Processamento de Linguagem Natural

Correção Ortográfica

Prof.: Hansenclever Bassani (Hans) hfb@cin.ufpe.br
Site da disciplina: www.cin.ufpe.br/~hfb/pln/

Baseado nos slides do curso de Stanford no Coursera
por Daniel Jurafsky e Christopher Manning.

Tradução: Ygor Sousa
Revisão: Hansenclever Bassani

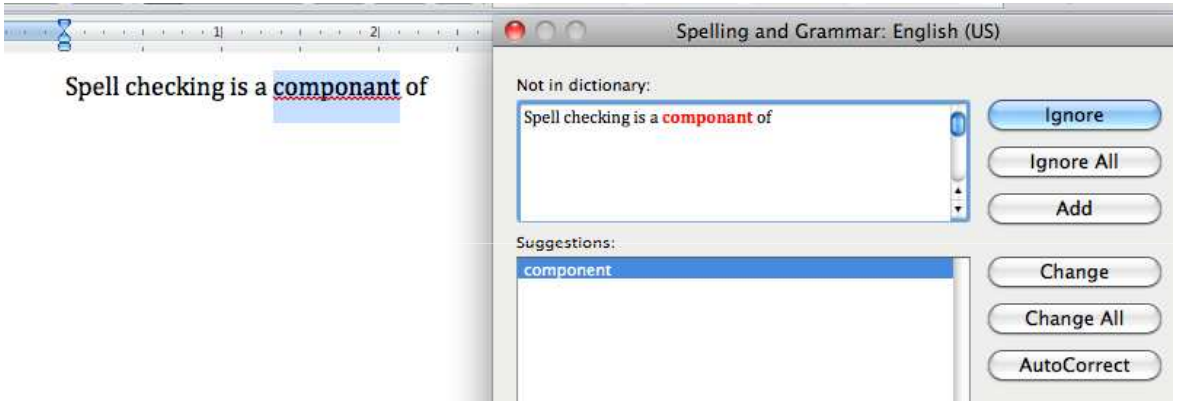
UNIVERSIDADE
FEDERAL
DE PERNAMBUCO

Correção Ortográfica e Canal com Ruído

A Tarefa de Correção Ortográfica

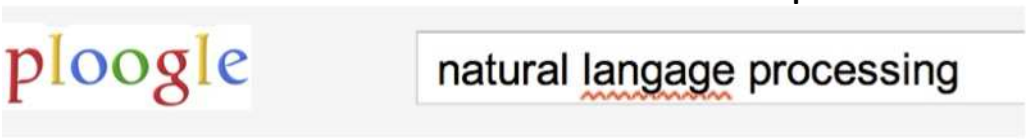
Aplicações para Correção Ortográfica

Processamento de Textos



Telefones

Pesquisa na Internet



Showing results for natural ***language*** processing
Search instead for natural langage processing

Tarefas de Ortografia

- Detecção de Erros de Ortografia
- Correção de Erros de Ortografia:
  - Correção automática
    - hte→the
  - Sugestão de correção
  - Listas de sugestões

Tipos de Erros de Ortografia

- Erros Non-word
  - *graffe →giraffe*
- Erros Real-word
  - Erros Tipográficos
    - *three →there*
  - Erros Cognitivos (homófono)
    - *piece→peace*,
    - *too → two*

Taxas de Erros Ortográficos

26%: Consultas na Web Wang *et al.* 2003

13%: Redigitação, sem backspace Whitelaw *et al.* English&German

7%: Palavras corrigidas redigitando em agendas eletrônicas

2%: Palavras não corrigidas em agendas eletrônicas Soukoreff &MacKenzie 2003

1-2%: Redigitação: Kane and Wobbrock 2007, Gruden et al. 1983

Erros de Ortografia Non-word

- Detecção de erros ortográficos Non-word:
  - Qualquer palavra que não esteja no ***dicionário*** é um erro
  - Quanto maior o dicionário melhor
- Correção de erros ortográficos Non-word:
  - Geração de ***candidatos***: palavras reais que são similares ao erro
  - Escolher o que é melhor:
    - Menor distância de edição ponderada
    - Maior probabilidade de canal com ruído

