Abordagem computacional para a questão do acento no português brasileiro

III WORKSHOP DE LINGUÍSTICA COMPUTACIONAL

Bruno Ferrari Guide

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

8 de dezembro de 2016

Introdução

- Tópicos dessa apresentação:
  - Objetivos
  - Resumo
  - Resultados

Objetivos do projeto

- Revisar abordagens para a questão do acento.
- A partir de dados coletados do idioma, apresentar uma análise quantitativa sobre o comportamento do acento em relação a algumas variáveis linguísticas.
- Implementar modelos computacionais preditivos do acento para fundamentar uma discussão sobre o tema.

Sobre o Acento - 2 tendências

- O acento no português brasileiro (quase) sempre ocupa uma das três últimas posições da palavra, criando as três categorias acentuais: Oxítona, Paroxítona, Proparoxítona.
- Duas tendências dão conta da maioria das palavras do PB:
  - Caso a sílaba final seja pesada, a palavra é oxítona.
  - Caso a sílaba final seja leve, a palavra é paroxítona.

2 tendências?

- Problemas com palavras oxítonas terminadas em sílaba leve, como *caqui, urubu*
- Problemas com paroxítonas terminadas em sílaba pesada, como em *mártir, câncer, difícil*.
- Problemas com as proparoxítonas de modo geral.
- O acento é regular, porém tem irregularidades.
- O acento é irregular, porém tem regularidades.

Sobre modelos probabilísticos

- Modelo é uma representação formal de um objeto.
- As vezes, o objeto possui comportamento imprevisível.
- Na matemática, a área que lida com a imprevisibilidade (ou seja, que a quantifica e formaliza) é a probabilidade.
- Existem muitas formas de tentar formalizar essa imprevisibilidade, cada uma possui suas vantagens e desvantagens.

Modelos implementados e Propostas analisadas

- Propostas analisadas:
  - Bisol (1992)
  - Lee (1995)
- Modelos probabilísticos:
  - N-gramas
  - Classificador Bayesiano Ingênuo
- Modelo arbitrário:
  - *Baseline*

Análise das propostas e definição de baseline

- Bisol (1992)
  - Se a palavra termina em sílaba pesada, é oxítona.
  - Senão, é paroxítona.
- Lee (1995)
  - Não-verbos: Oxítonas com sílaba final leve ou pesada são previstas. Paroxítonas com sílaba final leve também.
  - Verbos: Paroxítonas terminadas em sílaba leve e oxítonas terminadas em sílaba pesada são previstas.
- *Baseline*
  - Toda palavra é paroxítona.

Exemplo: N-gramas

- Palavra: 'casa'
- Representação: '&ka-za*'
- Probabilidade obtida a partir do cálculo:
  - BIGRAMAS P('&ka-za*') = P('&k') $\times$ P('ka') $\times$ P('a-') $\times$ P('-z') $\times$ P('za') $\times$ P('a*')
  - TRIGRAMAS P('&ka-za*') = P('&ka') $\times$ P('ka-') $\times$ P('a-z') $\times$ P('-za') $\times$ P('za*')

Exemplo: Classificador Bayesiano Ingênuo

- Palavra: 'perna'
- Vetor de traços: [Categoria morfossintática: 'Nome', estrutura silábica: 'CVC-CV', nivel de frequência no corpus: 3]
- Probabilidades:
  - Oxítona = p(oxítona)* p('nome'|oxítona) * p('leve'|oxítona)...
  - Paroxítona = p(paroxítona)* p('nome'|paroxítona) * p('leve'|paroxítona)...
  - Proparoxítona = p(proparoxítona)* p('nome'|proparoxítona) * p('leve'|proparoxítona)...
- Repare que os meus *priors* vão favorecer as categorias paroxítona, oxítona e proparoxítona nessa ordem.

Implementação

- Corpus utilizado: ABG $\rightarrow$ 98.000 palavras.
- Implementação feita em Python.
- Foi efeituada uma *Cross-validation* dos resultados.
- Acertos e erros em comparação com acentuação categórica já presente no corpus.

Resultados

Desempenho do modelo baseado em Bisol(1992) para os verbos

Bisol (1992)	% de tipos	% de ocorrências	Acerto
Verbos	48,60	33,12	
Oxítonas leves	15,31	23,20	Não
Oxítonas pesadas	24,99	32,97	Sim
Paroxítonas leves	58,10	42,48	Sim
Paroxítonas pesadas	1,26	1,27	Não
Proparoxítonas leves	0,33	0,06	Não
Proparoxítonas pesadas	0,00	0,01	Não
% Acerto verbos	83,09	75,45	
% do total	40,38	24,99	

Desempenho do modelo baseado em Lee (1995) para os verbos

Lee (1995)	% de tipos	% de ocorrências	Acerto
Verbos	48,60	33,12	
Oxítonas leves	15,31	23,20	Não
Oxítonas pesadas	24,99	32,97	Não
Paroxítonas leves	58,10	42,48	Sim
Paroxítonas pesadas	1,26	1,27	Sim
Proparoxítonas leves	0,33	0,06	Não
Proparoxítonas pesadas	0,00	0,01	Não
% Acerto verbos	59,36	43,75	
% do total	28,85	14,49	

