ĐẠI HỌC BÁCH KHOA HÀ NỘI
VIỆN CÔNG NGHỆ THÔNG TIN VÀ TRUYỀN THÔNG

Phân tích cú pháp xác suất

Viện Công nghệ Thông tin và Truyền thông

Làm cách nào chọn cây đúng?

- Ví dụ:

  I saw a man with a telescope.

- Khi số luật tăng, khả năng nhập nhằng tăng
- Tập luật NYU: bộ PTCP Apple pie : 20,000-30,000 luật cho tiếng Anh
- Lựa chọn luật AD: V DT NN PP

  (1) VP $\rightarrow$ V NP PP

  NP $\rightarrow$ DT NN

  (2) VP $\rightarrow$ V NP

  NP $\rightarrow$ DT NN PP

Kết hợp từ (bigrams pr)

Ví dụ:

Eat ice-cream (high freq)

Eat John (low, except on Survivor)

Nhược điểm:

- P(John decided to bake a) có xác suất cao
- Xét:

$P(w_3) = P(w_3|w_2w_1)=P(w_3|w_2)P(w_2|w_1)P(w_1)$

Giả thiết này quá mạnh: chủ ngữ có thể quyết định bổ ngữ trong câu

Clinton admires honesty

➢ sử dụng cấu trúc ngữ pháp để dừng việc lan truyền

- Xét Fred watered his mother's small garden. Từ *garden* có ảnh hưởng như thế nào?
  - Pr(*garden*|*mother's small*) thấp ⇒ mô hình trigram không tốt
  - Pr(*garden* | X là thành phần chính của bổ ngữ cho động từ *to water*) cao hơn

➢ sử dụng bigram + quan hệ ngữ pháp

Kết hợp từ (bigrams pr)

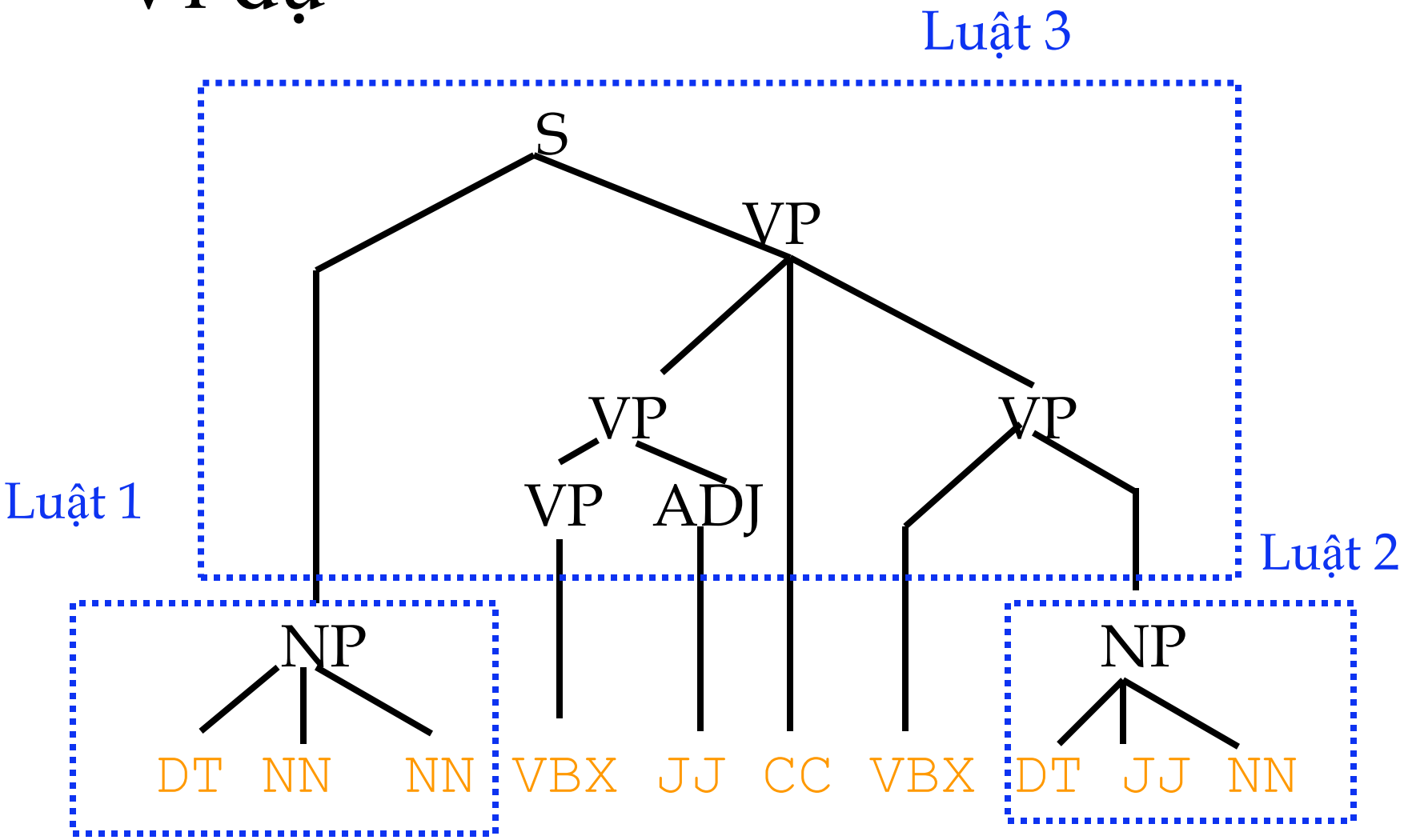
- V có một số loại bổ ngữ nhất định
  $\Rightarrow$ Verb-with-obj, verb-without-obj
- Sự tương thích giữa chủ ngữ và bổ ngữ:
  John admires honesty
  Honesty admires John ???