Erros de Ortografia Real Word

- Para cada palavra *w*, gerar um conjunto de candidatos:
  - Encontrar palavras candidatas com ***pronúncia*** similar
  - Encontrar palavras candidatas com ***ortografia*** similar
  - Incluir *w* no conjunto de candidatos
- Escolher o melhor candidato
  - Canal com Ruído
  - Classificador

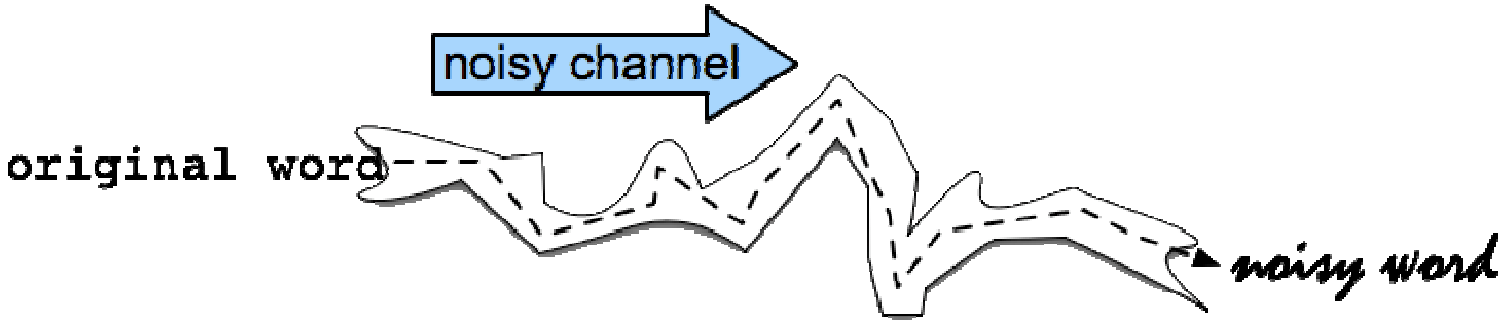
Correção Ortográfica e Canal com Ruído

A Tarefa de Correção Ortográfica

Correção Ortográfica e Canal com Ruído

Ortografia: Modelo de Canal com Ruído (Noisy Channel)

Canal com Ruído (Noisy Channel)



Canal com Ruído (Noisy Channel)

- Vemos uma observação x de uma palavra mal escrita
- Encontrar a palavra correta w

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(w \mid x)$$

$$= \underset{w \in V}{\operatorname{argmax}} \frac{P(x \mid w) P(w)}{P(x)}$$

$$= \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)$$

História: Noisy Channel para Ortografia proposto por volta de 1990

- **IBM**
  - Mays, Eric, Fred J. Damerau and Robert L. Mercer. 1991. Context based spelling correction. *Information Processing and Management*, 23(5), 517–522
- **AT&T Bell Labs**
  - Kernighan, Mark D., Kenneth W. Church, and William A. Gale. 1990. A spelling correction program based on a noisy channel model. Proceedings of COLING 1990, 205-210

Exemplo de Erro Ortográfico Non-word

acress

Geração de Candidato

- Palavras com ortografia similar
  - Distância pequena de edição ao erro
- Palavras com pronuncia similar
  - Distância pequena de edição de pronuncia ao erro

Distância de Edição Damerau-Levenshtein

- Distância mínima de edição entre duas strings, em que edições são:
  - Inserção
  - Remoção
  - Substituição
  - Transposição de duas letras adjacentes

Candidatos para `acress` com distância 1

Erro	Candidato de Correção	Letra Correta	Letra Errada	Tipo
acress	actress	t	-	remoção
acress	cress	-	a	inserção
acress	caress	ca	ac	transposição
acress	access	c	r	substituição
acress	across	o	e	substituição
acress	acres	-	s	inserção
acress	acres	-	s	inserção

Geração de Candidatos

- 80% dos erros tem distância de edição 1
- Quase todos os erros tem distância de edição 2

- Também permite inserção de **espaço** ou **hífen**
  - `thisidea` → `this idea`
  - `inlaw` → `in-law`

