N-gramas e o acento no Português Brasileiro

Bruno Ferrari Guide

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

12 de abril de 2015

Introdução

- Tópicos dessa apresentação:
  - Objetivos
  - Sobre o acento
  - Modelos de N-Gramas
  - Resultados
  - Perspectivas

Objetivos

- A partir da criação de modelos probabilísticos, eu pretendo apresentar uma discussão sobre o comportamento do acento no PB.
- Os modelos são baseados em corpus e podem trazer à tona algumas características quantitativas sobre esse comportamento.
- Os modelos retornam as probabilidades de uma determinada palavra pertencer a alguma categoria acentual (Oxítona, Paroxítona, Proparoxítona) e a partir disso é possível discutir os erros e os acertos de um modelo.

Sobre o acento

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.

Modelos de N-Gramas

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 de N-Gramas: princípios

- A noção de língua subjacente é simples.
- Poderíamos atribuir probabilidade levando em conta a palavra inteira.
- Mas a ideia é que se pode atribuir à sequência a sua probabilidade a partir da probabilidade condicional não dela inteira, mas somente da sequências de tamanho N que a compõe. (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:
  - $P(e|\&) = \frac{C(\&e)}{C(\&)}$

Modelos de N-Gramas: Entrada

- O modelo recebe como entrada uma palavra transcrita.
- A transcrição foi feita automaticamente.
- O corpus utilizado é o Corpus ABG, compilado em conjunto com a colega Aline Benevides.
- As palavras monossílabas foram descartadas para essa pesquisa.
- Com isso, o corpus tem cerca de 95 mil tipos de palavras diferentes.

Diferentes Modelos: Tamanho de N

- Quanto maior a cadeia de n-gramas, mais complexo e informativo é o modelo.
- Quanto maior a cadeia, maior o número de possíveis n-gramas:

$$C = \alpha^n$$

- Com mais possibilidades, é necessário um Corpus cada vez maior para o modelo funcionar.
- Por isso, os tamanhos de n escolhidos foram 2 (bi-gramas) e 3 (tri-gramas).

Diferentes Modelos: Tipos e Ocorrências

- O modelo baseado em tipos contabiliza toda palavra uma única vez.
- O modelo baseado em ocorrências inclui as repetições de uma mesma palavra.
- O corpus ABG já contem esse tipo de informação.
- No entanto, uma vez que foram eliminadas as palavras monossílabas, o comportamento do acento em termos quantitativos tanto nos tipos quanto nas ocorrências ficou bastante similar.

Diferentes Modelos: Conhecimento embutido

- Caso a probabilidade atribuída pelo modelo de n-gramas resulte em um empate entre duas palavras candidatas, qual medida tomar?
- Nos modelos sem heurística, tal situação foi considerada um erro.
- Nos modelos com heurística, o modelo escolhia uma das candidatas empatadas. No caso, a que pertencesse a uma das categorias mais comum.
- (Paroxítona > Oxítona > Proparoxítona)

Modelos de N-Gramas e a linguistica: Bigramas

r

	tokens		types
-r	238189	-r	15433
r-	128499	r-	8181
ra	125610	ra	5914
r*	120046	rE	4218
1r	69086	r*	4215

p

&p	215256	-p	8563
-p	110325	&p	6715
pr	64697	pr	3347
pe	48205	pe	3140
pa	40335	pa	2360

Modelos de N-Gramas e a linguistica Trigramas

r

	tokens		types
-ra	84376	a-r	3587
rA*	77924	ra-	3448
1r*	60944	-ra	3231
&pr	48882	re-	2999
-tr	48243	e-r	2715

p

&pr	48882	pE-	2175
pE-	36883	&pr	2078
&p1	36340	-pE	1884
&pE	31120	pa-	1732
p1-	28943	A-p	1721

Diferentes Modelos: Segmento e Sílaba

- Montei uma versão do corpus só com palavras com 3 sílabas ou mais.
- O que alimenta o modelo não são os segmentos, e sim as sílabas.
- A palavra 'pedrada' seria representada em bigramas da seguinte maneira então: ['pe-dra','dra-da']
- Expectativa de que os resultados fossem bastante interessantes, dado que é um modelo muito informativo.

Resultados

Obtendo os resultados

A metodologia para a obtenção dos resultados foi a seguinte:

- O processo começa com a separação randômica do corpus entre corpus treino (80%) e corpus teste (20%).
- O corpus treino alimenta o modelo.
- O modelo atribui probabilidades às diferentes versões acentuadas das palavras do corpus teste.
- A probabilidade mais alta para cada candidata é escolhida, a partir disso se compara com a acentuação correta da palavra.
- O modelo acerta se a categoria que elegeu mais provável é a categoria da qual a palavra faz parte.
- Esse processo foi repetido 100 vezes para cada modelo.

