

Processamento Estatístico da Linguagem Natural

Aula 20

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

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

Aula 20 - 20/11/2008

1

Aula de Hoje

- Cap. 14 – Jurafsky & Martin – Parsing Estatístico
 - Seções 14.1, 14.2, 14.3, 14.4.1

Aula 20 - 20/11/2008

2

Como resolver ambiguidades?

- Os métodos tradicionais de parsing geram todas as árvores compatíveis com uma frase.
 - Qual devemos escolher?
- Solução: métodos probabilísticos
 - Adicionar probabilidades às regras da gramática
 - Modificar o parser pra manter somente as árvores mais prováveis.
 - No final, retornar somente a árvore mais provável.

Aula 20 - 20/11/2008

3

Gramáticas livres de contexto probabilísticas (PCFGs)

- O modelo probabilístico
 - Atribuir probabilidades às regras da gramática e às árvores de parse.
- Parsing com probabilidades
 - Modificação pequena na programação dinâmica.
 - Tarefa é encontrar a árvore de maior probabilidade para a frase em questão.

Aula 20 - 20/11/2008

4

Modelo Probabilístico

- Atribuir probabilidades às regras da gramática.
- As probabilidades das regras para cada não-terminal (do lado esquerdo) tem que somar 1
 - VP → Verb .55
 - VP → Verb NP .40
 - VP → Verb NP NP .05
 - Ler cada probabilidade como $P(\text{Regra} \mid \text{Lado-Esquerdo})$

Aula 20 - 20/11/2008

5

Modelo Probabilístico

- Uma derivação (ou árvore) consiste do conjunto de regras gramaticais que estão na árvore.
- A probabilidade uma árvore é o produto das probabilidades das regras na derivação.

Aula 20 - 20/11/2008

6

Modelo Probabilístico

- A probabilidade de uma seqüência de palavras (frase) é a probabilidade de sua árvore se ela não for ambígua.
- É a soma das probabilidades de todas as árvores no caso ambíguo.

Aula 20 - 20/11/2008

7

Obtendo as probabilidades

- A partir de um corpus anotado (um treebank)
 - Então por exemplo, para obter a probabilidade de uma regra com VP do lado esquerdo:
 - Contamos o número de vezes que a regra é usada e dividimos pelo número de vezes que qualquer regra com VP do lado esquerdo é usada.

Aula 20 - 20/11/2008

8

Abordagem típica para o parsing probabilístico

- Programação dinâmica bottom-up (CYK)
- Atribuir probabilidades aos agrupamentos à medida que eles são completados e colocados na tabela.
- Usar a probabilidade máxima para cada agrupamento quando se vai para o próximo nível da árvore.

Aula 20 - 20/11/2008

9

CKY Probabilístico

- Da mesma forma que o algoritmo CKY, requer que a gramática esteja em forma normal de Chomsky.
 - Temos que modificar as probabilidades também ao colocar na forma normal.
 - Na prática também podemos utilizar uma versão do CKY que permite regras unitárias.

Aula 20 - 20/11/2008

10

CKY com pesos

- Antes de pensar no CKY probabilístico veremos um CKY com pesos.
- Cada regra tem um peso e o objetivo é encontrar a árvore mais “leve”, isso é, com menor peso.

Aula 20 - 20/11/2008

11

CKY com pesos

- Entrada: frase, gramática em CNF
- Saída: sim/não
- Estrutura de dados: tabela $n \times n$
 - Linhas vão de 0 a $n-1$
 - Colunas vão de 1 a n
 - Célula (i,j) lista agrupamentos que englobam as palavras de i a j .
- Inicialização
 - Para cada i de 1 a n
 - Adicione a $(i-1,i)$ todos os não-terminais que possam produzir a palavra em $(i-1,i)$
 - Anote o peso de cada não-terminal

Aula 20 - 20/11/2008

12

CKY com pesos

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3								
1			NP 4 VP 4						
2				P 2 V 5					
3						Det 1			
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

13

CKY

Para w de 2 a n

Para **início** de 0 a $n-1$

Defina **fim**=início+ w

Para **meio** de início+1 até **fim**-1

Para cada não-terminal em (**início**,**meio**)

Para cada não-terminal em (**meio**,**fim**)

Para todas as maneiras de combiná-los

Adicione o não-terminal resultante à posição (**início**,**fim**)

O peso resultante será a soma dos pesos de cada não-terminal, mais o peso das regras

Aula 20 - 20/11/2008

14

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3								
1			NP 4 VP 4						
2				P 2 V 5					
3						Det 1			
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

