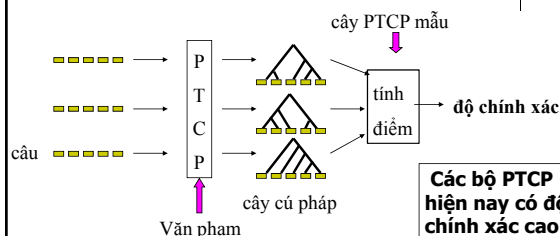


Phân tích cú pháp

Lê Thanh Hương
Bộ môn Hệ thống Thông tin
Viện CNTT & TT – Trường ĐHBKHN
Email: huonglt-fit@mail.hut.edu.vn

1

Bài toán PTCP



Các bộ PTCP hiện nay có độ chính xác cao
(Eisner, Collins, Charniak, etc.)

2

Khái niệm về văn phạm

- Phân tích câu “Bò vàng gặm cỏ non”
- Cây cú pháp:
- Tập luật
 - $C \rightarrow CN\ VN$
 - $CN \rightarrow DN$
 - $VN \rightarrow ĐgN$
 - $ĐgN \rightarrow ĐgT\ DN$
 - $DN \rightarrow DT\ TT$

3

Văn phạm

- Một văn phạm sản sinh là một hệ thống
- $G = (T, N, S, R)$, trong đó
- T (terminal) – tập ký hiệu kết thúc
- N (non terminal) – tập ký hiệu không kết thúc
- S (start) – ký hiệu khởi đầu
- R (rule) – tập luật
- $R = \{ \alpha \rightarrow \beta \mid \alpha, \beta \in (T \cup N)^* \}$
- $\alpha \rightarrow \beta$ gọi là luật sản xuất

4

Dạng chuẩn Chomsky

- Mọi NNPNC không chứa ϵ đều có thể sinh từ một văn phạm tndó mọi sản xuất đều có dạng $A \rightarrow BC$ hoặc $A \rightarrow a$, với $A, B, C \in N$ và $a \in T$
- Ví dụ: Tìm dạng chuẩn Chomsky cho văn phạm G với $T = \{a, b\}$, $N = \{S, A, B\}$, R như sau:
 - $S \rightarrow bA|aB$
 - $A \rightarrow bAA|aS|a$
 - $B \rightarrow aBB|bS|b$

5

Nhắc lại về văn phạm

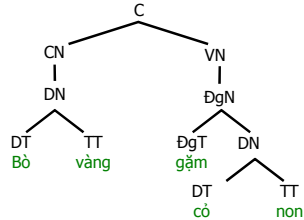
- Văn phạm: 1 tập luật viết lại
 - Ký hiệu kết thúc: các ký hiệu không thể phân rã được nữa.
 - Ký hiệu không kết thúc: các ký hiệu có thể phân rã được.
 - Xét văn phạm G :
 - $S \rightarrow NP\ VP$
 - $NP \rightarrow \text{John, garbage}$
 - $VP \rightarrow \text{laughed, walks}$
- G có thể sinh ra các câu sau:
- John laughed. John walks.*
Garbage laughed. Garbage walks.

6

Cấu trúc ngữ pháp

Cây cú pháp biểu diễn cấu trúc ngữ pháp của một câu.

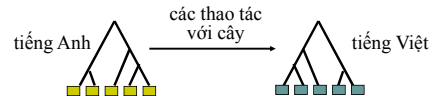
Bò vàng gặm cỏ non.



7

Các ứng dụng của PTCP

- Dịch máy (Alshawi 1996, Wu 1997, ...)



- Nhận dạng tiếng nói sử dụng PTCP (Chelba et al 1998)

Put the file in the folder.

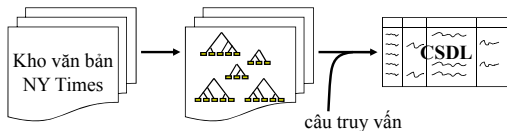
Put the file **and** the folder.

