



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

Gán nhãn từ loại

Lê Thanh Hương

Bộ môn Hệ thống Thông tin

Viện CNTT & TT – Trường ĐHBKHN

Email: huonglt@soict.hust.edu.vn

Định nghĩa

- Gán nhãn từ loại (Part of Speech tagging - POS tagging): mỗi từ trong câu được gán nhãn thể 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

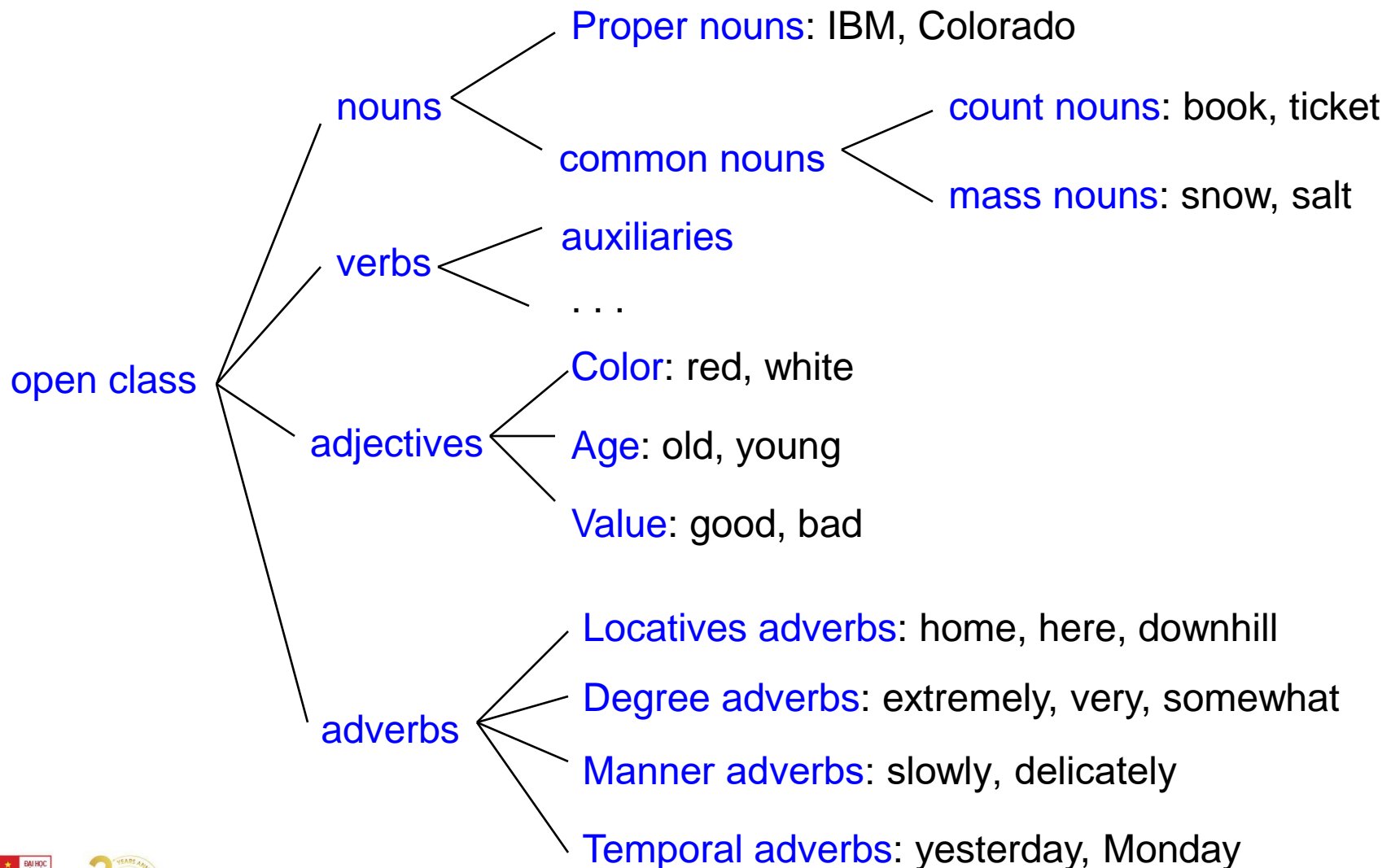
Tại sao cần gán nhãn?