15

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3								
1			NP 4 VP 4						
2				P 2 V 5					
3						Det 1			
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

16

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3								
1			NP 4 VP 4						
2				P 2 V 5					
3						Det 1			
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

17

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3								
1			NP 4 VP 4						
2				P 2 V 5					
3						Det 1			
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

18

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3		NP 10 S 8 S 13						
1			NP 4 VP 4						
2				P 2 V 5					
3					Det 1				
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

19

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3		NP 10 S 8 S 13						
1			NP 4 VP 4						
2				P 2 V 5					
3					Det 1		NP 10		
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

20

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3		NP 10 S 8 S 13						
1			NP 4 VP 4						
2				P 2 V 5					
3					Det 1		NP 10		
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

21

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3		NP 10 S 8 S 13						
1			NP 4 VP 4						
2				P 2 V 5			PP 12		
3					Det 1		NP 10		
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

22

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3		NP 10 S 8 S 13						
1			NP 4 VP 4						
2				P 2 V 5			PP 12 VP 16		
3					Det 1		NP 10		
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

23

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3		NP 10 S 8 S 13						
1			NP 4 VP 4						
2				P 2 V 5			PP 12 VP 16		
3					Det 1		NP 10		
4							N 8		

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

24

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							
1		NP 4 VP 4						NP 18 S 21	
2			P 2 V 5					PP 12 VP 16	
3						Det 1		NP 10	
4								N 8	

1 S → NP VP
 6 S → Vst NP
 2 S → S PP
 1 VP → V NP
 2 VP → VP PP
 1 NP → Det N
 2 NP → NP PP
 3 NP → NP NP
 0 PP → P NP

Aula 20 - 20/11/2008

25

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							
1		NP 4 VP 4						NP 18 S 21	
2			P 2 V 5					PP 12 VP 16	
3						Det 1		NP 10	
4								N 8	

1 S → NP VP
 6 S → Vst NP
 2 S → S PP
 1 VP → V NP
 2 VP → VP PP
 1 NP → Det N
 2 NP → NP PP
 3 NP → NP NP
 0 PP → P NP

Aula 20 - 20/11/2008

26

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							
1		NP 4 VP 4						NP 18 S 21 VP 18	
2			P 2 V 5					PP 12 VP 16	
3						Det 1		NP 10	
4								N 8	

1 S → NP VP
 6 S → Vst NP
 2 S → S PP
 1 VP → V NP
 2 VP → VP PP
 1 NP → Det N
 2 NP → NP PP
 3 NP → NP NP
 0 PP → P NP

Aula 20 - 20/11/2008

27

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							
1		NP 4 VP 4						NP 18 S 21 VP 18	
2			P 2 V 5					PP 12 VP 16	
3						Det 1		NP 10	
4								N 8	

1 S → NP VP
 6 S → Vst NP
 2 S → S PP
 1 VP → V NP
 2 VP → VP PP
 1 NP → Det N
 2 NP → NP PP
 3 NP → NP NP
 0 PP → P NP

Aula 20 - 20/11/2008

28

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22
1		NP 4 VP 4						NP 18 S 21 VP 18	
2			P 2 V 5					PP 12 VP 16	
3						Det 1		NP 10	
4								N 8	

1 S → NP VP
 6 S → Vst NP
 2 S → S PP
 1 VP → V NP
 2 VP → VP PP
 1 NP → Det N
 2 NP → NP PP
 3 NP → NP NP
 0 PP → P NP

Aula 20 - 20/11/2008

29

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22 S 27
1		NP 4 VP 4						NP 18 S 21 VP 18	
2			P 2 V 5					PP 12 VP 16	
3						Det 1		NP 10	
4								N 8	

1 S → NP VP
 6 S → Vst NP
 2 S → S PP
 1 VP → V NP
 2 VP → VP PP
 1 NP → Det N
 2 NP → NP PP
 3 NP → NP NP
 0 PP → P NP

Aula 20 - 20/11/2008

30

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 21 S 18
1		NP 4 VP 4							NP 18 S 21 VP 18
2			P 2 V 5						PP 12 VP 16
3					Det 1				NP 10 N 8
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

31

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22 S 27
1		NP 4 VP 4							NP 18 S 21 VP 18
2			P 2 V 5						PP 12 VP 16
3					Det 1				NP 10 N 8
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

32

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22 S 27 NP 24
1		NP 4 VP 4							NP 18 S 21 VP 18
2			P 2 V 5						PP 12 VP 16
3					Det 1				NP 10 N 8
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

