Instrumentos computacionais para a questão do acento no Português Brasileiro

Bruno Ferrari Guide

Orientador: Marcelo Barra

Contexto

- Acento no PB: Comportamento previsível em partes.
- Teorias: Bisol (1992) e Lee (1995)
- Pergunta 1: É possível criar um modelo que explique o comportamento do acento no PB sem se valer da marcação lexical?
- Pergunta 2: É possível identificar outras variáveis que tem relação com o comportamento do acento além de Categoria Morfossintática e Peso silábico?

Trabalhando com dados linguísticos

Linguística computacional -> Análise de grandes quantidades de dados linguísticos

O que implica na formatação de grande quantidade de dados linguísticos:

palavra -> [&pa-la-vra*, &pa-l1-vra*, &CV-CV-CCV*, nome, V, 1203, 203, 1000]

Corpus ABG - montagem

- Corpus Oral
  - C-oral do Brasil (UFMG)
  - Iboruna (UNESP – Rio Preto)
  - Projeto SP 2010(USP)
- Corpus Escrito
  - Textos Jornalísticos
  - Textos de Blogs
  - Textos Científicos

- Palavras foram transcritas, etiquetadas e acentuadas automaticamente e manualmente

Dados do Corpus ABG

Corpus Escrito		**Corpus Oral**	
Estadão	397.869	**C-oral**	12.079
Artigos	342.871	**Iboruna**	734.991
Blogs	215.126	**ProjetoSP2010**	1.216.728
Folha	819.381		
TOTAIS PARCIAIS			
Ocorrências	1.775.247	Ocorrências	1.963.798
Tipos	104.364	Tipos	40.586
TOTAL GERAL			
Ocorrências		3.739.045	
Tipos		123.245	

Análise Quantitativa

- Dados do corpus

CAT-ACENTUAL	% DE TYPES	% DE TOKENS
Monossílabas	2,39%	32,16%
Oxítonas	30,21%	27,31%
Paroxítonas	63,46%	38,89%
Proparoxítonas	3,90%	1,62%
4-sílabas	0,04%	0,03%
Total	**100,00%**	**100,00%**

N-Gramas

- *But it must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.*

N-Gramas

- *But it must be recognized that the notion “probability of a sentence” is an entirely useless one, under any known interpretation of this term.*

Noam Chomsky

N-Gramas: Explicação

Puramente Empírico.

N-gramas são sequências de tamanho N a que são atribuídas probabilidades.

Se extrai as frequências dos n-gramas do corpus de treinamento.

Aqui temos dois tipos de modelos:

Modelo Baseado em Tokens – ‘para’ > ‘paralelepípedo’

Modelo Baseado em Types – ‘para’ = ‘paralelepípedo’

N-Gramas: Acento

A probabilidade associada a uma atribuição de acento surge das cadeias de segmentos que a compõe:

Modelo de Trigramas (como uma cadeia de Markov):

P1(pa-l<u>a</u>) ≈ P(pa-|pa) x P(a-l|a-) x P(-l<u>a</u>|-l) x P(l<u>a</u>*|l<u>a</u>)

P2(p<u>a</u>-la) ≈ P(p<u>a</u>-|p<u>a</u>) x P(<u>a</u>-l|<u>a</u>-) x P(-la|-l) x P(la*|la)

Modelo = Argmax(P1,P2)

Classificador Bayesiano Ingênuo: Explicação

- Classificador probabilístico, pode conter conhecimento a priori, usado na área de aprendizado de máquina.
- Exige treinamento em um corpus para começar a funcionar.
- Classifica baseado na transformação do que se quer classificar em um vetor de características.
- Ingênuo pois assume que não existe nenhum tipo de relação entre as variáveis.

Classificador Bayesiano Ingênuo: Acento

- Uma palavra *w* é representada pelo vetor de traços *v*

palavra -> [&pa-la-vra*, &CV-CV-CCV*, nome, V, 1203, 203, 1000]

- A cada traço se atribui uma probabilidade em relação as classes:

  Nome -> Oxítona:0.3 | Paroxítona: 0.6 | Proparoxítona : 0.1

  V -> Oxítona: 0.1 | Paroxítona: 0.8 | Proparoxítona: 0.1

  ...

Classificador Bayesiano Ingênuo: Acento

- Com a regra de Bayes simplificada temos:
  - P (c|w) = P (w|c) * P(c)

  c = categoria, w = palavra

- Logo, podemos atribuir uma probabilidade da palavra pertencer a uma categoria levando em conta o vetor de traços que a compõe.

Conclusão

- Rodar aplicações baseadas nos modelos descritos.
- Coletar e analisar resultados.
- Testar novas versões do modelo CBI usando outros conjuntos de variáveis.

Referências

- Corpus:
  - C-oral
    - Raso, Tommaso, and Heliana Mello, eds. *C-oral-Brasil: corpus de referência do português brasileiro falado informal. I*. 2012.
  - Projeto SP
    - Mendes, Ronald Beline, and Lívia Oushiro. "O paulistano no mapa sociolinguístico brasileiro." *ALFA: Revista de Linguística* 56.3 (2012).
  - Iboruna
    - GONÇALVES, SCL. "Banco de dados Iboruna: amostras eletrônicas do português falado no interior paulista." *São José do Rio Preto:[sn]* (2007).
- Linguistica computacional e acento:
  - BIRD, S., KLEIN, E. and LOPER, E. (2009). *Natural Language Processing with Python.* Sebastopol, CA: O'Reilly.
  - JURAFSKY, D. and MARTIN, J. (2008). *Speech and Language Processing*. Upper Saddle River, NJ: Prentice Hall.
  - Bisol, Leda. "O acento e o pé métrico binário." *Cadernos de estudos lingüísticos* 22 (2012).
  - LEE, S.H (1995) – "Morfologia e Fonologia lexical do Português Brasileiro" – Tese de Doutorado – UNICAMP

Muito Obrigado!

bruno.fguide@gmail.com