8

Các ứng dụng của PTCP

- Kiểm tra ngữ pháp (Microsoft)

- Trích rút thông tin (Hobbs 1996)



9

Văn phạm phi ngữ cảnh (Context-Free Grammar)

... còn gọi là văn phạm cấu trúc đoạn

- $G = \langle T, N, P, S, R \rangle$
 - T – tập các ký hiệu kết thúc (terminals)
 - N – tập các ký hiệu không kết thúc (non-terminals)
 - P – ký hiệu tiên kết thúc (preterminals), khi viết lại trở thành ký hiệu kết thúc
 - S – ký hiệu bắt đầu
 - R: $X \rightarrow \gamma$, X là ký hiệu không kết thúc; γ là chuỗi các ký hiệu kết thúc và không kết thúc (có thể rỗng)
- Văn phạm G sinh ra ngôn ngữ L
- Bộ nhận dạng: trả về **yes** hoặc **no**
- Bộ PTCP: trả về tập các cây cú pháp

10

- Văn phạm ngữ cấu:

- $\alpha \rightarrow \beta$, với $\alpha \in V^+$, $\beta \in V^*$

- Văn phạm cảm ngữ cảnh:

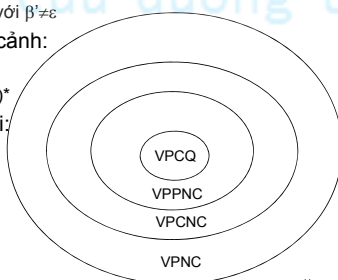
- $r = \alpha \rightarrow \beta$, với $\alpha \in V^+$, $\beta \in V^*$, $|\alpha| \leq |\beta|$
- và $\alpha_1 A \alpha_2 \rightarrow \alpha_1 \beta' \alpha_2$ với $\beta' \neq \epsilon$

- Văn phạm phi ngữ cảnh:

- $A \rightarrow \theta$, $A \in N$,
- với $\theta \in V^* = (T \cup N)^*$

- Văn phạm chính qui:

- $A \rightarrow aB$,
- $A \rightarrow Ba$,
- $A \rightarrow a$,
- với $A, B \in N$, $a \in T$.



11

Văn phạm phi ngữ cảnh

$S \rightarrow NP VP$	$DT \rightarrow the$
$NP \rightarrow \left\{ \begin{array}{l} DT NNS \\ DT NN \\ NP PP \end{array} \right\}$	$NNS \rightarrow \left\{ \begin{array}{l} children \\ students \\ mountains \end{array} \right\}$
$VP \rightarrow \left\{ \begin{array}{l} VP PP \\ VBD \end{array} \right\}$	$VBD \rightarrow \left\{ \begin{array}{l} slept \\ ate \\ saw \end{array} \right\}$
$PP \rightarrow IN NP$	$IN \rightarrow \left\{ \begin{array}{l} in \\ of \end{array} \right\}$
	$NN \rightarrow cake$

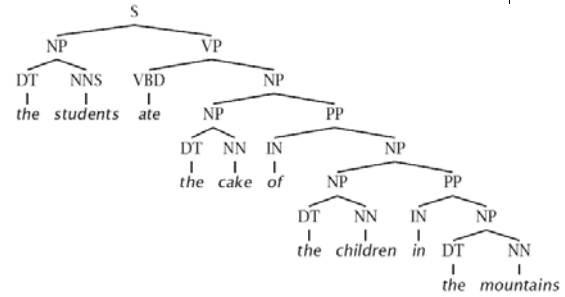
12

Áp dụng tập luật ngữ pháp

- S
 - NP VP
 - DT NNS VBD
 - *The children slept*
- S
 - NP VP
 - DT NNS VBD NP
 - DT NNS VBD DT NN
 - *The children ate the cake*