Modelo de Linguagem

- Usar qualquer um dos algoritmos de modelo de linguagem que vimos
- Unigram, bigram, trigram
- Correção Ortográfica Web-scale
  - Stupid backoff

Unigram: Probabilidade Prévia

Contagens de 404,253,213 palavras no Corpus of Contemporary English (COCA)

Palavra	Frequência da Palavra	P(palavra)
actress	9,321	.0000230573
cress	220	.0000005442
caress	686	.0000016969
access	37,038	.0000916207
across	120,844	.0002989314
acres	12,874	.0000318463

Probabilidade de Modelo de Canal

- **Error model probability, Edit probability**
- *Kernighan, Church, Gale 1990*

- *Palavra incorreta* $x = x_1, x_2, x_3 \ldots x_m$
- *Palavra correta* $w = w_1, w_2, w_3, \ldots, w_n$

- P(x|w) = probabilidade de edição
  - (remoção/inserção/substituição/transposição)

Calcular probabilidade de erro: matriz de confusão

```
del[x,y]:    contar (xy digitado como x)
ins[x,y]:    contar (x digitado como xy)
sub[x,y]:    contar (x digitado como y)
trans[x,y]:  contar (xy digitado como yx)
```

Inserção e remoção condicionada ao caractere anterior (poderia ser em relação ao posterior também).

Matriz de Confusão para Erros de Ortografia

sub[X, Y] = Substitution of X (incorrect) for Y (correct)

X \ Y (correct)	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
a	0	0	7	1	342	0	0	2	118	0	1	0	0	3	76	0	0	1	35	9	9	0	1	0	5	0
b	0	0	9	9	2	2	3	1	0	0	0	5	11	5	0	10	0	0	2	1	0	0	8	0	0	0
c	6	5	0	16	0	9	5	0	0	0	1	0	7	9	1	10	2	5	39	40	1	3	7	1	1	0
d	1	10	13	0	12	0	5	5	0	0	2	3	7	3	0	1	0	43	30	22	0	0	4	0	2	0
e	388	0	3	11	0	2	2	0	89	0	0	3	0	5	93	0	0	14	12	6	15	0	1	0	18	0
f	0	15	0	3	1	0	5	2	0	0	0	3	4	1	0	0	0	6	4	12	0	0	2	0	0	0
g	4	1	11	11	9	2	0	0	0	1	1	3	0	0	2	1	3	5	13	21	0	0	1	0	3	0
h	1	8	0	3	0	0	0	0	0	0	2	0	12	14	2	3	0	3	1	11	0	0	2	0	0	0
i	103	0	0	0	146	0	1	0	0	0	0	6	0	0	49	0	0	0	2	1	47	0	2	1	15	0
j	0	1	1	9	0	0	1	0	0	0	0	2	1	0	0	0	0	0	5	0	0	0	0	0	0	0
k	1	2	8	4	1	1	2	5	0	0	0	0	5	0	2	0	0	0	6	0	0	0	4	0	0	3
l	2	10	1	4	0	4	5	6	13	0	1	0	0	14	2	5	0	11	10	2	0	0	0	0	0	0
m	1	3	7	8	0	2	0	6	0	0	4	4	0	180	0	6	0	0	9	15	13	3	2	2	3	0
n	2	7	6	5	3	0	1	19	1	0	4	35	78	0	0	7	0	28	5	7	0	0	1	2	0	2
o	91	1	1	3	116	0	0	0	25	0	2	0	0	0	0	14	0	2	4	14	39	0	0	0	18	0
p	0	11	1	2	0	6	5	0	2	9	0	2	7	6	15	0	0	1	3	6	0	4	1	0	0	0
q	0	0	1	0	0	0	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
r	0	14	0	30	12	2	2	8	2	0	5	8	4	20	1	14	0	0	12	22	4	0	0	1	0	0
s	11	8	27	33	35	4	0	1	0	1	0	27	0	6	1	7	0	14	0	15	0	0	5	3	20	1
t	3	4	9	42	7	5	19	5	0	1	0	14	9	5	5	6	0	11	37	0	0	2	19	0	7	6
u	20	0	0	0	44	0	0	0	64	0	0	0	0	2	43	0	0	4	0	0	0	0	2	0	8	0
v	0	0	7	0	0	3	0	0	0	0	0	1	0	0	1	0	0	0	8	3	0	0	0	0	0	0
w	2	2	1	0	1	0	0	2	0	0	1	0	0	0	0	7	0	6	3	3	1	0	0	0	0	0
x	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0
y	0	0	2	0	15	0	1	7	15	0	0	0	2	0	6	1	0	7	36	8	5	0	0	1	0	0
z	0	0	0	7	0	0	0	0	0	0	0	7	5	0	0	0	0	2	21	3	0	0	0	0	3	0

