

Processamento Estatístico da Linguagem Natural

Aula 12

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

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

Aula 12 - 14/10/2008

1

Aula de Hoje

- Cap. 5 – Jurafsky & Martin – POS Tagging
– Seções 5.5.1, 5.6, 5.7

Aula 12 - 14/10/2008

2

Revisão: Abordagens existentes para rotulação POS

- Regras escritas por especialistas
 - Especificam, por exemplo, que uma palavra é um substantivo (não um verbo) quando vem depois de um determinante.
- Abordagem estatística
 - Usam um corpus de treinamento para calcular a probabilidade de cada rótulo em um determinado contexto.
 - Exemplo: rotulador HMM
- Sistemas híbridos
 - Exemplo: rotulador de Brill.

Aula 12 - 14/10/2008

3

Rotulação estatística: HMM

- Intuição: Escolher o rótulo mais provável.
- Rotuladores HMM escolhem a sequência de rótulos que maximiza a fórmula:
 - $P(\text{palavra}|\text{rótulo}) \times P(\text{rótulo} | n \text{ rótulos anteriores})$
- Seja $T = t_1, t_2, \dots, t_n$
Seja $W = w_1, w_2, \dots, w_n$
- Escolher a sequência de rótulos que tenha a maior probabilidade de gerar a sequência de palavras observadas.

Aula 12 - 14/10/2008

4

Rotulador HMM baseado em bigramas

- $\text{argmax}_T P(T|W)$
- $\text{argmax}_T P(T)P(W|T)$
- $\text{argmax}_T P(t_1 \dots t_n)P(w_1 \dots w_n | t_1 \dots t_n)$
- $\text{argmax}_i [P(t_1)P(t_2|t_1) \dots P(t_n|t_{n-1})][P(w_1|t_1)P(w_2|t_2) \dots P(w_n|t_n)]$
- Para rotular uma palavra: $t_i = \text{argmax}_j P(t_j|t_{i-1})P(w_i|t_j)$
- Como calcular $P(t_i|t_{i-1})$?
 - $c(t_i, t_i)/c(t_{i-1})$
- Como calcular $P(w_i|t_i)$?
 - $c(w_i, t_i)/c(t_i)$
- Como calcular a sequência de rótulos mais provável?
 - Viterbi (Cap. 6)

Aula 12 - 14/10/2008

5

Rotuladores HMM

- Rotulador baseado em bigramas
 - Faz previsões usando apenas o rótulo anterior
- Rotulador baseado em trigramas
 - Pode fazer previsões mais corretas levando em consideração um contexto maior (se os dados são suficientes).
 - RB(advérbio) VBD(passado) OU RB VBN(passado particípio) ?
 - Exemplo : "clearly marked"
 - Is clearly marked : $P(\text{BEZ RB VBN}) > P(\text{BEZ RB VBD})$
 - He clearly marked : $P(\text{PN RB VBD}) > P(\text{PN RB VBN})$

Aula 12 - 14/10/2008

6

- Secretariat/NNP is/VBZ expected/VBN to/TO **race**/VB tomorrow/NN
- People/NNS continue/VBP to/TO inquire/VB the DT reason/NN for/IN the/DT **race**/NN for/IN outer/JJ space/NN
- to/TO **race**/???
- the/DT **race**/???
- $t_i = \operatorname{argmax}_t P(t_i|t_{-i})P(w_i|t_i)$
 $= \max\{P(VB|TO)P(\text{race}|VB), P(NN|TO)P(\text{race}|NN)\}$
- Brown:
 - $P(NN|TO) = .00047 \times P(\text{race}|NN) = .00057 = .0000003$
 - $P(VB|TO) = .83 \times P(\text{race}|VB) = .00012 = .0001$

Aula 12 - 14/10/2008

7

- Combinação de rotulador baseado em regras e rotulador estatístico.
 - Regras são usadas para especificar transformações em um certo contexto.
 - Usa um corpus de treinamento para *aprender* as regras.
- Entrada:
 - Corpus de treinamento rotulado
 - Dicionário (com os rótulos mais frequentes)
 - Normalmente construído a partir do corpus de treinamento.