33

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22 S 27 NP 24 S 27
1		NP 4 VP 4							NP 18 S 21 VP 18
2			P 2 V 5						PP 12 VP 16
3					Det 1				NP 10 N 8
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

34

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22 S 27 NP 24 S 27 S 22
1		NP 4 VP 4							NP 18 S 21 VP 18
2			P 2 V 5						PP 12 VP 16
3					Det 1				NP 10 N 8
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

35

CKY

time	1	flies	2	like	3	an	4	arrow	5
0	NP 3 Vst 3	NP 10 S 8 S 13							NP 24 S 22 S 27 NP 24 S 27 S 22 S 27
1		NP 4 VP 4							NP 18 S 21 VP 18
2			P 2 V 5						PP 12 VP 16
3					Det 1				NP 10 N 8
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

36

CKY - Backtracking

time	1	flies	2	like	3	an	4	arrow	5
0									
1									
2									
3									
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

37

CKY - Backtracking

time	1	flies	2	like	3	an	4	arrow	5
0									
1									
2									
3									
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

38

CKY - Backtracking

time	1	flies	2	like	3	an	4	arrow	5
0									
1									
2									
3									
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

39

CKY - Backtracking

time	1	flies	2	like	3	an	4	arrow	5
0									
1									
2									
3									
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

40

CKY - Backtracking

time	1	flies	2	like	3	an	4	arrow	5
0									
1									
2									
3									
4									

- 1 S → NP VP
- 6 S → Vst NP
- 2 S → S PP
- 1 VP → V NP
- 2 VP → VP PP
- 1 NP → Det N
- 2 NP → NP PP
- 3 NP → NP NP
- 0 PP → P NP

Aula 20 - 20/11/2008

41

CKY Probabilístico

- Para obter a árvore de maior probabilidade basta fazer o peso w de cada regra ser $-\log p(\text{Regra} \mid \text{Lado-Esquerdo})$

$$W\left(\begin{array}{c} S \\ \swarrow \downarrow \\ NP \quad VP \\ \swarrow \downarrow \swarrow \downarrow \\ VP \quad flies \quad PP \\ \swarrow \downarrow \swarrow \downarrow \\ P \quad like \quad NP \\ \swarrow \downarrow \swarrow \downarrow \\ Det \quad an \quad N \end{array} \mid s\right) = W(S \rightarrow NP \quad VP) + W(NP \rightarrow flies \mid NP) + W(VP \rightarrow VP \quad NP) + W(VP \rightarrow flies) + \dots$$

Aula 20 - 20/11/2008

42

CKY Probabilístico

$$p(\begin{array}{c} S \\ \swarrow \searrow \\ NP \quad VP \\ \swarrow \searrow \swarrow \searrow \\ time \quad flies \quad PP \\ \swarrow \searrow \swarrow \searrow \\ like \quad Det \quad N \\ \swarrow \searrow \\ an \quad arrow \end{array} \mid s) = p(s \rightarrow NP \ VP \mid s) * p(NP \rightarrow flies \mid NP)$$

$$* p(VP \rightarrow VP \ NP \mid VP)$$

$$* p(VP \rightarrow flies \mid VP) * \dots$$

Aula 20 - 20/11/2008

43

Problemas com PCFGs

- Hipótese de independência entre as regras.
 - PCFG supõe que a probabilidade de expansão de uma regra é independente de seu contexto (isto é, de outras regras).
- Falta de dependência léxica.
 - PCFG só leva as palavras em consideração para as regras em que os símbolos terminais aparecem.

Aula 20 - 20/11/2008

44

Hipótese de independência entre as regras

- A expansão de um não-terminal é independente do seu contexto, isto é, de outros não-terminais próximos na árvore.
- Porém, em linguagem natural, a escolha de que nó deve ser expandido depende da posição do nó na árvore.
 - Por exemplo: em inglês NPs que são sujeitos tem maior probabilidade de serem pronomes, enquanto NPs que são objetos tem maior probabilidade de não serem pronomes.

Aula 20 - 20/11/2008

45

NP Sujeito vs. Pronome

	Pronoun	Non-Pronoun
Subject	91%	9%
Object	34%	66%

- Não podemos usar as probabilidades da tabela acima porque elas dependem do contexto. A PCFG seria:

$$NP \rightarrow DT \ NN \ .28$$

$$NP \rightarrow PRP \ .25$$

Aula 20 - 20/11/2008

46