Geração da matriz de confusão

- Lista de erros de Peter Norvig
- Lista de contagens de erros de edição única de Peter Norvig

http://norvig.com/ngrams/spell-errors.txt

Modelo de Canal

Kernighan, Church, Gale 1990

$$P(x|w) = \begin{cases} \frac{\text{del}[w_{i-1}, w_i]}{\text{count}[w_{i-1} w_i]}, & \text{if deletion} \\ \frac{\text{ins}[w_{i-1}, x_i]}{\text{count}[w_{i-1}]}, & \text{if insertion} \\ \frac{\text{sub}[x_i, w_i]}{\text{count}[w_i]}, & \text{if substitution} \\ \frac{\text{trans}[w_i, w_{i+1}]}{\text{count}[w_i w_{i+1}]}, & \text{if transposition} \end{cases}$$

Modelo de Canal para acress

Candidato de correção	Letra Correta	Letra Errada	x|w	P(x|word)
actress	t	-	c|ct	.000117
cress	-	a	a|#	.00000144
caress	ca	ac	ac|ca	.00000164
access	c	r	r|c	.000000209
across	o	e	e|o	.0000093
acres	-	s	es|e	.0000321
acres	-	s	ss|s	.0000342

Probabilidade de Noisy Channel para acress

Candidato de correção	Letra Correta	Letra Errada	x|w	P(x|word)	P(word)	10^9 * P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	o	e	e|o	.0000093	.000299	2.8
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Probabilidade de Noisy Channel para acress

Candidato de correção	Letra Correta	Letra Errada	x|w	P(x|word)	P(word)	10^9 * P(x|w)P(w)
actress	t	-	c|ct	.000117	.0000231	2.7
cress	-	a	a|#	.00000144	.000000544	.00078
caress	ca	ac	ac|ca	.00000164	.00000170	.0028
access	c	r	r|c	.000000209	.0000916	.019
across	**o**	**e**	**e|o**	**.0000093**	**.000299**	**2.8**
acres	-	s	es|e	.0000321	.0000318	1.0
acres	-	s	ss|s	.0000342	.0000318	1.0

Usando um Modelo de Linguagem Bigram

- "a stellar and versatile **acress** whose combination of sass and glamour…"
- Contagens do *Corpus of Contemporary American English* com suavização add-1
- P(actress|versatile)=.000021 P(whose|actress) = .0010
- P(across|versatile) =.000021 P(whose|across) = .000006

- P("versatile actress whose") = .000021*.0010 = 210 x10^{-10}
- P("versatile across whose") = .000021*.000006 = 1 x10^{-10}

Usando um Modelo de Linguagem Bigram

- "a stellar and versatile **acress** whose combination of sass and glamour…"
- Contagens do *Corpus of Contemporary American English* com suavização add-1
- P(actress|versatile)=.000021 P(whose|actress) = .0010
- P(across|versatile) =.000021 P(whose|across) = .000006

- **P("versatile actress whose") = .000021*.0010 = 210 x10^{-10}**
- P("versatile across whose") = .000021*.000006 = 1 x10^{-10}

Avaliação

- Alguns conjuntos de teste de erros de ortografia
  - Lista de erros comuns em Inglês da Wikipedia
  - Versão filtrada da lista Aspell
  - Conjunto de Erros Ortográficos de Birkbeck
  - Lista de Erros de Peter Norvig (inclui Wikipedia e Birkbeck, para treinamento e teste)

Correção Ortográfica e Canal com Ruído

Ortografia: Modelo de Canal com Ruído (Noisy Channel)

Correção Ortográfica e Canal com Ruído

Correção Ortográfica Real-Word

Erros Ortográficos Real-word

- …leaving in about fifteen ***minuets*** to go to her house.
- The design ***an*** construction of the system…
- Can they ***lave*** him my messages?
- The study was conducted mainly ***be*** John Black.

