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

Gán nhãn từ loại

Viện Công nghệ Thông tin và Truyền thông

Định nghĩa

- Gán nhãn từ loại (Part of Speech tagging - POS tagging): mỗi từ trong câu được gán nhãn thẻ từ loại tương ứng của nó
  - Vào : 1 đoạn văn bản đã tách từ + tập nhãn
  - Ra: cách gán nhãn chính xác nhất

Ví dụ 1
Ví dụ 2
Ví dụ 3
Ví dụ 4
Ví dụ 5

➢ Gán nhãn làm cho việc phân tích văn bản dễ dàng hơn

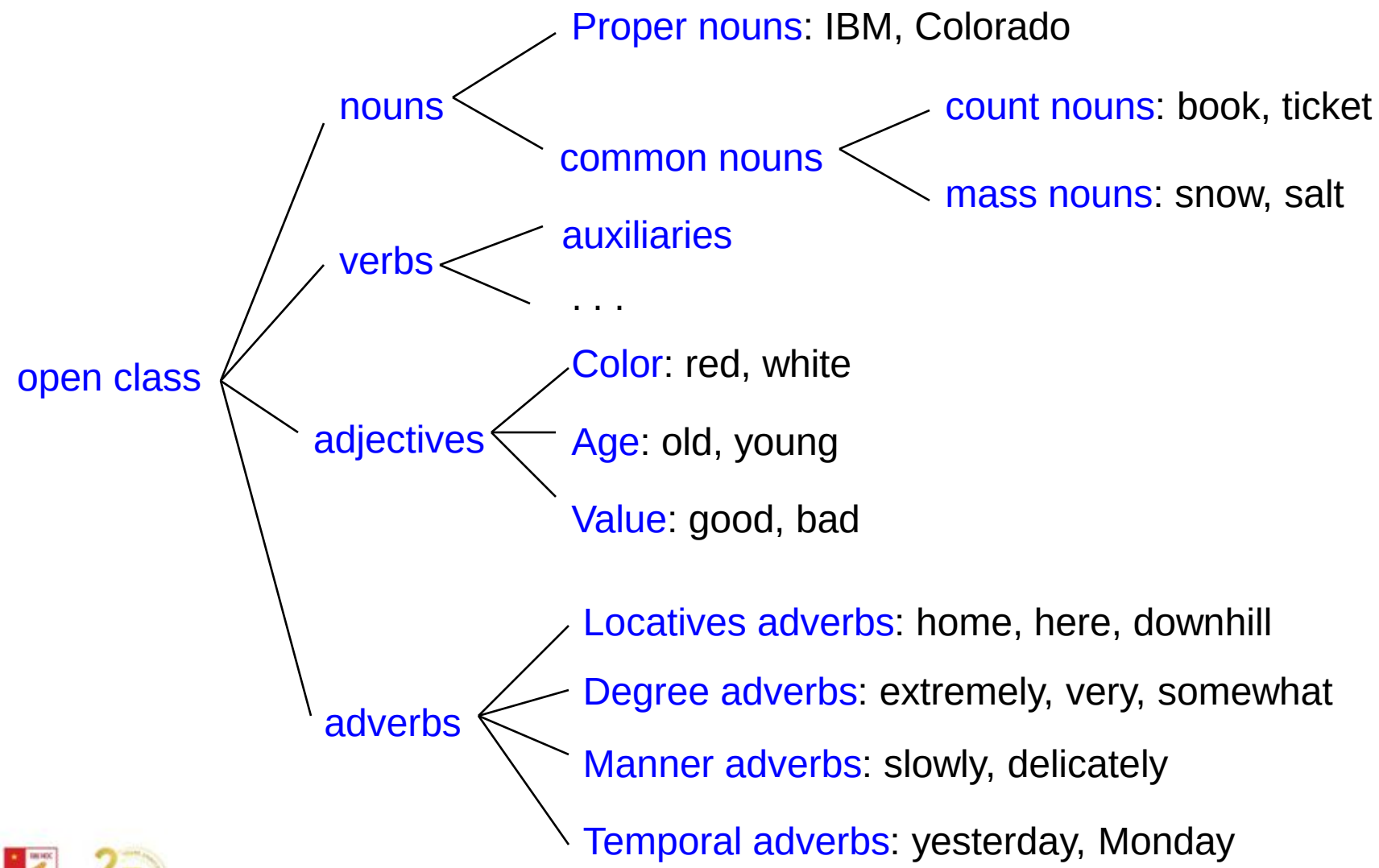
Tại sao cần gán nhãn?