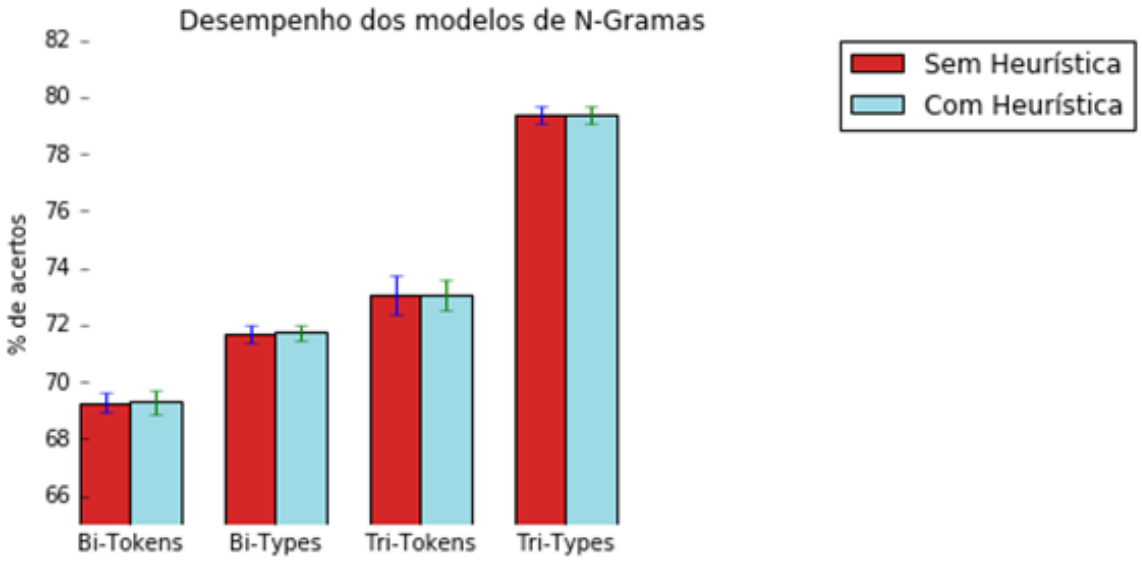
Desempenho do modelo baseado em Bisol (1992) para os não-verbos

Bisol (1992)	**% de tipos**	**% de ocorrências**	**Acerto**
Não-verbos	51,40	66,88	
Oxítonas leves	10,01	4,78	Não
Oxítonas pesadas	8,76	23,48	Sim
Paroxítonas leves	68,51	66,25	Sim
Paroxítonas pesadas	4,98	1,41	Não
Proparoxítonas leves	7,46	3,64	Não
Proparoxítonas pesadas	0,28	0,44	Não
% Acerto não-verbos	77,27	89,73	
% do total	39,72	60,01	
% Acerto total	**80,10**	**85,00**	

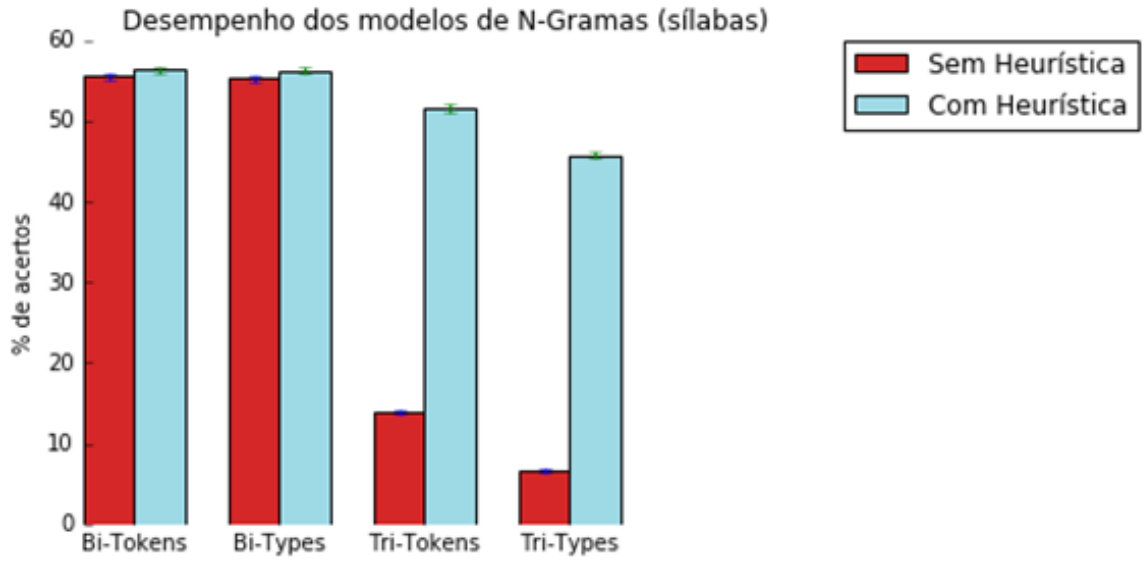
Desempenho do modelo baseado em Lee (1995) para os não-verbos