- 25-40% dos erros ortográficos são palavras reais Kukich 1992

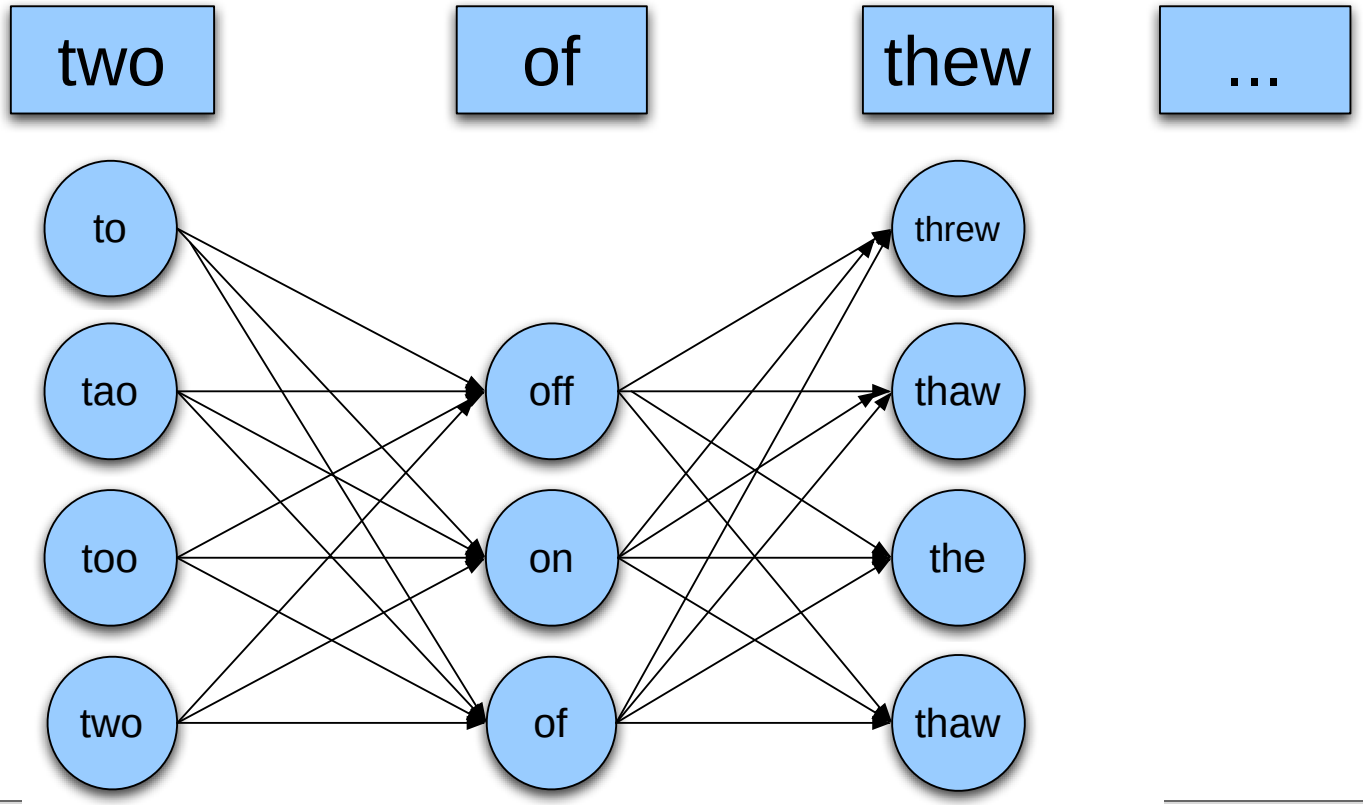
Resolvendo Erros Ortográficos Real-world

- Para cada palavra em uma sentença
  - Gerar um *conjunto de candidatos*
    - A palavra em si
    - Todas edições de letra única que são palavras em Inglês
    - Palavras que são homófonas
- Escolher melhores candidatos
    - Modelo Noisy channel
    - Classificador para tarefa específica

