Modelo de N-gramas para a questão do acento no Português Brasileiro

Bruno Ferrari Guide

Orientador: Marcelo Barra Ferreira
Departamento de Linguística - FFLCH - USP

17 de dezembro de 2015

Introdução

- Tópicos dessa apresentação:
  - Objetivos
  - Sobre o acento
  - Modelos
  - N-gramas
  - Perspectivas

Objetivos

Objetivo Geral

- Implementar modelos computacionais capazes de atribuir probabilidades às diferentes versões acentuadas de uma palavra.

Objetivos Específicos

- Traçar um perfil detalhado do acento no Português Brasileiro (PB).
- Implementar e testar algoritmos baseados nas soluções teóricas estudadas.
- Aplicar ferramentas computacionais disponíveis para desenvolver abordagens ao problema.

Sobre o acento

Sobre o acento

- O acento no PB pode recair nas últimas três sílabas da palavra, a partir disso temos as três categorias acentuais: oxítonas, paroxítonas e proparoxítonas.
- O acento é regular, porém tem irregularidades.
- O acento é irregular, porém tem regularidades.

Panorama do acento no PB

Categoria Acentual	Ocorrências	%	Tipos	%
Oxítona	985.587	40,20%	28.037	30,50%
Paroxítona	1.407.433	57,41%	60.186	65,48%
Proparoxítona	58.447	2,38%	3.694	4,02%
Total	**2.451.467**	**100%**	**91.917**	**100%**

Comportamento do acento

Padrão Previsível	Padrão Imprevisível
Verbos – todos	Não-verbos não-derivados
Não– Verbos derivados morfologicamente	Proparoxítonas
Oxítonas terminadas em consoante	Oxítonas terminados em vogal
Paroxítonas terminadas em vogal	Paroxítonas terminadas em consoantes

Teorias sobre a questão

- Abordagem Lexicalista - Câmara (1975)
- Abordagens Métricas:
  - Bisol (1992)
  - Lee (1995)
- Existe sempre em alguma medida arbitrariedade no comportamento do acento.

Modelos

Definição de modelo

- ***model*** *(n): a miniature representation of something; a pattern of something to be made; an example for imitation or emulation; a description or analogy used to help visualize something (e.g., an atom) that cannot be directly observed; a system of postulates, data and inferences presented as a mathematical description of an entity or state of affairs.*
- ***mathematical model*** *(n): a representation in mathematical terms of the behavior of real devices and objects*
- Comparar modelos distintos não é uma tarefa trivial.

Modelos Probabilísticos

- A ideia é que modelos probabilísticos podem dar um tratamento mais sensível às irregularidades do comportamento do acento do que modelos categóricos que necessariamente exigem algum tipo de marcação arbitrária para cobrir os dados.
- Algum objeto investigado de comportamento mais regular não exigiria um modelo probabilístico para formalizar seu funcionamento, como é o caso da silabificação. (Guide, 2013)

N-gramas

N-Gramas: Ponto de Partida

- Esse modelo parte do pressuposto que o objeto de estudo é uma sequência de eventos aleatórios. (O modelo de língua subjacente é, portanto, muito simples.)
- Além disso, a ideia é que se pode atribuir à um desses eventos a sua probabilidade a partir da probabilidade condicional não da sequência inteira da qual ele faz parte, mas somente da sequência composta pelos N elementos que o antecedem. (Cadeia de Markov de tamanho N).

$$P(X_{i+1}=x|\ X_1=x_1,X_2=x_2,\ldots,X_i=x_i)=P(X_{i+1}=x|\ X_i=x_i\ \ldots\ X_{i-N}=x_{i-N})$$

Funcionamento do modelo para a questão do acento

- Se atribui uma probabilidade a partir do modelo para cada uma das versões acentuadas de uma palavra a partir das frequências extraídas do Corpus.
- A palavra é quebrada nos n-gramas que a compõe e a probabilidade final é o produto das probabilidades dos n-gramas que a compõe.
- Ex: (num modelo de bi-gramas)
  - P(&**es**trela*) = P(**e**|&)* P(**s**|**e**)* P(t|**s**)...
  - P(&es**tre**la*) = P(e|&)* P(s|e)* P(t|s)...
  - P(&estre**la**) = P(e|&)* P(s|e)* P(t|s)...
    a partir da noção de que:
  - $P(e|\&) = \frac{C(\&e)}{C(\&)}$

N-Gramas: Implementação e questões computacionais

- Feita em Python
- Qual o tamanho de N?
  - Quanto maior N, mais informativo é o modelo.
  - Quanto maior N, mais esparsos os dados, maior o corpus necessário para conseguir trabalhar. (explosão exponencial)

N-Gramas: Implementação e questões linguísticas

- O que é considerado uma palavra?
- O modelo de língua vai dar mais peso para palavras mais frequentes? Ou todas serão consideradas iguais?
- Palavras com menos de 3 sílabas não acabam criando um enviesamento contra as proparoxítonas?

A partir desses questionamentos, uma série de modelos de N-gramas distintos é proposta.

Modelos de N-gramas Propostos

- Modelo de Bigramas baseado em tipos
- Modelo de Bigramas baseado em ocorrências
- Modelo de Trigramas baseado em tipos
- Modelo de Trigramas baseado em ocorrências

Cada um dos modelos ainda vai ser testado também levando em conta uma versão do Corpus SEM as palavras com menos de duas sílabas.
Além disso, também é interessante testar modelos que considerem a palavra não uma sequência de sons, mas sim de sílabas.

Perspectivas

Modelo de N-gramas: Resultados

- Treinar os modelos no corpus de treino e depois testá-los.
- Tabular os desempenhos dos modelos distintos e analisar os resultados.

Modelo do Classificador Bayesiano Ingênuo

- Este modelo será a análise probabilística proposta mais completa.
- Implementações comuns desse modelo estão sendo estudadas.
- O script do Classificador irá atribuir probabilidades a possíveis formas acentuadas de uma palavra baseado na relação entre todos os diversos *features* representados no corpus com as categorias acentuais.
- A partir disso, análises de desempenho irão promover novas versões do classificador, considerando um conjunto de variáveis distinto (por exemplo, excluindo a frequência da palavra como *feature*)

Obrigado!