13

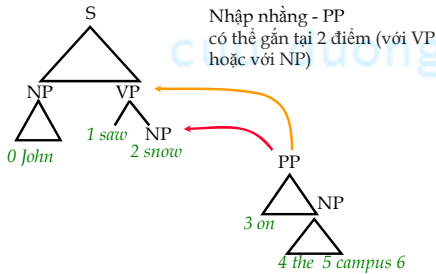
Cấu trúc đoạn đệ quy



14

Vấn phạm cho ngôn ngữ tự nhiên có nhập nhằng

John saw snow on the campus



15

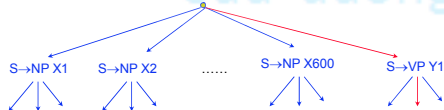
PTCP kiểu trên xuống

- Hướng đích
- Khởi đầu với 1 danh sách các ký hiệu cần triển khai (S, NP, VP, ...)
- Viết lại các đích trong tập đích bằng cách:
 - tìm luật có vẻ trái trùng với đích cần triển khai
 - triển khai nó với vẻ phải luật, tìm cách khớp với câu đầu vào
- Nếu 1 đích có nhiều cách viết lại → chọn 1 luật để áp dụng (bài toán tìm kiếm)
- Có thể sử dụng tìm kiếm rộng (breadth-first search) hoặc tìm kiếm sâu (depth-first search)

16

Khó khăn với PTCP trên xuống

- Các luật đệ quy trái
- PTCP trên xuống rất bất lợi khi có nhiều luật có cùng vẻ trái



- Nhiều thao tác thừa: triển khai tất cả các nút có thể phân tích trên xuống
- PTCP trên xuống sẽ làm việc tốt khi có chiến lược điều khiển ngữ pháp phù hợp
- PTCP trên xuống không thể triển khai các ký hiệu tiên kết thúc thành các ký hiệu kết thúc. Trên thực tế, người ta thường sử dụng phương pháp dưới lên để làm việc này.
- Lập lại công việc: bất cứ chỗ nào có cấu trúc giống nhau

17

PTCP dưới lên

- Hướng dữ liệu
- Khởi tạo với xâu cần phân tích
- Nếu chuỗi trong tập đích phù hợp với vẻ phải của 1 luật → thay nó bằng vẻ trái của luật
- Kết thúc khi tập đích = {S}.
- Nếu vẻ phải của các luật khớp với nhiều luật trong tập đích, cần lựa chọn luật áp dụng (bài toán tìm kiếm)
- Có thể sử dụng tìm kiếm rộng (breadth-first search) hoặc tìm kiếm sâu (depth-first search)

18

Khó khăn với PTCP dưới lên

- Không hiệu quả khi có nhiều nhập nhằng mức từ vựng
- Lặp lại công việc: bất cứ khi nào có cấu trúc con chung
- Cả PTCP TD (LL) và BU (LR) đều có độ phức tạp là hàm mũ của độ dài câu.

19

Thuật toán CKY (bộ nhận dạng)

- **Vào:** xâu n từ
- **Ra:** yes/no
- **Cấu trúc ngữ pháp:** bảng n x n (chart table)
 - hàng đánh số 0 đến n-1
 - cột đánh số 1 đến n
 - cell [i,j] liệt kê tất cả các nhãn cú pháp giữa i và j

20

Thuật toán CKY (bottom-up)

- for i := 1 to n
 - Thêm tất cả từ loại của từ thứ i vào ô [i-1,i]
- for width := 2 to n
 - for start := 0 to n-width
 - end := start + width
 - for mid := start+1 to end-1
 - for mọi nhãn cú pháp X trong [start,mid]
 - for mọi nhãn cú pháp Y trong [mid,end]
 - for mọi cách kết hợp X và Y (nếu có)
 - Thêm nhãn kết quả vào [start,end] nếu chưa có nhãn này

21

Ví dụ

	Bò	vàng	gặm	cỏ	non
	1	2	3	4	5
0	DT	→ DN	→	→	C
1		↑ TT			↑
2			DgT	→	DgN
3				DT	→ DN
4					↑ TT