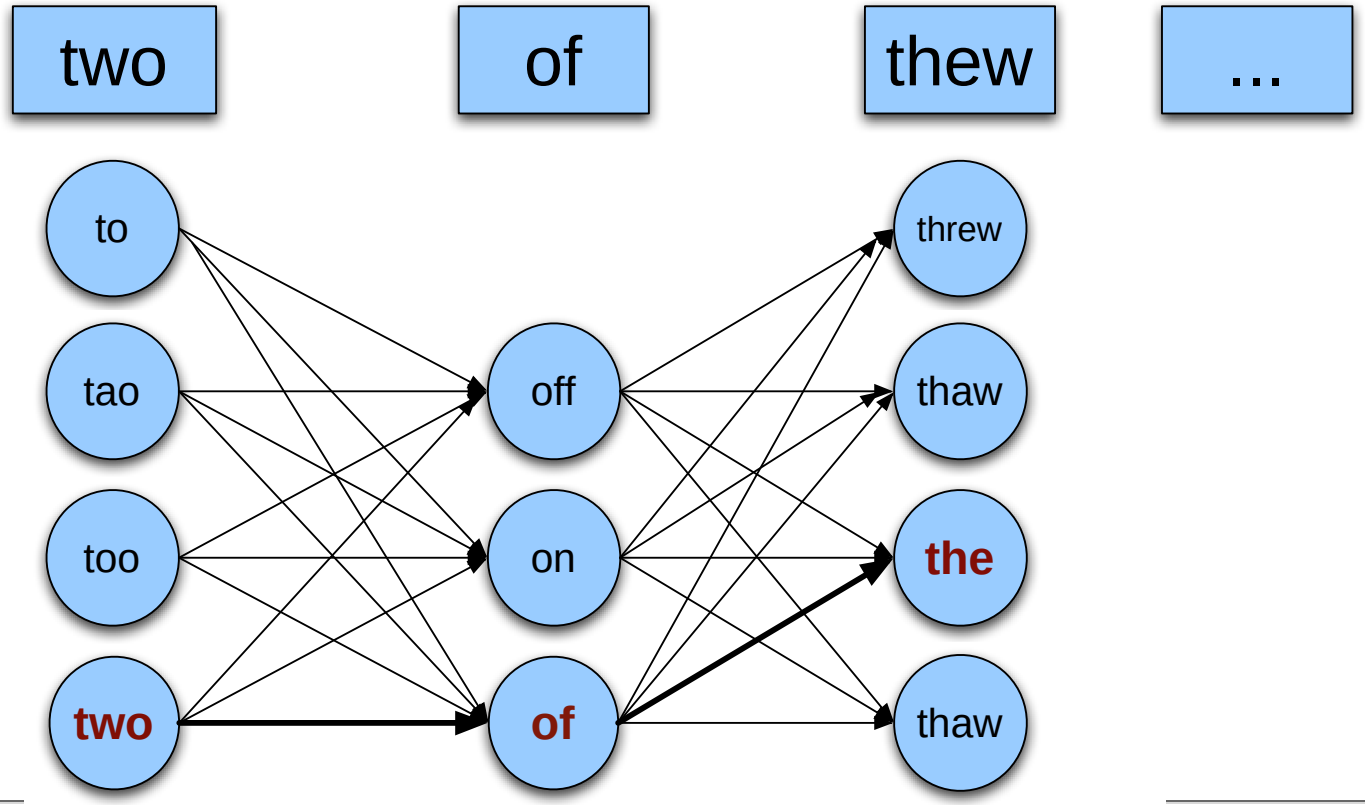
Noisy Channel para Correção Ortográfica em Real-Word

- Dada a sentença W = $w_1, w_2, w_3, \ldots, w_n$
- Gerar um conjunto de candidatos para cada palavra w_i
  - Candidate(w_1) = {w_1, w'_1 , w''_1 , w'''_1 ,…}
  - Candidate(w_2) = {w_2, w'_2 , w''_2 , w'''_2 ,…}
  - Candidate(w_n) = {w_n, w'_n , w''_n , w'''_n ,…}
- Escolher a sequência W que maximiza P(W)

