

Processamento Estatístico da Linguagem Natural

Aula 11

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

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

Aula 11 - 09/10/2008

1

Aula de Hoje

- Cap. 5 – Jurafsky & Martin – POS Tagging
– Seções 5.1, 5.2, 5.3, 5.5

Aula 11 - 09/10/2008

2

O que é uma classe gramatical?

(Em inglês, *part-of-speech* (POS) ou *word class*).

- É um conjunto de palavras que “se comporta” da mesma maneira:
 - Aparecem em contextos similares.
 - Têm funções parecidas nas frases.
 - Passam por transformações morfológicas parecidas.

Aula 11 - 09/10/2008

3

Exemplos de Classes Gramaticais

- Substantivos: pessoas, animais, conceitos, coisas
- Verbos: expressam ações nas frases
- Adjetivos: descrevem propriedades dos substantivos

The {
sad
intelligent
green
fat
...
} one is in the corner.

“Substitution test”

Aula 11 - 09/10/2008

4

Por que identificar classes gramaticais?

- Modelos de linguagem mais precisos.
 - N-gramas podem usar classe gramatical como “backoff”.
- Pronúncia
 - Exemplo: *deSERT* (V) e *DEsert* (N)
- *Stemming* (identificar o radical)
 - A classe gramatical ajuda a determinar que afixos são possíveis.
- Determinar o significado de palavras com mais de um significado (quando são de classes gramaticais diferentes).
 - Exemplo: interest
- ...

Aula 11 - 09/10/2008

5

Quantas classes gramaticais existem?

- Em inglês, conjunto básico é:
 - Substantivos, Verbos, Adjetivos, Advérbios, Preposições, Determinantes, Pronomes, Conjunções, Verbos Auxiliares, Partículas, Numerais
- São divididas em duas categorias:
 - Classes abertas: admitem novas palavras.
 - Substantivos, Verbos, Adjetivos, Advérbios.
 - Classes fechadas: formadas por um conjunto fixo de palavras.
 - Preposições, Artigos, Pronomes, Conjunções, Verbos Auxiliares, Partículas, Numerais

Aula 11 - 09/10/2008

6

Classes Gramaticais (em Inglês)

- **Substantivos** denotam pessoas, lugares ou coisas e podem ser precedidos por artigos, mas...
 - *My typing is very bad*
 - *The Mary loves John** (*=frase não-gramatical)
- Podem ser divididos em
 - Substantivos **próprios** e substantivos **comuns**.
 - Substantivos **contáveis** e **não-contáveis**.

Aula 11 - 09/10/2008

7

Classes Gramaticais (em Inglês)

- **Verbos** denotam ações e processos.
 - Têm várias formas morfológicas.
 - É uma classe aberta.
 - *I will have **emailed** everyone by noon.*
 - Há uma classe de verbos especial chamada de auxiliares que é fechada.
- **Adjetivos** descrevem propriedades ou qualidades, mas...
 - *a cat sitter, a child seat*

Aula 11 - 09/10/2008

8

Classes Gramaticais (em Inglês)

- **Advérbios** modificam verbos, outros advérbios ou locuções verbais.
 - Podem especificar local ou direção da ação (*here, downhill*).
 - Podem especificar a extensão ou grau da ação (*very, extremely, somewhat*)
 - Podem especificar a maneira da ação (*slowly, delicately*)
 - Podem especificar quando uma ação ocorreu (*yesterday, ago*)
 - *Monday* é um advérbio ou substantivo?

Aula 11 - 09/10/2008

9

Classes Gramaticais (em Inglês)

- **Advérbios** modificam verbos, outros advérbios ou orações.
 - Podem especificar local ou direção da ação (*here, downhill*).
 - Podem especificar a extensão ou grau da ação (*very, extremely, somewhat*)
 - Podem especificar a maneira da ação (*slowly, delicately*)
 - Podem especificar quando uma ação ocorreu (*yesterday, ago*)
 - *Monday* é um advérbio ou substantivo?

Aula 11 - 09/10/2008

10

Classes Gramaticais (em Inglês)

- **Preposições** ocorrem antes de locuções substantivas, indicando relações, normalmente temporais ou espaciais.
- **Partículas** parecem preposições ou advérbios e são usadas em combinações com verbos, modificando seu significado.
- Preposições vs. Partículas
 - *So it went on talking to him* (on = partícula)
 - *The book is on the table* (on = preposição)

Aula 11 - 09/10/2008

11

Classes Gramaticais (em Inglês)

- **Determinantes** marcam o início de uma locução substantiva.
 - Podem ser artigos.
 - *A, an e the.*
 - *The* é a palavra mais frequente do inglês (6%)
 - Outros determinantes são *this, that*.
- **Conjunções** juntam duas orações.
 - Coordenadas (*and, but*) ou subordinadas (*that, because, unless*).

Aula 11 - 09/10/2008

12

Classes Gramaticais (em Inglês)

- **Pronomes** substituem locuções substantivas, entidades ou eventos.
 - Podem ser pessoais (*I, he, she*), possessivos (*my, his*), ou wh (*what, who, whom*).
- **Verbos auxiliares** incluem os verbos de acoplamento (*be, do e have*) e os verbos modais (*can, will, shall, must*).
- E mais
 - Interjeições
 - *There* existencial
 - Saudações e termos de boas maneiras (*hello, please*)

Aula 11 - 09/10/2008

13

A tarefa de rotulação POS

Input: *the lead paint is unsafe*

Output: *the/Det lead/N paint/N is/V unsafe/Adj*

- Que conjunto de rótulos utilizar?
 - Brown = 87 rótulos
 - C5 = 62 rótulos
 - Penn Treebank = 45 rótulos
 - É o mais utilizado atualmente.