22

Văn phạm phi ngữ cảnh

1. Start → S
2. S → NP VP
3. NP → Det Noun
4. NP → Name
5. NP → Name PP
6. PP → Prep NP
7. VP → V NP
8. VP → V NP PP
9. V → ate
10. Name → John
11. Name → ice-cream, snow
12. Noun → ice-cream, pizza
13. Noun → table, guy, campus
14. Det → the
15. Prep → on

23

Luật kết hợp

- Ô Cell[i,j] chứa nhãn X nếu
 - Có luật $X \rightarrow YZ$;
 - Cell[i,k] chứa nhãn Y và ô Cell[k,j] chứa nhãn Z, với k nằm giữa i và j;
- VD: NP → DT [0,1] NN[1,2]

24

CKY phải sử dụng luật nhị phân

- Chuyển $VP \rightarrow V \ NP \ PP$ thành:
 - 8.a. $VP \rightarrow V \text{ Arguments}$
 - 8.b. $\text{Arguments} \rightarrow NP \ PP$

25

CKY chart

"The guy ate the ice-cream on the table"

	1	2	3	4	5	6	7	8
0	DT							
1		NN						
2			VBD					
3				DT				
4					NN			
5						IN		
6							DT	
7								NN

26

Áp dụng thao tác 'dán'

	1	2	3	4	5	6	7	8
0	DT	NP						
1		NN						
2			VBD					
3				DT				
4					NN			
5						IN		
6							DT	
7								NN

27

Nhập nhằng!

- 5. $NP \rightarrow NN \ PP$
- 8.a. $VP \rightarrow V \text{ Arguments}$
- 8.b. $\text{Arguments} \rightarrow NP \ PP$

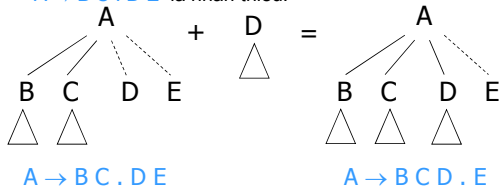
	1	2	3	4	5	6	7	8
0	DT	NP						S
1		NN						VP
2			VBD					NP
3				DT	NP			Arg
4				NN				PP
5						IN		DT
6							DT	NN
7								

28

Thuật toán Earley (top-down)

- Tìm các nhãn và các nhãn thiếu (partial constituents) từ đầu vào

- $A \rightarrow BC.DE$ là nhãn thiếu:



- Tiến hành dần từ trái sang phải

29

Ví dụ

ROOT $\rightarrow S$
 S $\rightarrow NP \ VP$
 NP $\rightarrow Det \ N$
 NP $\rightarrow NP \ PP$
 VP $\rightarrow V \ PP$
 VP $\rightarrow V \ NP$
 PP $\rightarrow P \ NP$
 NP $\rightarrow Papa$
 N $\rightarrow caviar$
 N $\rightarrow spoon$
 V $\rightarrow ate$
 P $\rightarrow with$
 Det $\rightarrow the$
 Det $\rightarrow a$

30

Thuật toán Earley

- Thuật toán Earley giống thuật toán đệ qui nói trên, nhưng giải quyết được vấn đề đệ qui trái.
- Sử dụng bảng phân tích giống thuật toán CKY, nhằm lưu lại các thông tin đã tìm thấy → lập trình động “**Dynamic programming.**”

Các thao tác của thuật toán

- Xử lý phần đi sau dấu . theo kiểu đệ qui :
 - Nếu là từ, quét (**scan**) đầu vào để xem có phù hợp không
 - Nếu là ký hiệu không kết thúc, đoán (**predict**) các khả năng để khớp nó (giảm số phép tiên đoán bằng cách nhìn trước k ký hiệu từ đầu vào và chỉ sử dụng các luật phù hợp với k ký hiệu đó)
 - Nếu rỗng, ta đã hoàn thành một thành phần ngữ pháp, gắn (**attach**) nó vào những chỗ liên quan

