

Processamento Estatístico da Linguagem Natural

Aula 9

Professora Bianca
(Sala 302 – Bloco E)
bianca@ic.uff.br

<http://www.ic.uff.br/~bianca/peln/>

Aula 9 - 02/10/2008

1

Aula de Hoje

- Explicação sobre o trabalho
- Cap. 4 – Jurafsky & Martin
 - Seções 4.3.1, 4.3.2, 4.4 e 4.5

Aula 9 - 02/10/2008

2

O método de visualização de Shannon

- Para gerar sentenças aleatórias:
- Escolha um bigrama aleatório <s>, w de acordo com a sua probabilidade.
- Escolha um bigrama (w, x) de acordo com a sua probabilidade.
- Até que </s> seja escolhido.

- <s> I
I want
want to
to eat
eat Chinese
Chinese food
food </s>

Aula 9 - 02/10/2008

3

Unigram	• To him swallowed confess hear both. Which. Of save on trail for are ay device and rote life have • Every enter now severally so, let • Hill he late speaks; or! a more to leg less first you enter • Are where exeunt and sighs have rise excellency took of.. Sleep knave we, near; vile like
Bigram	• What means, sir. I confess she? then all sorts, he is trim, captain. • Why dost stand forth thy canopy, forsooth; he is this palpable hit the King Henry. Live king. Follow. • What we, hath got so she that I rest and sent to scold and nature bankrupt, nor the first gentleman? • Enter Menenius, if it so many good direction found't thou art a strong upon command of fear not a liberal largess given away, Falstaff! Exeunt
Trigram	• Sweet prince, Falstaff shall die. Harry of Monmouth's grave. • This shall forbid it should be branded, if renown made it empty. • Indeed the duke; and had a very good friend. • Fly, and will rid me these news of price. Therefore the sadness of parting, as they say, 'tis done.
Quadrigram	• King Henry. What! I will go seek the traitor Gloucester. Exeunt some of the watch. A great banquet serv'd in; • Will you not tell me who I am? • It cannot be but so. • Indeed the short and the long. Marry, 'tis a noble Lepidus.

Shakespeare como corpus

- N=884.647 tokens, V=29.066
- Shakespeare produziu 300.000 tipos de bigramas dos $V^2 = 844$ milhões de possíveis bigramas: então, 99.96% dos bigramas possíveis nunca foram vistos. (tem entrada zero na tabela)
- Quadrigramas: o que sai parece com Shakespeare porque é Shakespeare.

Aula 9 - 02/10/2008

5

O Wall Street Journal não é Shakespeare

unigram: Months the my and issue of year foreign new exchange's september were recession exchange new endorsed a acquire to six executives

bigram: Last December through the way to preserve the Hudson corporation N. B. E. C. Taylor would seem to complete the major central planners one point five percent of U. S. E. has already old M. X. corporation of living on information such as more frequently fishing to keep her

trigram: They also point to ninety nine point six billion dollars from two hundred four oh six three percent of the rates of interest stores as Mexico and Brazil on market conditions

Aula 9 - 02/10/2008

6

Palavras desconhecidas: Vocabulário fechado e aberto

- Se sabemos todas as palavras de antemão
 - Vocabulário V é fixo
 - Tarefa de vocabulário **fechado**
- Muitas vezes não sabemos todas as palavras de antemão
 - Fora do vocabulário = OOV words
 - Tarefa de vocabulário **aberto**

Aula 9 - 02/10/2008

7

Vocabulário Aberto

- Usar o token <UNK> para palavras desconhecidas
 - Escolher um vocabulário fixo $L \in V$
 - Na fase de normalização, cada palavra que não esteja em L é modificada para <UNK>
 - As probabilidades de <UNK> são estimadas como as de qualquer outra palavra.
 - As probabilidades <UNK> são usadas para qualquer palavra que não estiver no treinamento.
 - Alternativa: não escolher um vocabulário e substituir a primeira ocorrência de cada palavra por <UNK>

Aula 9 - 02/10/2008

8

Avaliação de N-gramas

- Melhor avaliação para um modelo de N-grama
 - Utilizar o modelo A em uma aplicação (ex: reconhecimento de voz)
 - Utilizar o modelo B na mesma aplicação,
 - Comparar o desempenho de A e B na mesma aplicação
 - **Avaliação "in-vivo" ou extrínseca**
- Mas esse tipo de avaliação consome muito tempo
 - O experimento pode levar dias.
- Então
 - Queremos uma métrica independente da aplicação que possa ser usada de forma rápida.
 - Para avaliar N-gramas normalmente usamos uma métrica chamada **perplexidade**
 - **Avaliação intrínseca**
 - Só é utilizada para desenvolvimento, não é suficiente para publicação de resultados.