Aula 12 - 14/10/2008

- **Idéia básica:**
 - Inicializar o rótulo de cada palavra com o rótulo mais provável
 - Mudar rótulos de acordo com regras do tipo "se palavra-1 for um determinante e a palavra for um verbo então mude rótulo para substantivo" em uma ordem específica.
- **Treinamento é feito em um corpus rotulado:**
 1. Escrever um conjunto de templates de regras
 2. Dentre esse conjunto, encontrar a regra com o maior score.
 3. Continuar de 2 até ultrapassar um score mínimo
 4. Manter um conjunto ordenado de regras
- **Regras que erram podem ser corrigidas por outras regras.**

Aula 12 - 14/10/2008

©

- Rotulador rotula cada palavra com seu rótulo mais provável.
 - Por exemplo: *race* tem as seguintes probabilidades no corpus de Brown:
 - $P(NN|race) = .98$
 - $P(VB|race) = .02$
- Regras de transformação fazem mudanças nos rótulos
 - “Mudar NN para VB quando o rótulo anterior é TO”
... is/VBZ expected/VBN to/TO race/NN tomorrow/NN
fica... is/VBZ expected/VBN to/TO race/VB tomorrow/NN

Aula 12 - 14/10/2008

1

The preceding (following) word is tagged z. The word two before (after) is tagged z. One of the two preceding (following) words is tagged z. One of the three preceding (following) words is tagged z. The preceding word is tagged z and the following word is tagged w. The preceding (following) word is tagged z and the word two before (after) is tagged w.

Figure 8.9 Brill's (1995) templates. Each begins with 'Change tag **a** to tag **b** when:'. The variables **a**, **b**, **z**, and **w** range over parts of speech.

Aula 12 - 14/10/2008

11

Schema	τ_{i-3}	τ_{i-2}	τ_{i-1}	τ_i	τ_{i+1}	τ_{i+2}	τ_{i+3}
1							
2				*			
3				*			
4				*			
5				*			
6				*			
7				*			
8				*			
9				*			

Aula 12 - 14/10/2008

1

Exemplos: Rotulador de Brill

#	Change tags		Condition	Example
	From	To		
1	NN	VB	Previous tag is TO	to/TO race/NN → VB
2	VBP	VB	One of the previous 3 tags is MD	might/MD vanish/VBP → VB
3	NN	VB	One of the previous 2 tags is MD	might/MD not reply/NN → VB
4	VB	NN	One of the previous 2 tags is DT	
5	VBD	VBN	One of the previous 3 tags is VBZ	

Figure 8.11 The first 20 nonlexicalized transformations from the Brill tagger (Brill, 1995).

Aula 12 - 14/10/2008

13

Avaliação

- Para qualquer problema de NLP, precisamos saber avaliar as soluções.
- Primeiro, separamos o corpus em treinamento e teste.
 - O corpus de teste anotado manualmente.
 - Chamado de "Gold Standard".
 - Calcula-se a acurácia.
 - Número de palavra em que o rótulo é igual ao "gold standard".
 - Para os algoritmos atuais fica em torno de 96-7%
 - Importante usar testes estatísticos para determinar se diferenças são significativas.
 - "Teto humano": verificar a taxa de concordância entre rotuladores humanos diferentes.
 - A concordância fica em torno de 96-7%.
 - Quando os humanos não concordam, escolhe-se o rótulo com a maior frequência no corpus de treinamento.

Aula 12 - 14/10/2008

14

Análise de Erros

- Matriz de confusão
 - Linhas mostram rótulo verdadeiro
 - Colunas mostram rótulo previsto pelo rotulador
 - Cada célula indica a percentagem de erros desse tipo.

	IN	JJ	NN	NNP	RB	VBD	VBN
IN	-	.2			.7		
JJ	.2	-	3.3	2.1	1.7	.2	2.7
NN		8.7	-				.2
NNP	.2	3.3	4.1	-	.2		
RB	2.2	2.0	.5		-		
VBD		.3	.5			-	4.4
VBN		2.8				2.6	-

Aula 12 - 14/10/2008

15

Fatores que afetam o desempenho do rotulador

- Quantidade de dados de treinamento disponível.
- O conjunto de rótulos utilizado.
- Diferenças entre o corpus de treinamento e de teste.
- O dicionário utilizado.
 - Lista as classes gramaticais possíveis para cada palavra.
- Palavras desconhecidas.

Aula 12 - 14/10/2008

16