37

0	
0 ROOT . S	khởi tạo

tương đương với $(0, \text{ROOT} \rightarrow . S)$

38

0	
0 ROOT . S	
0 S . NP VP	predict luật có vế trái là S

$(0, S \rightarrow . NP VP)$

39

0	
0 ROOT . S	
0 S . NP VP	
0 NP . Det N	predict luật có VT = NP
0 NP . NP PP	(có 3 luật phù hợp)
0 NP . Papa	

40

0	
0 ROOT . S	
0 S . NP VP	
0 NP . Det N	
0 NP . NP PP	
0 NP . Papa	
0 Det . the	predict luật có VT = Det (2 luật)
0 Det . a	

41

0	
0 ROOT . S	
0 S . NP VP	
0 NP . Det N	
0 NP . NP PP	
0 NP . Papa	
0 Det . the	predict luật có VT = NP
0 Det . a	ta đã làm việc này ở bước trước, vì vậy không làm lại
	Chú ý: ta phải làm lại việc này với luật đệ qui trái

42

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP		
0 NP . Det N		
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

scan: từ phù hợp từ đầu vào

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP		
0 NP . Det N		
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

scan: không phù hợp

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP		
0 NP . Det N		
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

scan: không phù hợp

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP	0 S NP . VP	
0 NP . Det N	0 NP NP . PP	
0 NP . NP PP		
0 NP . Papa		
0 Det . the		
0 Det . a		

attach NP mới tạo (bắt đầu từ 0) với các phần liên quan (các phần chưa hoàn thành kết thúc tại 0 và có NP sau đầu .)

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP	0 S NP . VP	
0 NP . Det N	0 NP NP . PP	
0 NP . NP PP	1 VP . V NP	
0 NP . Papa	1 VP . VP PP	
0 Det . the		
0 Det . a		

predict

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP	0 S NP . VP	
0 NP . Det N	0 NP NP . PP	
0 NP . NP PP	1 VP . V NP	
0 NP . Papa	1 VP . VP PP	
0 Det . the	1 PP . P NP	
0 Det . a		

predict

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP	0 S NP . VP	
0 NP . Det N	0 NP NP . PP	
0 NP . NP PP	1 VP . V NP	predict
0 NP . Papa	1 VP . VP PP	
0 Det . the	1 PP . P NP	
0 Det . a	1 V . ate	

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP	0 S NP . VP	
0 NP . Det N	0 NP NP . PP	
0 NP . NP PP	1 VP . V NP	
0 NP . Papa	1 VP . VP PP	predict
0 Det . the	1 PP . P NP	
0 Det . a	1 V . ate	

0	Papa	1
0 ROOT . S	0 NP Papa .	
0 S . NP VP	0 S NP . VP	
0 NP . Det N	0 NP NP . PP	
0 NP . NP PP	1 VP . V NP	
0 NP . Papa	1 VP . VP PP	
0 Det . the	1 PP . P NP	predict
0 Det . a	1 V . ate	
	1 P . with	

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP			
0 NP . Det N	0 NP NP . PP			
0 NP . NP PP	1 VP . V NP			
0 NP . Papa	1 VP . VP PP			
0 Det . the	1 PP . P NP			
0 Det . a	1 V . ate	scan: thành công!		
	1 P . with			

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP			
0 NP . Det N	0 NP NP . PP			
0 NP . NP PP	1 VP . V NP			
0 NP . Papa	1 VP . VP PP			
0 Det . the	1 PP . P NP			
0 Det . a	1 V . ate			
	1 P . with	scan: không hợp		

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP			
0 NP . NP PP	1 VP . V NP			
0 NP . Papa	1 VP . VP PP			
0 Det . the	1 PP . P NP			
0 Det . a	1 V . ate			
	1 P . with			

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP			
0 Det . a	1 V . ate			
	1 P . with			

predict

55

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP	2 Det . the		
0 Det . a	1 V . ate	2 Det . a		
	1 P . with			

predict (các bước sau tương tự)