Aula 9 - 02/10/2008

9

Perplexidade

- Perplexidade é o inverso da probabilidade do conjunto de teste (segundo um modelo), normalizada pelo número de palavras:

$$PP(W) = P(w_1 w_2 \dots w_N)^{-\frac{1}{N}}$$

$$= \sqrt[N]{\frac{1}{P(w_1 w_2 \dots w_N)}}$$
- Regra da Cadeia: $PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_1 \dots w_{i-1})}}$
- Para bigramas: $PP(W) = \sqrt[N]{\prod_{i=1}^N \frac{1}{P(w_i | w_{i-1})}}$
- Minimizar perplexidade é o mesmo que maximizar a probabilidade
 - **O melhor modelo é o que melhor prevê um conjunto de teste desconhecido.**

Aula 9 - 02/10/2008

10

Outra maneira de ver a perplexidade

- A perplexidade pode ser vista como um fator de ramificação ponderado.
- O fator de ramificação de uma linguagem é o número possível de palavras que podem seguir qualquer outra palavra.
 - Uma linguagem de 10 dígitos aleatórios tem perplexidade 10.
 - Se o dígito 0 é muito mais freqüente que os outros a perplexidade cai.
 - O fator de ramificação ainda é 10, mas o fator de ramificação ponderado de acordo com a freqüência cai.

Aula 9 - 02/10/2008

11

Problemas com MLE

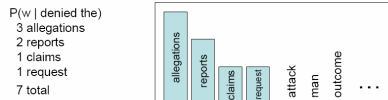
- Lei de Zipf:
 - Um pequeno número de eventos ocorre com alta freqüência
 - Um grande número de eventos ocorre com baixa freqüência
 - Podemos coletar estatísticas para os eventos de alta freqüência
 - Temos que ter dados arbitrariamente grandes para poder coletar estatísticas sobre eventos de baixa freqüência.
- Resultado:
 - As estimativas são esparsas = zero para a maioria dos eventos
 - Alguns são realmente eventos que não ocorrem nunca, mas outros são eventos de baixa freqüência que não ocorreram naquele corpus em particular.
- Solução:
 - Utilizar "smoothing" = dar um pouco de probabilidade aos eventos que não aconteceram nos dados de treinamento.

Aula 9 - 02/10/2008

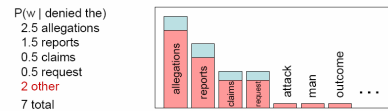
12

Smoothing = Robin Hood: Roubar dos ricos e dar para os pobres (em probabilidade)

- We often want to make predictions from sparse statistics:



- Smoothing flattens spiky distributions so they generalize better



- Very important all over NLP, but easy to do badly!

Slide from Dan Klein

Aula 9 - 02/10/2008

13

Laplace smoothing

- Também chamado de “add-one smoothing”
- Simplesmente adiciona 1 a todas as contagens.

Estimativa MLE: $P(w_i) = \frac{c_i}{N}$

Estimativa Laplace: $P_{\text{Laplace}}(w_i) = \frac{c_i + 1}{N + V}$

Contagens ajustadas: $c_i^* = (c_i + 1) \frac{N}{N + V}$

Aula 9 - 02/10/2008

14

Exemplo Adicionar 1 às contagens

	i	want	to	eat	chinese	food	lunch	spend
i	6	828	1	10	1	1	1	3
want	3	1	609	2	7	7	6	2
to	3	1	5	687	3	1	7	212
eat	1	1	3	1	17	3	43	1
chinese	2	1	1	1	1	83	2	1
food	16	1	16	1	2	5	1	1
lunch	3	1	1	1	1	2	1	1
spend	2	1	2	1	1	1	1	1

Aula 9 - 02/10/2008

15

Exemplo Calcular probabilidades

$$P^*(w_n | w_{n-1}) = \frac{C(w_{n-1}w_n) + 1}{C(w_{n-1}) + V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Aula 9 - 02/10/2008