- **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	Tag	Description	Example
CC	Coordin. Conjunction	<i>and, but, or</i>	SYM	Symbol	<i>+, %, &</i>
CD	Cardinal number	<i>one, two, three</i>	TO	“to”	<i>to</i>
DT	Determiner	<i>a, the</i>	UH	Interjection	<i>ah, oops</i>
EX	Existential ‘there’	<i>there</i>	VB	Verb, base form	<i>eat</i>
FW	Foreign word	<i>mea culpa</i>	VBD	Verb, past tense	<i>ate</i>
IN	Preposition/sub-conj	<i>of, in, by</i>	VBG	Verb, gerund	<i>eating</i>
JJ	Adjective	<i>yellow</i>	VCN	Verb, past participle	<i>eaten</i>
JJR	Adj., comparative	<i>bigger</i>	VBP	Verb, non-3sg pres	<i>eat</i>
JJS	Adj., superlative	<i>wildest</i>	VBZ	Verb, 3sg pres	<i>eats</i>
LS	List item marker	<i>1, 2, One</i>	WDT	Wh-determiner	<i>which, that</i>
MD	Modal	<i>can, should</i>	WP	Wh-pronoun	<i>what, who</i>
NN	Noun, sing. or mass	<i>llama</i>	WP\$	Possessive wh-	<i>whose</i>
NNS	Noun, plural	<i>llamas</i>	WRB	Wh-adverb	<i>how, where</i>
NNP	Proper noun, singular	<i>IBM</i>	\$	Dollar sign	<i>\$</i>
NNPS	Proper noun, plural	<i>Carolinas</i>	#	Pound sign	<i>#</i>
PDT	Predeterminer	<i>all, both</i>	“	Left quote	<i>(‘ or “)</i>
POS	Possessive ending	<i>’s</i>	”	Right quote	<i>(’ or ”)</i>
PP	Personal pronoun	<i>I, you, he</i>	(Left parenthesis	<i>([, (, {, <)</i>
PP\$	Possessive pronoun	<i>your, one’s</i>)	Right parenthesis	<i>(],), }, >)</i>
RB	Adverb	<i>quickly, never</i>	,	Comma	<i>,</i>
RBR	Adverb, comparative	<i>faster</i>	.	Sentence-final punc	<i>(. ! ?)</i>
RBS	Adverb, superlative	<i>fastest</i>	:	Mid-sentence punc	<i>(: ; ... – -)</i>
RP	Particle	<i>up, off</i>			

Tag	Description	Example	Tag	Description	Example
CC	Coordin. Conjunction	<i>and, but, or</i>	SYM	Symbol	<i>+, %, &</i>
CD	Cardinal number	<i>one, two, three</i>	TO	“to”	<i>to</i>
DT	Determiner	<i>a, the</i>	UH	Interjection	<i>ah, oops</i>
EX	Existential ‘there’	<i>there</i>	VB	Verb, base form	<i>eat</i>
FW	Foreign word	<i>mea culpa</i>	VBD	Verb, past tense	<i>ate</i>
IN	Preposition/sub-conj	<i>of, in, by</i>	VBG	Verb, gerund	<i>eating</i>
JJ	Adjective	<i>yellow</i>	VCN	Verb, past participle	<i>eaten</i>
JJR	Adj., comparative	<i>bigger</i>	VBP	Verb, non-3sg pres	<i>eat</i>
JJS	Adj., superlative	<i>wildest</i>	VBZ	Verb, 3sg pres	<i>eats</i>
LS	List item marker	<i>1, 2, One</i>	WDT	Wh-determiner	<i>which, that</i>
MD	Modal	<i>can, should</i>	WP	Wh-pronoun	<i>what, who</i>
NN	Noun, sing. or mass	<i>llama</i>	WP\$	Possessive wh-	<i>whose</i>
NNS	Noun, plural	<i>llamas</i>	WRB	Wh-adverb	<i>how, where</i>
NNP	Proper noun, singular	<i>IBM</i>	\$	Dollar sign	<i>\$</i>
NNPS	Proper noun, plural	<i>Carolinas</i>	#	Pound sign	<i>#</i>
PDT	Predeterminer	<i>all, both</i>	“	Left quote	<i>(‘ or “)</i>
POS	Possessive ending	<i>’s</i>	”	Right quote	<i>(’ or ”)</i>
PP	Personal pronoun	<i>I, you, he</i>	(Left parenthesis	<i>([, (, { , <)</i>
PP\$	Possessive pronoun	<i>your, one’s</i>)	Right parenthesis	<i>(],), }, >)</i>
RB	Adverb	I know that blocks the sun. He always books the violin concert tickets early. He says that book is interesting.			
RBR	Adverb, co				
RBS	Adverb, su				
RP	Particle				