Noisy channel para Correção Ortográfica em Real-word



Noisy channel para Correção Ortográfica em Real-word



Simplificação: Um erro por sentença

- Retorno de todas as sentenças possíveis com uma palavra substituída
  - w_1, **w''$_2$**,w_3,w_4 two **off** thew
  - w_1,w_2,**w'$_3$**,w_4 two of **the**
  - **w'''$_1$**,w_2,w_3,w_4 **too** of thew
  - …
- Escolher uma sequência de W que maximiza P(W)

Onde conseguir as probabilidades

- Modelo de Linguagem
  - Unigram
  - Bigram
  - Etc.
- Modelo de Canal
  - Mesmo usado para correção ortográfica non-word
  - Além disso, precisa de probabilidade para nenhum erro, P(w|w)

Probabilidade de nenhum erro

- Qual é a probabilidade de canal para uma palavra escrita corretamente?
- P(“the”|“the”)
- Obviamente depende da aplicação
  - .90 (1 error in 10 words)
  - .95 (1 error in 20 words)
  - .99 (1 error in 100 words)
  - .995 (1 error in 200 words)

Exemplo “thew” de Peter Norvig

x	w	x|w	P(x|w)	P(w)	10^9 P(x|w)P(w)
thew	the	ew|e	0.000007	0.02	144
thew	thew		0.95	0.00000009	90
thew	thaw	e|a	0.001	0.0000007	0.7
thew	threw	h|hr	0.000008	0.000004	0.03
thew	thwe	ew|we	0.000003	0.00000004	0.0001

Correção Ortográfica e Canal com Ruído

Correção Ortográfica Real-Word

Correção Ortográfica e Canal com Ruído

Sistemas Estado-da-Arte

UNIVERSIDADE FEDERAL DE PERNAMBUCO

Questões HCI em Ortografia

- Se bastante confiante na correção
  - Correção Automática
- Se não tão confiante
  - Apresentar melhor correção
- Ainda menos confiante
  - Apresentar uma lista de correções
- Se não tem confiança
  - Basta marcar como erro

Noisy channel: Estado da Arte

- Em geral não apenas se multiplica o anterior e o modelo de erro.

- Ao invés: Pondere-os

$$\hat{w} = \underset{w \in V}{\operatorname{argmax}} P(x \mid w) P(w)^{\lambda}$$

- Aprenda λ de um conjunto de teste de desenvolvimento

Modelo de Erro Fonético

- Metaphone, usado no GNU aspell
  - Converte erro ortográfico em pronúncia metaphone
    - “Remover letras adjacentes duplicadas com exceção de C”
    - “se a palavra começa com 'KN', 'GN', 'PN', 'AE', 'WR', remova a primeira letra.”
    - “Remova 'B‘ se após 'M' e se ele estiver no final da sentença”
    - …
  - Encontra palavras as quais a pronúncia tem distância de edição entre 1-2 de erros ortográficos
  - Lista de Resultados de Score
    - Distância de edição ponderada de candidato ao erro
    - Distância de edição da pronúncia do candidato a pronúncia do erro

Melhorias no Modelo de Canal

- Permite edições mais ricas (Brill and Moore 2000)
  - ent→ant
  - ph→f
  - le→al
- Incorpora pronúncia ao canal (Toutanova and Moore 2002)

Modelo de Canal

- Fatores que poderiam influenciar p(erro de ortografia|palavra)
  - A letra original
  - A letra alvo
  - Letras ao redor
  - A posição na palavra
  - Teclas próximas no teclado
  - Homologia no teclado
  - Pronúncias
  - Transformações prováveis de morfemas

Teclas próximas

Métodos baseados em classificadores para correções ortográficas real-word

- Ao invés de apenas modelo de canal e modelo de linguagem
- Usar muitas características em um classificador (próxima aula)
- Construir um classificador para um par específico como:

  whether/weather
  - "cloudy" em +- 10 palavras
  - ___ to VERB
  - ___ or not

Correção Ortográfica e Canal com Ruído

Sistemas Estado-da-Arte

UNIVERSIDADE FEDERAL DE PERNAMBUCO