



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



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

Todos os direitos reservados. É proibida a reprodução total ou parcial do trabalho sem autorização da universidade, do autor e do orientador.

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 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 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 MORPHOSYNTACTIC 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

1	Introdução	10
2	Motivação	12
2.1	Twitter	12
3	Tarefa	14
3.1	Classificação de palavras	14
3.2	ANOTADOR MORFOSSINTÁTICO	14
3.3	ANOTADOR MORFOSSINTÁTICO para o Português-Twitter	15
3.4	Português-Twitter	16
4	Algoritmos de Aprendizado	19
4.1	Classificador Bayesiano Ingênuo	19
4.2	Classificador Unigrama	20
4.3	Árvores de Decisão	20
4.4	Transformation-Based Learning	21
4.5	Entropy Guided Transformation Learning	23
5	Criando o Corpus Anotado	26
5.1	Criando o Corpus	26
5.2	Anotando os exemplos	31
5.3	Estatísticas do Corpus	38
6	Modelagem	40
6.1	O Modelo	40
6.2	Experimentos	41
7	Conclusão	43
8	Referências Bibliográficas	45
A	Português-Twitter	48
A.1	Contrações e Combinações	48

Lista de figuras

3.1	Frase com anotação morfossintática.	15
3.2	Anotação automática simples de um <i>tweet</i> .	16
3.3	Mensagem em português-twitter.	16
3.4	Exemplo de um <i>tweet</i> escrito em português e alguns de seus possíveis equivalentes em português-twitter.	17
4.1	Exemplo de árvore de decisão que classifica meios de transporte.	21
4.2	Exemplo de um gabarito de geração de regras de correção.	22
4.3	Treinamento do TBL.	22
4.4	Uma regra gerada a partir do gabarito da Figura 4.2.	23
4.5	Treinamento do ETL.	24
4.6	Geração dos templates.	24
4.7	Exemplo de janela no ETL.	25
5.1	Pipeline de geração do corpus.	26
5.2	Pipeline de anotação do corpus.	34
5.3	Pipeline do tradutor semi-automático.	36
5.4	Exemplo de tradução do português-twitter para o português.	36
5.5	Exemplo de mapeamento entre uma frase escrita em português-twitter e sua tradução para o português.	37
5.6	Exemplo de mapeamento entre <i>tweet</i> escrito em português-twitter e sua versão traduzida e anotada.	37
5.7	Transferindo o POS da versão traduzida para o <i>tweet</i> escrito em português-twitter.	37
5.8	Anotando elementos que faltam com sua categoria padrão.	38
5.9	Revisando e corrigindo o <i>tweet</i> escrito em português-twitter.	38

Lista de tabelas

5.1	Percentual de brasileiros no TWITTER de acordo com a faixa etária.	31
5.2	Etiquetas para anotação POS do português-twitter.	32
6.1	Acurácia do corpus de teste.	42
A.1	Contrações e combinações do português-twitter.	49