Nhược điểm:

- Kích thước tập ngữ pháp tăng
- Các bài báo của tạp chí Wall Street Journal trong 1 năm: 47,219 câu, độ dài trung bình 23 từ, gán nhãn bằng tay: chỉ có 4.7% hay 2,232 câu có cùng cấu trúc ngữ pháp

➢ Không thể dựa trên việc tìm các cấu trúc cú pháp đúng cho cả câu. Phải xây dựng tập các mẫu ngữ pháp nhỏ

Ví dụ


Luật 3
S
VP
VP
VP
VP
ADJ
Luật 1
Luật 2
NP
NP
DT NN NN
VBX JJ CC VBX
DT JJ NN
This apple pie looks good and is a real treat

Luật

1. NP→DT NN NN
2. NP→DT JJ NN
3. S→NP VBX JJ CC VBX NP

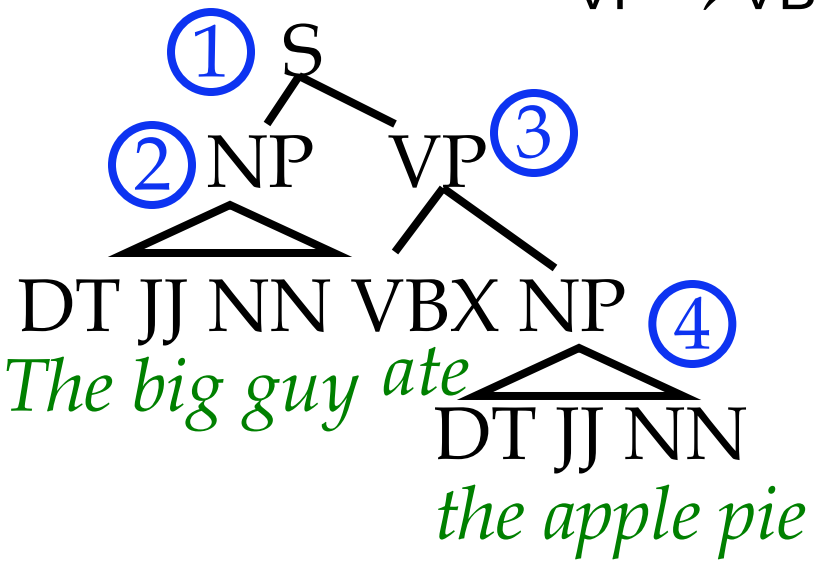
- Nhóm (NNS, NN) thành NX; (NNP, NNPs)=NPX; (VBP, VBZ, VBD)=VBX;
- Chọn các luật theo tần suất của nó

Tính xác suất

$$\Pr(X \rightarrow Y) = \frac{X \Rightarrow NP}{Y \quad NP} = \frac{DT\ JJ\ NN}{NP} = \frac{1470}{9711} = 0.1532$$

Tính Pr

S → NP VP; 0.35
NP → DT JJ NN; 0.1532
VP → VBX NP; 0.302



Luật áp dụng	*Chuỗi Pr*
1 S →NP VP	0.35
2 NP → DT JJ NN	0.1532 x 0.35 = 0.0536
3 VP → VBX NP	0.302 x 0.0536= 0.0162
4 NP → DT JJ NN	0.1532 x 0.0162=0.0025

Pr = 0.0025

Văn phạm phi ngữ cảnh xác suất

- 1 văn phạm phi ngữ cảnh xác suất (Probabilistic Context Free Grammar) gồm các phần thông thường của CFG
- Tập ký hiệu kết thúc $\{w^k\}$, $k = 1, \ldots, V$
- Tập ký hiệu không kết thúc $\{N^i\}$, $i = 1, \ldots, n$
- Ký hiệu khởi đầu N^1
- Tập luật $\{N^i \rightarrow \zeta^j\}$, ζ^j là chuỗi các ký hiệu kết thúc và không kết thúc
- Tập các xác suất của 1 luật là:

$$\forall i \sum_j P(N^i \rightarrow \zeta^j) = 1$$

- Xác suất của 1 cây cú pháp:

$$P(T) = \Pi_{i=1..n} p(r(i))$$

Các giả thiết

- ***Độc lập vị trí:*** Xác suất 1 cây con không phụ thuộc vào vị trí của các từ của cây con đó ở trong câu

  $\forall k, P(N_{jk}(k+c) \rightarrow \zeta)$ là giống nhau

- ***Độc lập ngữ cảnh:*** Xác suất 1 cây con không phụ thuộc vào các từ ngoài cây con đó

  $P(N_{jkl} \rightarrow \zeta |$ các từ ngoài khoảng k đến l) = $P(N_{jkl} \rightarrow \zeta)$

- ***Độc lập tổ tiên:*** Xác suất 1 cây con không phụ thuộc vào các nút ngoài cay con đó

  $P(N_{jkl} \rightarrow \zeta |$ các nút ngoài cây con N_{jkl}) = $P(N_{jkl} \rightarrow \zeta)$