Resultados N-Gramas

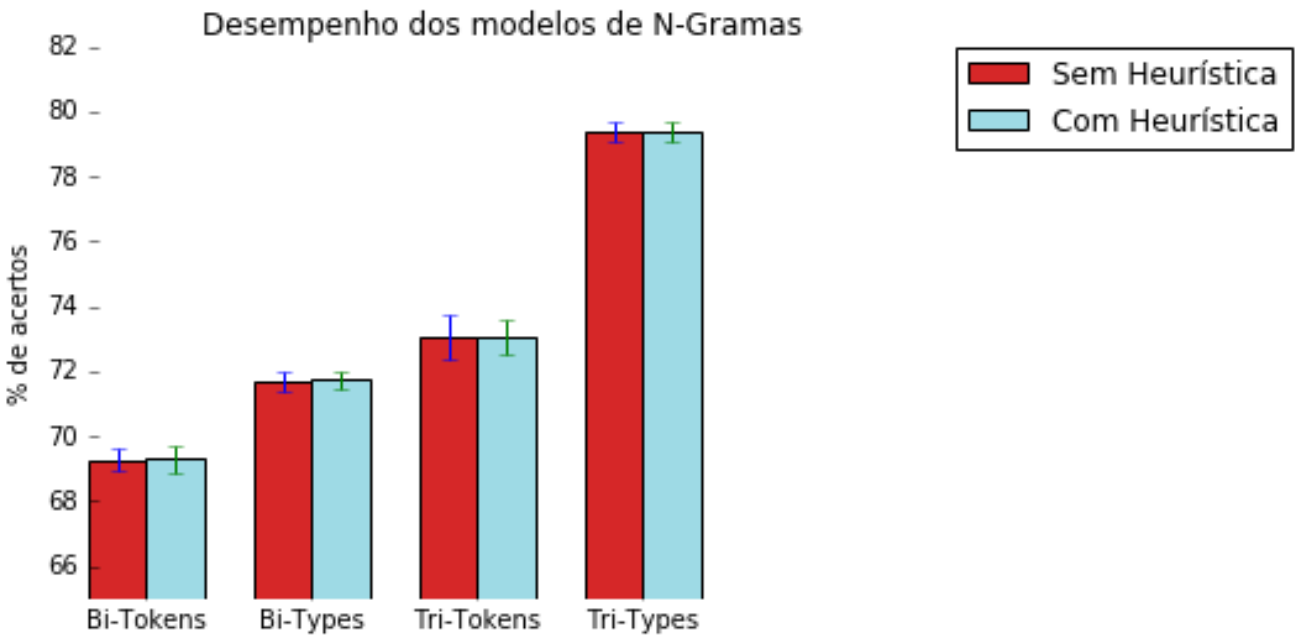

Figura: Desempenho dos modelos de N-gramas

Análise N-Gramas baseado em segmentos

Modelos		Media Acertos	Desvio Padrão
Bi-tok	Sem H	69.29	0.36
	Com H	69.30	0.43
Bi-typ	Sem H	71.69	0.30
	Com H	71.74	0.28
Tri-tok	Sem H	73.05	0.66
	Com H	73.05	0.54
Tri-typ	Sem H	79.40	0.30
	Com H	79.38	0.27