16

Exemplo Contagens ajustadas

$$c^*(w_{n-1}w_n) = \frac{[C(w_{n-1}w_n) + 1] \times C(w_{n-1})}{C(w_{n-1}) + V}$$

	i	want	to	eat	chinese	food	lunch	spend
i	3.8	527	0.64	6.4	0.64	0.64	0.64	1.9
want	1.2	0.39	238	0.78	2.7	2.7	2.3	0.78
to	1.9	0.63	3.1	430	1.9	0.63	4.4	133
eat	0.34	0.34	1	0.34	5.8	1	15	0.34
chinese	0.2	0.098	0.098	0.098	0.098	8.2	0.2	0.098
food	6.9	0.43	6.9	0.43	0.86	2.2	0.43	0.43
lunch	0.57	0.19	0.19	0.19	0.19	0.38	0.19	0.19
spend	0.32	0.16	0.32	0.16	0.16	0.16	0.16	0.16

Aula 9 - 02/10/2008

17

Grandes mudanças nas contagens

- C(count to) foi de 608 para 238!
- P(to|want) foi de 0.66 para 0.26!
- Desconto $d = c^*/c$
 - d para “chinese food” = 10! Uma redução de 10x
 - Em geral, Laplace smoothing é muito drástico
 - Poderíamos adicionar um número menor que 1 (add- δ)
 - Ainda há problemas (como escolher o δ ?)

Aula 9 - 02/10/2008

18

Técnicas de “smoothing” melhores

- Intuição usada por muitos algoritmos de smoothing como
 - Good-Turing
 - Kneser-Ney
 - Witten-Bell
- é usar a contagem de eventos que vimos apenas uma vez para estimar a contagem de eventos que nunca vimos.

Aula 9 - 02/10/2008

19

Good-Turing

- Imagine que você está pescando
 - Existem 8 espécies: carpa, bagre, traíra, truta, salmão, enguia, tilápia, corvina
- You have caught
 - 10 carps, 3 bagres, 2 traíras, 1 truta, 1 salmão, 1 enguia
- Qual é a probabilidade da próxima espécie ser nova (i.e. tilápia ou corvina)
 - 3/18
- Qual é a probabilidade da nova espécie ser truta?
 - Menos de 1/18

Aula 9 - 02/10/2008

20

Good-Turing

- Notação: N_x é a o número de itens com frequência x (frequência-da-frequência)
 - Logo no exemplo $N_0=1$, $N_1=3$, etc
- Para estimar o número total de espécies
 - Usar o número de espécies (ou palavras) que vimos uma vez
 - $c_0^* = c_1$ $p_0 = N_1/N = 3/18$
- Todas as outras contagens são ajustadas para baixo de acordo com a seguinte fórmula:

$$c^* = (c+1) \frac{N_{c+1}}{N_c} \quad c(\text{enguia}) = c^*(1) = (1+1) 1/3 = 2/3$$

Aula 9 - 02/10/2008

21

Good-Turing

	unseen (bass or catfish)	trout
c	0	1
MLE p	$p = \frac{0}{18} = 0$	$\frac{1}{18}$
c^*		$c^*(\text{trout}) = 2 \times \frac{N_2}{N_1} = 2 \times \frac{1}{3} = .67$
GT p_{GT}^*	$p_{GT}^*(\text{unseen}) = \frac{N_1}{N} = \frac{3}{18} = .17$	$p_{GT}^*(\text{trout}) = \frac{.67}{18} = \frac{1}{27} = .037$

Aula 9 - 02/10/2008

22

Frequências-de-frequências e estimativas de Good-Turing

AP Newswire			Berkeley Restaurant—		
c (MLE)	N_c	c^* (GT)	c (MLE)	N_c	c^* (GT)
0	74,671,100,000	0.0000270	0	2,081,496	0.002553
1	2,018,046	0.446	1	5315	0.533960
2	449,721	1.26	2	1419	1.357294
3	188,933	2.24	3	642	2.373832
4	105,668	3.24	4	381	4.081365
5	68,379	4.22	5	311	3.781350
6	48,190	5.19	6	196	4.500000

Aula 9 - 02/10/2008

23

Good-Turing - complicações

- Na prática, é comum supor que as contagens grandes ($c > k$ para algum k) são confiáveis:

$$c^* = c \text{ for } c > k$$

- Isso complica c^* :

$$c^* = \frac{(c+1) \frac{N_{c+1}}{N_c} - c \frac{(k+1)N_{k+1}}{N_1}}{1 - \frac{(k+1)N_{k+1}}{N_1}}, \text{ for } 1 \leq c \leq k.$$

- Também: supor que contagem $c=1$ não são confiáveis, entra tratamos N-gramas com contagem 1 como se fossem N-gramas com contagem zero.
- Também, precisamos que os N_k sejam não-nulos, então temos que interpolar as contagens N_k antes de computar c^* a partir deles.

Aula 9 - 02/10/2008

24