Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Larronda Asti

Anotador Morfossintático para o Português-Twitter

Dissertação de Mestrado

Dissertação apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do grau de Mestre em Informática.

Orientador: Prof. Ruy Luiz Milidiú

Rio de Janeiro
Abril de 2011

Pontifícia Universidade Católica
do Rio de Janeiro

Pedro Larronda Asti

Anotador Morfossintático para o Português-Twitter

Dissertação apresentada ao Programa de Pós–graduação em Informática do Departamento de Informática do Centro Técnico Científico da PUC–Rio como requisito parcial para obtenção do grau de Mestre em Informática. Aprovada pela Comissão Examinadora abaixo assinada.

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

Prof. Bruno Feijó
Departamento de Informática – PUC-Rio

Profa. Maria Cláudia de Freitas
Departamento de Letras – PUC-Rio

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

Rio de Janeiro, 05 de Abril de 2011

Pedro Larronda Asti

Graduou–se em Engenharia de Computação na Pontifícia Universidade Católica do Rio de Janeiro.

Ficha Catalográfica

Asti, Pedro Larronda

Anotador morfossintático para o português-twitter / Pedro Larronda Asti; orientador: Ruy Luiz Milidiú. — Rio de Janeiro : PUC–Rio, Departamento de Informática, 2011.

v., 49 f: il. ; 29,7 cm

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

Inclui referências bibliográficas.

1. Informática – Dissertação; 2. Twitter; 3. Português-Twitter; 4. POS; 5. Part-Of-Speech; 6. Anotador de POS; 7. Anotador de Part-Of-Speech; 8. Anotador Morfossintático; 9. Processamento de Linguagem Natural; 10. Aprendizado de Máquina; 11. ETL. I Milidiú, Ruy Luiz; II Pontifícia Universidade Católica do Rio de Janeiro. Departamento de Informática; III Título.

CDD: 004

Agradecimentos

Primeiramente, agradeço aos meus pais por tudo o que me ensinaram e me proporcionaram, que me levou a ser quem sou hoje.

Agradeço a minha família, a minha namorada Renata e aos meus amigos por seu apoio e sua compreensão com a minha ausência.

Agradeço ao Hugo Roenick por seu apoio nessa reta final e ao Allan Valeriano por ter enfrentado esse desafio comigo.

Agradeço a CAPES por me possibilitar fazer esse mestrado.

Por sua ajuda e ensinamentos, agradeço ao meu orientador Ruy Luiz Milidiú. Agradeço também a Eraldo Fernandes, Carlos Crestana, Eduardo Motta e demais membros do laboratório LEARN por sua ajuda e sugestões.

Agradeço a Cristina Ururahy, Carlos Cassino e o laboratório Tecgraf por sua paciência, apoio e compreensão.

Agradeço a linguista Maria Cláudia de Freitas por me ajudar a definir as classes dos elementos do Twitter e ao linguista Tiago da Silva Ribeiro por me ajudar na revisão do corpus.

Resumo

Asti, Pedro Larronda; Milidiú, Ruy Luiz. **Anotador Morfossintático para o Português-Twitter**. Rio de Janeiro, 2011. 49p. Dissertação de Mestrado — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

Nesta dissertação, apresentamos um processador linguístico que resolve a tarefa de Anotação morfossintática de mensagens em português postadas no Twitter. Ao analisar as mensagens escritas por brasileiros no Twitter, é fácil verificar que novos caracteres são introduzidos no alfabeto e também que novas palavras são adicionadas ao idioma. Além disso, observamos que essas mensagens são sintaticamente mal formadas. Isto impossibilita o uso nessas mensagens de diversos processadores linguísticos existentes para o português. Resolvemos esse problema considerando essas mensagens como escritas em uma nova língua, o português-twitter. O alfabeto dessa nova língua contém o alfabeto do português e o seu vocabulário contém o vocabulário da língua portuguesa. Porém, suas gramáticas são diferentes. Para construir os processadores desta nova linguagem, utilizamos a técnica de aprendizado supervisionado denominada Entropy Guided Transformation Learning (ETL). Adicionalmente, para treinar os processadores ETL, construímos um corpus anotado de mensagens em português-twitter. Não temos conhecimento da existência de outros Anotadores Morfossintáticos para o português-twitter. Porém, sabemos que, no estado-da-arte da Anotação Morfossintática para o português, a acurácia é de aproximadamente 96%, variando de acordo com o conjunto de classes escolhido. Construímos o processador composto de dois estágios, um morfológico e um contextual. Como métrica de avaliação, adotamos a acurácia, que mede quantos por cento do corpus foi anotado corretamente. Nossos resultados experimentais apresentam uma acurácia de 90,24% para o anotador proposto. Isto corresponde a um aprendizado significativo, pois o sistema inicial tem uma acurácia de apenas 76,58%. Este resultado é compatível com o aprendizado observado nos correspondentes processadores na língua portuguesa.

Palavras–chave

Twitter; Português-Twitter; POS; Part-Of-Speech; Anotador de POS; Anotador de Part-Of-Speech; Anotador Morfossintático; Processamento de Linguagem Natural; Aprendizado de Máquina; ETL.

Abstract

Asti, Pedro Larronda; Milidiú, Ruy Luiz (Advisor). **Morphosyntactic Tagger for Portuguese-Twitter**. Rio de Janeiro, 2011. 49p. MSc Dissertation — Departamento de Informática, Pontifícia Universidade Católica do Rio de Janeiro.

In this paper we present a language processor that solves the task of Morphosyntactic Tagging of messages posted in Portuguese on Twitter. By analyzing the messages written by Brazilian on Twitter, it is easy to notice that new characters are introduced in the alphabet and also that new words are added to the language. Furthermore, we note that these messages are syntactically malformed. This precludes the use of existing Portuguese processors in these messages, nevertheless this problem can be solved by considering these messages as written in a new language, the Portuguese-Twitter. Both the alphabet and the vocabulary of such idiom contain features of Portuguese. However, the grammar is are different. In order to build the processors for this new language, we have used a supervised learning technique known as Entropy Guided Transformation Learning (ETL). Additionally, to train ETL processors, we have built an annotated corpus of messages in Portuguese-Twitter. We are not aware of any other taggers for the Morphosyntactic Portuguese-Twitter task, thus we have compared our tagger to the the accuracy of state-of-art Morphosyntactic Annotation for Portuguese, which has accuracy around 96% depending on the tag set chosen. To assess the quality of the processor, we have used accuracy, which measures how many tokens were tagged correctly. Our experimental results show an accuracy of 90,24% for the proposed Morphosyntatic Tagger. This corresponds to significant learning, since the initial baseline system has an accuracy of only 76,58%. This finding is consistent with the observed learning for the corresponding regular Portuguese taggers.

Keywords

Twitter; Portuguese-Twitter; POS; Part-Of-Speech; POS Tagging; POS Tagger; Part-Of-Speech Tagging; Part-Of-Speech Tagger; Morphosyntactic Tagger; Natural Language Processing; Machine Learning; ETL; Entropy Guided Transformation Learning; Morphosyntactic Tagging.

Sumário

Lista de figuras

Lista de tabelas