- **Dễ thực hiện**: có thể thực hiện bằng nhiều phương pháp khác nhau
  - Các phương pháp sử dụng ngữ cảnh có thể đem lại kết quả tốt
  - Mặc dù nên thực hiện bằng phân tích văn bản
- **Các ứng dụng**:
  - Text-to-speech: record - N: [‘reko:d], V: [ri’ko:d]; lead – N [led], V: [li:d]
  - Tiền xử lý cho PTCP. PTCP thực hiện việc gán nhãn tốt hơn nhưng đắt hơn
  - Nhận dạng tiếng nói, PTCP, tìm kiếm, v.v…
- **Dễ đánh giá** *(có bao nhiêu thẻ được gán nhãn đúng?)*

Tập từ loại tiếng Anh

- **Lớp đóng** (các từ chức năng): số lượng cố định
  - Giới từ (Prepositions): on, under, over,…
  - Tiểu từ (Particles): abroad, about, around, before, in, instead, since, without,…
  - Mạo từ (Articles): a, an, the
  - Liên từ (Conjunctions): and, or, but, that,…
  - Đại từ (Pronouns): you, me, I, your, what, who,…
  - Trợ động từ (Auxiliary verbs): can, will, may, should,…
- **Lớp mở**: có thể có thêm từ mới

Lớp từ mở trong tiếng Anh



Tập nhãn cho tiếng Anh

- tập ngữ liệu Brown: 87 nhãn
- 3 tập thường được sử dụng:
  - Nhỏ: 45 nhãn - Penn treebank (slide sau)
  - Trung bình: 61 nhãn, British national corpus
  - Lớn: 146 nhãn, C7

Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*
CD	Cardinal number	*one, two, three*
DT	Determiner	*a, the*
EX	Existential 'there'	*there*
FW	Foreign word	*mea culpa*
IN	Preposition/sub-conj	*of, in, by*
JJ	Adjective	*yellow*
JJR	Adj., comparative	*bigger*
JJS	Adj., superlative	*wildest*
LS	List item marker	*1, 2, One*
MD	Modal	*can, should*
NN	Noun, sing. or mass	*llama*
NNS	Noun, plural	*llamas*
NNP	Proper noun, singular	*IBM*
NNPS	Proper noun, plural	*Carolinas*
PDT	Predeterminer	*all, both*
POS	Possessive ending	*'s*
PP	Personal pronoun	*I, you, he*
PP$	Possessive pronoun	*your, one's*
RB	Adverb	*quickly, never*
RBR	Adverb, comparative	*faster*
RBS	Adverb, superlative	*fastest*
RP	Particle	*up, off*

Tag	Description	Example
SYM	Symbol	+,%, &
TO	"to"	*to*
UH	Interjection	*ah, oops*
VB	Verb, base form	*eat*
VBD	Verb, past tense	*ate*
VBG	Verb, gerund	*eating*
VBN	Verb, past participle	*eaten*
VBP	Verb, non-3sg pres	*eat*
VBZ	Verb, 3sg pres	*eats*
WDT	Wh-determiner	*which, that*
WP	Wh-pronoun	*what, who*
WP$	Possessive wh-	*whose*
WRB	Wh-adverb	*how, where*
$	Dollar sign	*$*
#	Pound sign	*#*
"	Left quote	(' or ")
"	Right quote	(' or ")
(	Left parenthesis	([, (, {, <)
)	Right parenthesis	( ],), }, >)
,	Comma	,
.	Sentence-final punc	(. ! ?)
:	Mid-sentence punc	(: ; ... – -)

Tag	Description	Example
CC	Coordin. Conjunction	*and, but, or*
CD	Cardinal number	*one, two, three*
DT	Determiner	*a, the*
EX	Existential 'there'	*there*
FW	Foreign word	*mea culpa*
IN	Preposition/sub-conj	*of, in, by*
JJ	Adjective	*yellow*
JJR	Adj., comparative	*bigger*
JJS	Adj., superlative	*wildest*
LS	List item marker	*1, 2, One*
MD	Modal	*can, should*
NN	Noun, sing. or mass	*llama*
NNS	Noun, plural	*llamas*
NNP	Proper noun, singular	*IBM*
NNPS	Proper noun, plural	*Carolinas*
PDT	Predeterminer	*all, both*
POS	Possessive ending	*'s*
PP	Personal pronoun	*I, you, he*
PP$	Possessive pronoun	*your, one's*
RB	Adverb	
RBR	Adverb, co	
RBS	Adverb, su	
RP	Particle	