56

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP	2 Det . the		
0 Det . a	1 V . ate	2 Det . a		
	1 P . with			

predict

57

0	Papa	1	ate	2
0 ROOT . S	0 NP Papa .	1 V ate .		
0 S . NP VP	0 S NP . VP	1 VP V . NP		
0 NP . Det N	0 NP NP . PP	2 NP . Det N		
0 NP . NP PP	1 VP . V NP	2 NP . NP PP		
0 NP . Papa	1 VP . VP PP	2 NP . Papa		
0 Det . the	1 PP . P NP	2 Det . the		
0 Det . a	1 V . ate	2 Det . a		
	1 P . with			

scan (lúc này thất bại vì Papa không phải là từ tiếp theo)

58

0	Papa	1	ate	2	the	3
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .			
0 S . NP VP	0 S NP . VP	1 VP V . NP				
0 NP . Det N	0 NP NP . PP	2 NP . Det N				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa				
0 Det . the	1 PP . P NP	2 Det . the				
0 Det . a	1 V . ate	2 Det . a				
	1 P . with					

scan: thành công!

59

0	Papa	1	ate	2	the	3
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .			
0 S . NP VP	0 S NP . VP	1 VP V . NP				
0 NP . Det N	0 NP NP . PP	2 NP . Det N				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa				
0 Det . the	1 PP . P NP	2 Det . the				
0 Det . a	1 V . ate	2 Det . a				
	1 P . with					

60

0	Papa	1	ate	2	the	3
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .			
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N			
0 NP . Det N	0 NP NP . PP	2 NP . Det N				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa				
0 Det . the	1 PP . P NP	2 Det . the				
0 Det . a	1 V . ate	2 Det . a				
	1 P . with					

61

0	Papa	1	ate	2	the	3
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .			
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N			
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar			
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon			
0 NP . Papa	1 VP . VP PP	2 NP . Papa				
0 Det . the	1 PP . P NP	2 Det . the				
0 Det . a	1 V . ate	2 Det . a				
	1 P . with					

62

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N					
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar					
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon					
0 NP . Papa	1 VP . VP PP	2 NP . Papa						
0 Det . the	1 PP . P NP	2 Det . the						
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

63

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N					
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar					
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon					
0 NP . Papa	1 VP . VP PP	2 NP . Papa						
0 Det . the	1 PP . P NP	2 Det . the						
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

64

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det . N				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar					
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon					
0 NP . Papa	1 VP . VP PP	2 NP . Papa						
0 Det . the	1 PP . P NP	2 Det . the						
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

65

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det . N				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa						
0 Det . the	1 PP . P NP	2 Det . the						
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

66

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP				
0 Det . a	1 V . ate	2 Det . a						
	1 P . with							

attach

67

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP				
	1 P . with							

68

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP				
	1 P . with			0 ROOT S .				

attach

69

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP				
	1 P . with			0 ROOT S .				

70

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP				
	1 P . with			0 ROOT S .				
				4 P . with				

71

0	Papa	1	ate	2	the	3	caviar	4
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP				
	1 P . with			0 ROOT S .				
				4 P . with				

72

[illegible][illegible]



[illegible][illegible][illegible]

0	Papa	1	ate	2	the	3	caviar	4	with a spoon	
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						91				

0	Papa	1	ate	2	the	3	caviar	4	with a spoon	
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						92				

0	Papa	1	ate	2	the	3	caviar	4	with a spoon	
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						93				

0	Papa	1	ate	2	the	3	caviar	4	with a spoon	
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						0 ROOT S .				

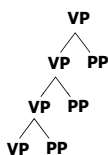
0	Papa	1	ate	2	the	3	caviar	4	with a spoon	
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						0 ROOT S .				

0	Papa	1	ate	2	the	3	caviar	4	with a spoon	
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .				
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .				
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .				
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP				
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .				
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .				
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP				
	1 P . with			0 ROOT S .		1 VP V NP .				
				4 P . with		2 NP NP . PP				
						0 S NP VP .				
						1 VP VP . PP				
						7 P . with				
						0 ROOT S .				

