing, by creating a special tagging set that helps to correctly find the head of a token. Using this tagging style, any classification algorithm can be trained to identify the syntactic head of each word in a sentence. In addition, this classification model treats projective and non-projective dependency graphs equally, avoiding pseudo-projective approaches. To evaluate its effectiveness, we apply the Entropy Guided Transformation Learning algorithm to the publicly available corpora from the CoNLL 2006 Shared Task. These computational experiments are performed on three *corpora* in different languages, namely: Danish, Dutch and Portuguese. We use the Unlabelled Attachment Score as the accuracy metric. Our results show that the generated models are above the average CoNLL system performance. Additionally, these findings also indicate that the token classification approach is a promising one.

Keywords

dependency parsing, machine learning, natural language processing, token classification, syntactic parsing

Resumo

Crestana, Carlos; Milidiú, Ruy. **Uma Abordagem Por Classificação Token-a-Token para o Parsing de Dependência**. Rio de Janeiro, 2010. 66p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Uma das tarefas mais importantes em Processamento de Linguagem Natural é a *análise sintática*, onde a estrutura de uma sentença é determinada de acordo com uma dada gramática, informando o significado de uma sentença a partir do significado das palavras nela contidas. A Análise Sintática baseada em Gramáticas de Dependência consiste em identificar para cada palavra a outra palavra na sentença que a governa. Assim, a saída de um analisador sintático de dependência é uma árvore onde os nós são as palavras da sentença. Esta estrutura simples, mas rica, é utilizada em uma grande variedade de aplicações, entre elas Sistemas de Pergunta-Resposta, Tradução Automática, Extração de Informação, e Identificação de Papéis Semânticos. Os sistemas estado-da-arte em análise sintática de dependência utilizam modelos baseados em transições ou modelos baseados em grafos. Essa dissertação apresenta uma abordagem por classificação *token-a-token* para a análise sintática de dependência ao criar um conjunto especial de classes que permitem a correta identificação de uma palavra na sentença. Usando esse conjunto de classes, qualquer algoritmo de classificação pode ser treinado para identificar corretamente a palavra governante de cada palavra na sentença. Além disso, este conjunto de classes permite tratar igualmente relações de dependência projetivas e não-projetivas, evitando abordagens pseudo-projetivas. Para avaliar a sua eficácia, aplicamos o algoritmo *Entropy Guided Transformation Learning* aos *corpora* disponibilizados publicamente na tarefa proposta durante a CoNLL 2006. Esses experimentos foram realizados em três *corpora* de diferentes idiomas: dinamarquês, holandês e português. Para avaliação de desempenho foi utilizada a métrica de *Unlabeled Attachment Score*. Nossos resultados mostram que os modelos gerados atingem resultados acima da média dos sistemas do CoNLL. Ainda, nossos resultados indicam que a abordagem por classificação *token-a-token* é uma abordagem promissora para o problema de análise sintática de dependência.

Palavras–chave

análise sintática de dependência, processamento de linguagem natural, aprendizado de máquina, classificação token-a-token

Contents

List of Figures

List of Tables