Resultados Sílabas

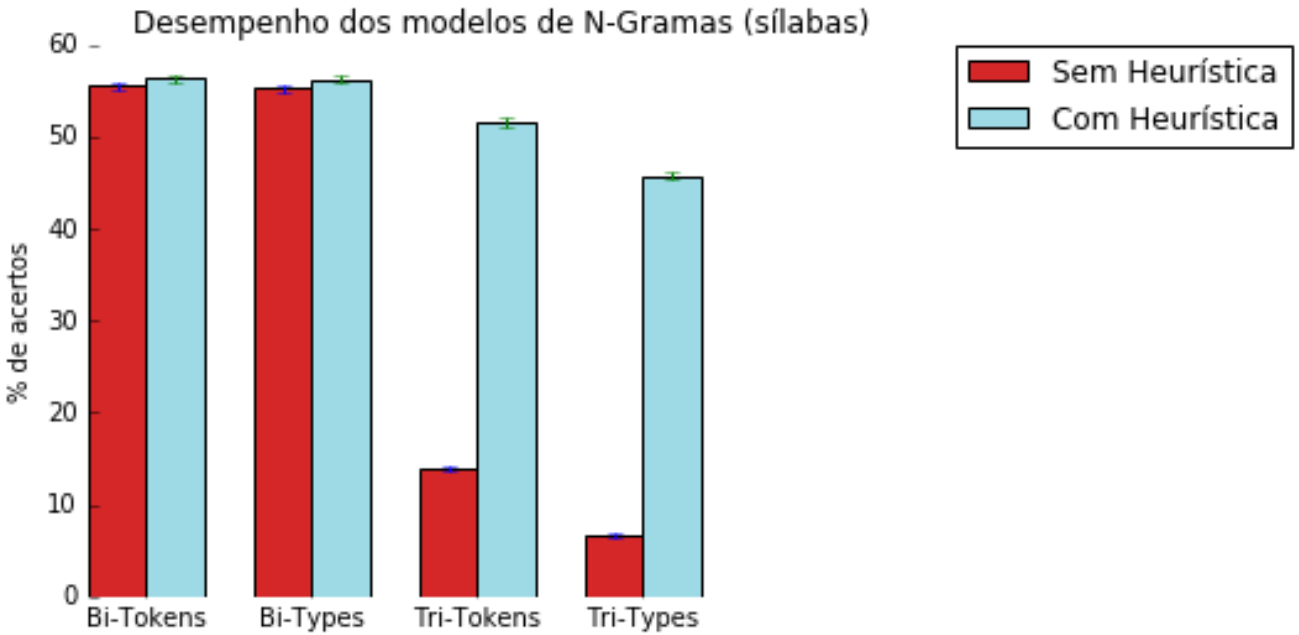

Figura: Desempenho dos modelos de N-gramas baseados em Sílaba.

Análise N-Gramas baseados em sílabas

Modelos		Média de Acertos	Desvio Padrão
Bigramas (Tokens)	Sem H	55.53	0.43
	Com H	56.31	0.39
Bigramas (Types)	Sem H	55.24	0.39
	Com H	56.22	0.35
Trigramas (Tokens)	Sem H	13.96	0.25
	Com H	51.54	0.42
Trigramas (Types)	Sem H	6.72	0.19
	Com H	45.75	0.42

Resultados Gerais

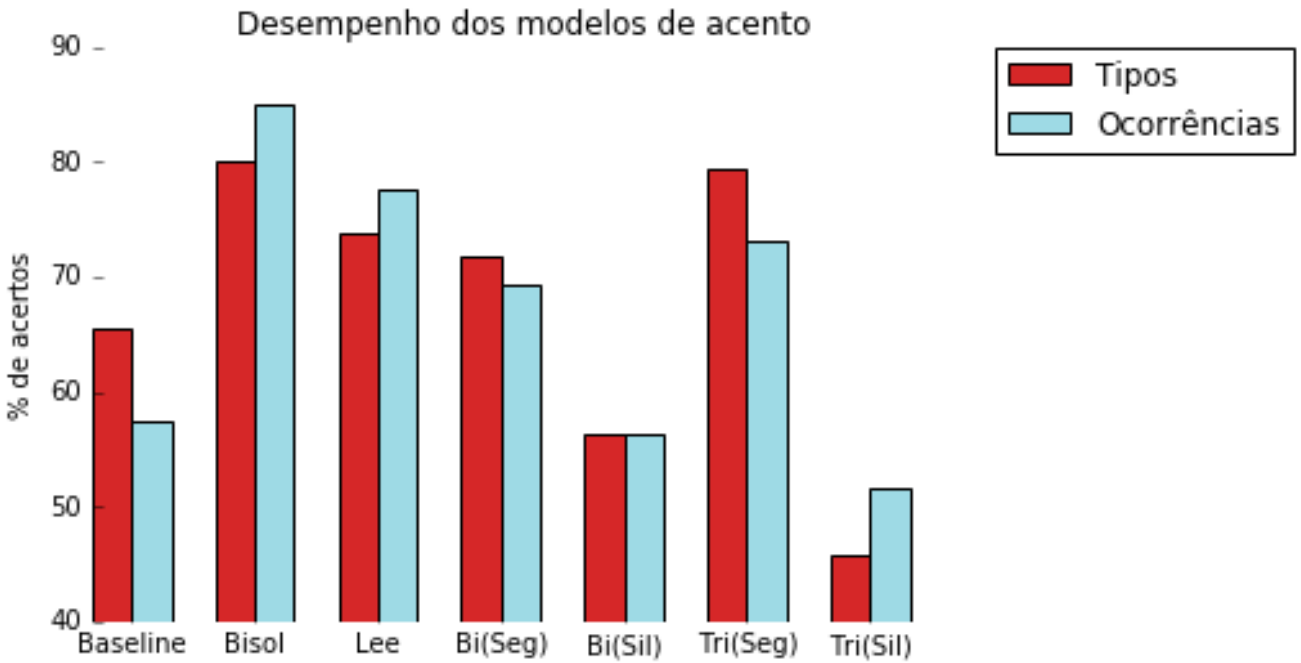

Figura: Desempenho dos modelos de N-gramas em comparação com os modelos baseados em Lee (1995), Bisol (1992) e o Baseline.

Próximos passos

- Implementar outro modelo probabilístico, que leva em conta dados linguísticos além da frequência dos segmentos. No caso, o modelo a ser implementado será o Classificador Bayesiano Ingênuo.
- Escrever, escrever, escrever...

Agradecimento

Muito Obrigado!