0	Papa	1	ate	2	the	3	caviar	4	with a spoon	5	...	6	N spoon
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .							
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .							
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .							
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP							
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .							
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .							
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP							
	1 P . with			0 ROOT S .		1 VP V NP .							
				4 P . with		2 NP NP . PP							
						0 S NP VP .							
						1 VP VP . PP							
						7 P . with							
						0 ROOT S .							

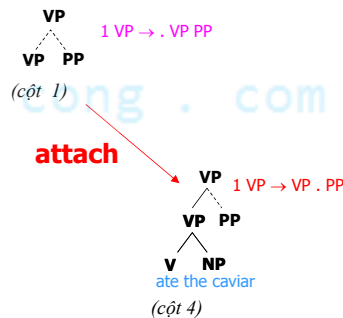
0	Papa	1	ate	2	the	3	caviar	4	with a spoon	5	...	6	N spoon
0 ROOT . S	0 NP Papa .	1 V ate .	2 Det the .	3 N caviar .	...	6 N spoon .							
0 S . NP VP	0 S NP . VP	1 VP V . NP	2 NP Det . N	2 NP Det N .		5 NP Det N .							
0 NP . Det N	0 NP NP . PP	2 NP . Det N	3 N . caviar	1 VP V NP .		4 PP P NP .							
0 NP . NP PP	1 VP . V NP	2 NP . NP PP	3 N . spoon	2 NP NP . PP		5 NP NP . PP							
0 NP . Papa	1 VP . VP PP	2 NP . Papa		0 S NP VP .		2 NP NP PP .							
0 Det . the	1 PP . P NP	2 Det . the		1 VP VP . PP		1 VP VP PP .							
0 Det . a	1 V . ate	2 Det . a		4 PP . P NP		7 PP . P NP							
	1 P . with			0 ROOT S .		1 VP V NP .							
				4 P . with		2 NP NP . PP							
						0 S NP VP .							
						1 VP VP . PP							
						7 P . with							
						0 ROOT S .							

Vấn đề với PTCP trên xuống: đệ quy trái

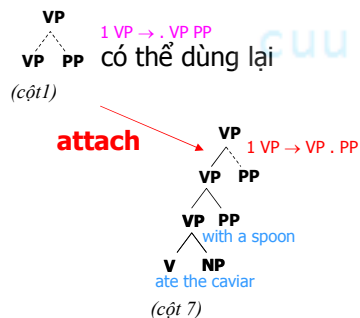


gắn liên tục các luật mới vào cây trước khi thấy PPs
→ cần đoán trước số PP cần ở đầu vào

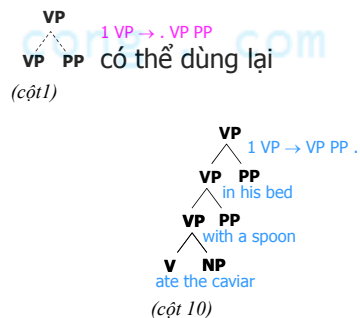
... nhưng thuật toán Earley Ok!



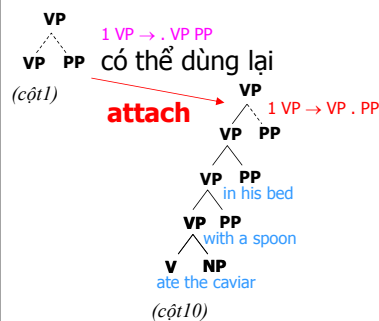
... nhưng thuật toán Earley Ok!



... nhưng thuật toán Earley Ok!



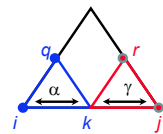
... nhưng thuật toán Earley Ok!