Aula 11 - 09/10/2008

14

Diferenças entre conjuntos de rótulos

- Grau de granularidade
- Decisões idiossincráticas
 - Exemplos:
 - Penn Treebank não distingue to/Prep de to/Inf. I/PP want/VBP to/TO go/VB to/TO Zanzibar/NNP ./.
 - Maneira de tratar palavras como *Monday* e substantivos usados como adjetivos (*cotton candy*)
- Não ter rótulos diferentes quando é possível distinguir usando a própria palavra.
 - Exemplos: não ter rótulos diferentes para cada forma do verbo *do*.

Aula 11 - 09/10/2008

15

Exemplos de rótulos do Penn Treebank

PART-OF-SPEECH	TAG	EXAMPLES
• Adjective	JJ	happy, bad
• Adjective, comparative	JJR	happier, worse
• Adjective, cardinal number	CD	3, fifteen
• Adverb	RB	often, particularly
• Conjunction, coordination	CC	and, or
• Conjunction, subordinating	IN	although, when
• Determiner	DT	this, each, other, the, a, some
• Determiner, postdeterminer	JJ	many, same
• Noun	NN	aircraft, data
• Noun, plural	NNS	women, books
• Noun, proper, singular	NNP	London, Michael
• Noun, proper, plural	NNPS	Australians, Methodists
• Pronoun, personal	PRP	you, we, she, it
• Pronoun, question	WP	who, whoever
• Verb, base present form	VB	take, live

Aula 11 - 09/10/2008

16

Ambigüidade

- Muitas palavras têm somente um rótulo possível.
 - Exemplos: *is, Mary, very, smallest*
- Porém muitas são ambíguas, isto é, podem ter mais de um rótulo.
 - Brown: 11,5% das palavras do vocabulário são ambíguas, 40% dos tokens são ambíguas.
 - Exemplos:
 - Time/N flies/V like/PPREP an/DET arrow/N
 - Fruit/N flies/N like/V a/DET banana/N
 - Fruit/N flies/V like/PPREP a/DET banana/N

Aula 11 - 09/10/2008

17

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.
- Sistemas híbridos
 - Exemplo: rotulador de Brill.

Aula 11 - 09/10/2008

18

Rotulação POS estatística

- Objetivo: escolher a melhor seqüência de rótulos T para uma seqüência de palavras W numa frase.

$$T' = \arg \max_{T \in \tau} P(T|W)$$

- Pela regra de Bayes

$$P(T|W) = \frac{P(T)P(W|T)}{P(W)}$$

- Como podemos ignorar $P(W)$, temos

$$T' = \arg \max_{T \in \tau} P(T)P(W|T)$$

Aula 11 - 09/10/2008

19

Rotulação POS estatística

$$P(T) = P(t_1, t_2, \dots, t_{n-1}, t_n)$$

Pela regra da cadeia:

$$\begin{aligned} &= P(t_n | t_1, \dots, t_{n-1}) P(t_1, \dots, t_{n-1}) \\ &= \prod_{i=1}^n P(t_i | t_{i-1}^{i-1}) \end{aligned}$$

Usando a hipótese de Markov: $\approx P(t_i | t_{i-N+1}^{i-1})$

$$\text{Para bigramas: } = \prod_{i=1}^n P(t_i | t_{i-1})$$

Aula 11 - 09/10/2008

20

Rotulação POS Estatística

$$P(W|T) = P(w_1, w_2, \dots, w_n | t_1, t_2, \dots, t_n)$$

Pela regra da cadeia:

$$= \prod_{i=1}^n P(w_i | w_{i-1} \dots w_{i-i+1} t_{i-1}^{i-1})$$

Hipótese simplificadora: probabilidade de uma palavra depende só da sua etiqueta $P(w_i | t_i)$

$$\approx \prod_{i=1}^n P(w_i | t_i)$$

Fazendo a multiplicação...

$$T' = \arg \max_{T \in \tau} \prod_{i=1}^n P(t_i | t_{i-1}) \prod_{i=1}^n P(w_i | t_i)$$

Aula 11 - 09/10/2008

21

Estimando as probabilidades

- Maximum-Likelihood Estimation (MLE)

- Para bigramas:

$$P(t_i | t_{i-1}) = c(t_{i-1}, t_i) / c(t_{i-1})$$

$$P(w_i | t_i) = \frac{c(w, t)}{c(t)}$$

Aula 11 - 09/10/2008

22

Exemplo: Brown Corpus

$$\begin{aligned} \bullet P(NN|DT) &= C(DT, NN) / C(DT) \\ &= 56509 / 116454 \\ &= 0.49 \end{aligned}$$

$$\begin{aligned} \bullet P(is|VBZ) &= C(VBZ, is) / C(VBZ) \\ &= 10073 / 21627 \\ &= 0.47 \end{aligned}$$

Aula 11 - 09/10/2008

23