Lee (1995)	**% de tipos**	**% de ocorrências**	**Acerto**
Não-verbos	51,40	66,88	
Oxítonas leves	10,01	4,78	Sim
Oxítonas pesadas	8,76	23,48	Sim
Paroxítonas leves	68,51	66,25	Sim
Paroxítonas pesadas	4,98	1,41	Não
Proparoxítonas leves	7,46	3,64	Não
Proparoxítonas pesadas	0,28	0,44	Não
% Acerto não-verbos	87,28	94,51	
% do total	44,86	63,21	
% Acerto total	**73,71**	**77,70**	

Resultados: N-gramas



Resultados: N-gramas Sílabas



Resultados: CBI

- N = Nível de frequência
- C = Categoria Morfossintática
- E = Estrutura silábica (CV-CV)

Tabela: CBI

Tabela: Desempenho dos classificadores bayesianos ingênuos montados a partir de diferentes subconjuntos de variáveis.

Nome	**Média**	**Desvio**
E	91,93	0,20
CE	90,18	0,22
EN	90,09	0,22
CEN	86,45	0,26
N	66,64	0,36
CN	66,62	0,35
C	66,59	0,33

Resultados: Geral

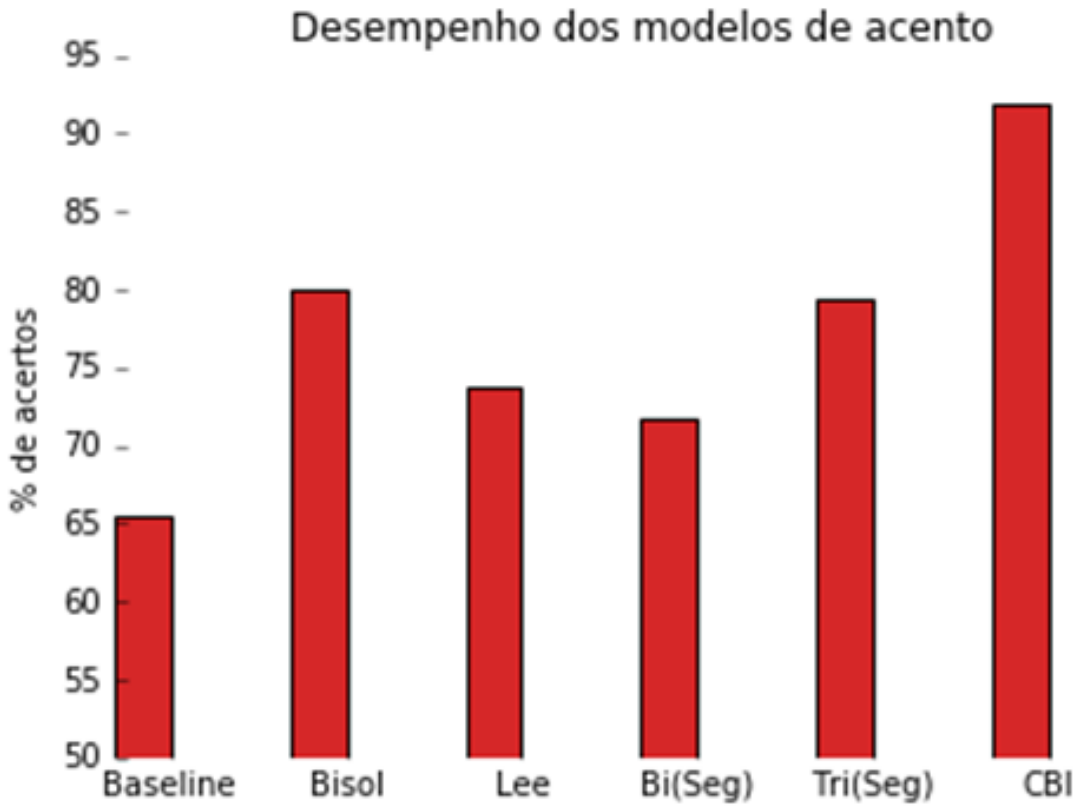


Tabela: Desempenho geral

Tabela: Desempenho Geral

Parâmetros	Média
CBI E	91.93
Bisol	80.10
Trigramas	79.40
Lee	73.71
Bigramas	71.69
Baseline	65.59

Conclusões

- Após as análises feitas, a marcação arbitrária para descrever o comportamento do acento no PB não pôde ser descartada. Mas o funcionamento da marcação arbitrária para a previsão do acento não é uma questão trivial.
- O peso da sílaba final é a variável linguística estudada que mais possui relação estatística com o comportamento do acento.
- A melhoria do modelo baseado em trigramas quando comparado com o modelo de bigramas sugere que há motivos para se conduzir uma nova análise usando modelos baseados em tetragramas.
- O CBI que se vale da probabilidade da estrutura silábica da palavra pertencer a uma categoria se mostrou bastante eficiente para prever o acento, ainda que não tenha sido capaz de capturar o comportamento que gera as palavras proparoxítonas.

Muito Obrigado!