103

Phục hồi cây cú pháp

$[s, i]$ trong tập trạng thái j Sử dụng thuật toán dùng queue đơn giản dựa trên các thành phần có ích



- 1 thành phần ở trạng thái kết thúc là *có ích*
- If $s = [A \rightarrow \alpha \bullet B, i]$ trong tập đích k & *có ích*
- then $q = [A \rightarrow \alpha B \bullet, k]$ & item $r = [B \rightarrow \gamma \bullet, j]$ là *có ích*

$[s, i]$: một thành phần với luật s & trả về con trỏ i .

Đánh dấu tất cả các thành phần trong tập trạng thái S_n ở dạng $Start \rightarrow \alpha S_n \bullet, 0$

for $j = n$ downto 0 do
for $i = 0$ to j do
for mọi bộ đã đánh dấu $[s, i]$ trong tập trạng thái j do
for $k = i$ to j do
if $[q, i] \in S_k$ & $[r, k] \in S_j$ & $s = q \otimes r$ then
đánh dấu $[q, i]$ và $[r, k]$

104

Ưu điểm

- Thuật toán Earley thực hiện một vài phép lọc *top-down*: bất cứ thành phần nào (state, or triple) được đưa vào tập trạng thái cần tương thích với phần đã được sinh ra ở bên trái. Ví dụ: $S \xrightarrow{*} w_i$ trong đó w_i là phần của câu đã được duyệt qua



105

Nhược điểm

- Biểu diễn luật: Explicit representation of rules: wastes time building them.
- Thực hiện phép lọc bên trái nhưng không lọc bên phải

Phép lọc nhìn trước cho ký hiệu không kết thúc A :

$FIRST(A) = \{x | A \Rightarrow x\delta\}$, $x = 1$ token

v.d., $FIRST(S) = \text{who, did, the, etc.}$

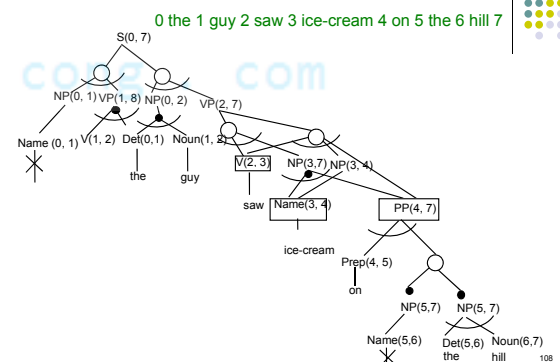
106

Các phương pháp khác

- Các phương pháp khác ứng với các cách khác nhau để tìm các đoạn
- Đoạn $X[i, j]$ là đoạn có nhãn X phủ đầu vào từ i đến j
Example:
 $John_1 \text{ ate } ice-cream_2 \text{ on } the_3 \text{ table}_4$
 $PP[3,6]; S[0,6]; \dots$
- Biểu diễn không gian tìm kiếm như cây and-or
 - Disjuncts (or) = các đường phân tích khác nhau
 - Conjuncts (and) = về phải của luật, ví dụ về phải của S là $NP VP$

107

PTCP là việc tìm kiếm

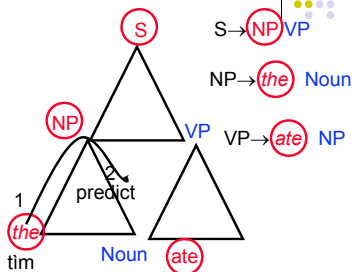


108

PTCP góc trái (Left-corner parsing)

- Nhìn từ dưới lên để tìm ký hiệu đầu tiên (left-corner) của đoạn, sau đó phân tích phần còn lại theo kiểu trên xuống

- Tìm cách kết hợp các đặc trưng tốt nhất của tìm phân tích trên xuống và dưới lên



Phương pháp này làm việc tốt với ngôn ngữ với thành phần quan trọng đặt ở đầu như tiếng Anh. Các tiếng Đức, Hà Lan, Nhật là ngôn ngữ có phần quan trọng đặt cuối.

cuu duong than cong . com

cuu duong than cong . com