Tag	Description	Example
SYM	Symbol	+,%, &
TO	"to"	*to*
UH	Interjection	*ah, oops*
VB	Verb, base form	*eat*
VBD	Verb, past tense	*ate*
VBG	Verb, gerund	*eating*
VBN	Verb, past participle	*eaten*
VBP	Verb, non-3sg pres	*eat*
VBZ	Verb, 3sg pres	*eats*
WDT	Wh-determiner	*which, that*
WP	Wh-pronoun	*what, who*
WP$	Possessive wh-	*whose*
WRB	Wh-adverb	*how, where*
$	Dollar sign	*$*
#	Pound sign	*#*
"	Left quote	(' or ")
"	Right quote	(' or ")
(	Left parenthesis	([, (, {, <)
)	Right parenthesis	( ],), }, >)

I know that blocks the sun.
He always books the violin concert tickets early.
He says that book is interesting.

Penn Treebank – ví dụ

- The grand jury commented on a number of other topics.

⇨ The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

Khó khăn trong gán nhãn từ loại?

… là xử lý nhập nhằng

Các phương pháp gán nhãn từ loại

- **Dựa trên xác suất**: dựa trên xác suất lớn nhất, dựa trên mô hình Markov ẩn (hidden markov model – HMM)

  Pr (Det-N) > Pr (Det-Det)

- **Dựa trên luật**

  If <mẫu>

  Then … <gán nhãn thẻ từ loại>

Các cách tiếp cận

- **Sử dụng HMM :** "Sử dụng tất cả thông tin đã có và đoán"
- **Dựa trên chuyển đổi:** "Đoán trước, sau đó có thể thay đổi"

=> Có thể dựa trên ràng buộc ngữ pháp để loại trừ những khả năng sai"

Gán nhãn dựa trên xác suất

Cho câu hoặc 1 xâu các từ, gán nhãn từ loại thường xảy ra nhất cho các từ trong xâu đó.

Cách thực hiện:

- Hidden Markov model (HMM):

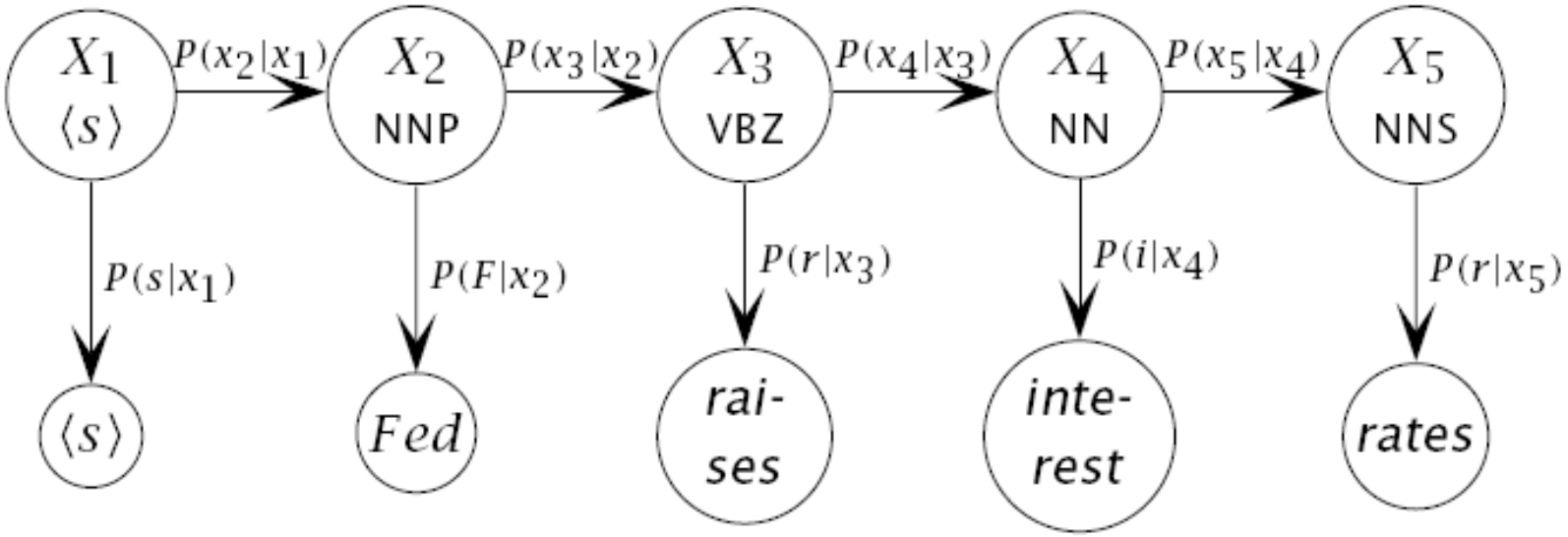
  Chọn thẻ từ loại làm tối đa xác suất:

  P(từ|từ loại)•P(từ loại| n từ loại phía trước)

The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

$\Rightarrow$ P(jury|NN) = 1/2

Ví dụ -HMMs



Thực hiện học có giám sát, sau đó suy diễn để xác định thẻ từ loại

Gán nhãn HMM

- **Công thức Bigram HMM**: chọn t_i cho w_i có nhiều khả năng nhất khi biết t_{i-1} và w_i :

  $t_i = \text{argmax}_j \, P(t_j \mid t_{i-1}, w_i)$ (1)

- **Giả thiết đơn giản hóa HMM:** vấn đề gán nhãn có thể giải quyết bằng cách dựa trên các từ và thẻ từ loại bên cạnh nó

$$t_i = \text{argmax}_j \, P(t_j \mid t_{i-1}) P(w_i \mid t_j) \quad (2)$$

xs chuỗi thẻ
(các thẻ đồng xuất hiện)

xs từ thường xuất hiện với thẻ tj

Ví dụ

1. Secretariat/NNP is/VBZ expected/VBN to/TO race/VB tomorrow/NN
2. People/NNS continue/VBP to/TO inquire/VB the/DT reason/NN for/IN the/DT race/NN for/IN outer/JJ space/NN

- Không thể đánh giá bằng cách chỉ đếm từ trong tập ngữ liệu (và chuẩn hóa)
- Muốn 1 động từ theo sau TO nhiều hơn 1 danh từ (*to race, to walk*). Nhưng 1 danh từ cũng có thể theo sau TO (*run to school*)

Giả sử chúng ta có tất cả các từ loại trừ từ race

- Chỉ nhìn vào từ đứng trước(bigram):
  to/TO race/??? NN or VB?
  the/DT race/???
- Áp dụng (2): $t_i = \text{argmax}_j P(t_j \mid t_{i-1}) P(w_i \mid t_j)$
- Chọn thẻ có xác suất lớn hơn giữa 2 xác suất:
  P(VB|TO)P(race|VB) hoặc P(NN|TO)P(race|NN)

  xác suất của 1 từ là race khi biết từ loại là VB.

I/PP know/VBP that/WDT block/NN blocks/NNS?VBZ? the/DT sun/NN.

Tính xác suất

Xét *P*(VB|TO) và *P*(NN|TO)

- Từ tập ngữ liệu Brown

  P(NN|TO)= .021

  P(VB|TO)= .340

  P(race|NN)= 0.00041

  P(race|VB)= 0.00003

- *P*(VB|TO)*P*(race|VB) = 0.00001
- *P*(NN|TO) *P*(race|NN) = 0.000007

➢ *race cần phải là động từ nếu đi sau* “TO”

Bài tập

$$t_i = \text{argmax}_j \, P(t_j \mid t_{i-1}) P(w_i \mid t_j)$$

- I know that blocks the sun.
- He always books the violin concert tickets early.
- He says that book is interesting.

- I/PP know/VBP that/WDT blocks/NNS block/VBP the/DT sun/NN.
- I/PP know/VBP that/WDT blocks/VBZ the/DT sun/NN.
- He/PP always/RB books/VBZ the/DT violin/NN concert/NN tickets/NNS early/RB.
- He/PP says/VBZ that/WDT book/NN is/VBZ interesting/JJ.

- I know that block blocks the sun.
- I/PP know/VBP that/DT block/NN blocks/NNS?VBZ? the/DT sun/NN.
- I/PP know/VBP that/WDT block/NN blocks/VBZ the/DT sun/NN.

Mô hình đầy đủ

- Chúng ta cần tìm chuỗi thẻ tốt nhất cho toàn xâu
- Cho xâu từ *W,* cần tính chuỗi từ loại có xác suất lớn nhất
  $T=t_1, t_2 ,\ldots, t_n$ hoặc,

$$\hat{T} = \arg\max_{T \in \tau} P(T \mid W)$$

(nguyên lý Bayes)

$$= \arg\max_{T \in \tau} \frac{P(T)P(W \mid T)}{P(W)}$$

$$= \arg\max_{T \in \tau} P(T)P(W \mid T)$$