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)

$$\text{Pr (Det-N)} > \text{Pr (Det-Det)}$$

- **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 ràng buộc ngữ pháp**: “không đoán, chỉ loại trừ những khả năng sai”
- **Dựa trên chuyển đổi**: “Đoán trước, sau đó có thể thay đổi”

Gán nhãn dựa trên xác suất

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

Cách thực hiện:

- Hidden Markov model (HMM):

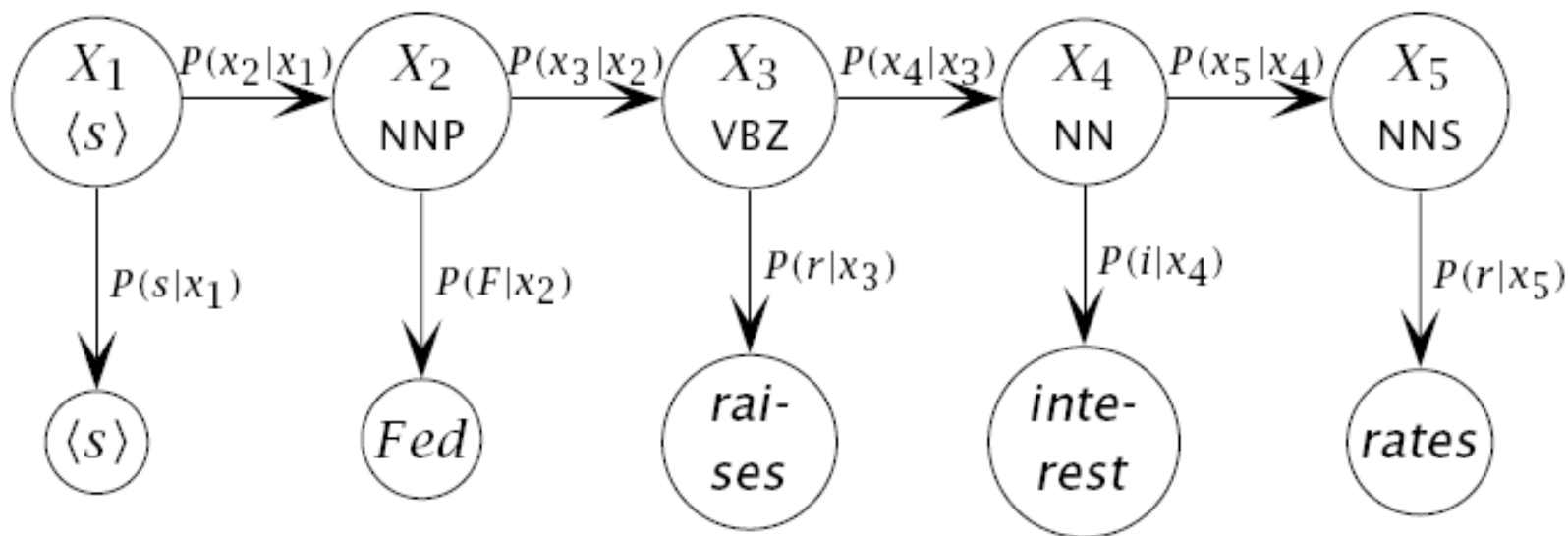
Chọn thẻ từ loại làm tối đa xác suất:

$P(\text{từ}|\text{từ loại}) \cdot P(\text{từ loại} | n \text{ từ loại phía trước})$

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

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

Ví dụ -HMMs



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

Gán nhãn HMM

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

$$t_j = \operatorname{argmax}_j P(t_j | t_{i-1}, w_i) \quad (1)$$

- **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

$$t_i = \operatorname{argmax}_j P(t_j | t_{i-1}) P(w_i | t_j) \quad (2)$$

xs chuỗi thẻ

(các thẻ đồng xuất hiện)

xs từ thường xuất hiện với thẻ t_j

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): $t_i = \operatorname{argmax}_j P(t_j | t_{i-1}) P(w_i | t_j)$

- Chọn thẻ có xác suất lớn hơn giữa 2 xác suất:
 $P(\text{VB}|\text{TO})P(\text{race}|\text{VB})$ hoặc $P(\text{NN}|\text{TO})P(\text{race}|\text{NN})$

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



Tính xác suất

Xét $P(\text{VB}|\text{TO})$ và $P(\text{NN}|\text{TO})$

- Từ tập ngữ liệu Brown

$$P(\text{NN}|\text{TO}) = .021$$

$$P(\text{VB}|\text{TO}) = .340$$

$$P(\text{race}|\text{NN}) = 0.00041$$

$$P(\text{race}|\text{VB}) = 0.00003$$

- $P(\text{VB}|\text{TO})P(\text{race}|\text{VB}) = 0.00001$
 - $P(\text{NN}|\text{TO})P(\text{race}|\text{NN}) = 0.000007$
- *race cần phải là động từ nếu đi sau “TO”*

Bài tập $t_i = \operatorname{argmax}_j P(t_j | t_{i-1}) P(w_i | t_j)$

- 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

$T=t_1, t_2, \dots, t_n$ hoặc,

$$\hat{T} = \arg \max_{T \in \tau} P(T | W) \quad (\text{nguyên lý Bayes})$$

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

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

Mở rộng sử dụng luật chuỗi

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

$$\begin{aligned} P(A,B,C) &= P(B,C|A)P(A) = P(C|A,B)P(B|A)P(A) \\ &= P(A)P(B|A)P(C|A,B) \end{aligned}$$

$$P(A,B,C,D...) = P(A)P(B|A)P(C|A,B)P(D|A,B,C..)$$

$$P(T)P(W | T) = \prod_{i=1}^n \underbrace{P(w_i | w_1 t_1 \dots w_{i-1} t_{i-1} t_i)}_{\text{pr từ}} \underbrace{P(t_i | w_1 t_1 \dots w_{i-1} t_{i-1})}_{\text{lịch sử nhấn}}$$

Giả thiết trigram

- Xác suất 1 từ chỉ phụ thuộc vào nhãn của nó

$$P(w_i | w_1 t_1 \dots t_{i-1} t_i) = P(w_i | t_i)$$

- Ta lấy lịch sử nhãn thông qua 2 nhãn gần nhất (trigram: 2 nhãn gần nhất + nhãn hiện tại)

$$P(t_i | w_1 t_1 \dots t_{i-1}) = P(t_i | t_{i-2} t_{i-1})$$

Thay vào công thức

$$P(T)P(W|T) =$$

$$P(t_1)P(t_2 | t_1) \prod_{i=3}^n P(t_i | t_{i-2}t_{i-1}) \left[\prod_{i=1}^n P(w_i | t_i) \right]$$

Đánh giá xác suất

- Sử dụng quan hệ xác suất từ tập ngữ liệu để đánh giá xác suất:

$$P(t_i \mid t_{i-1}t_{i-2}) = \frac{c(t_{i-2}t_{i-1}t_i)}{c(t_{i-2}t_{i-1})}$$

$$P(w_i \mid t_i) = \frac{c(w_i, t_i)}{c(t_i)}$$

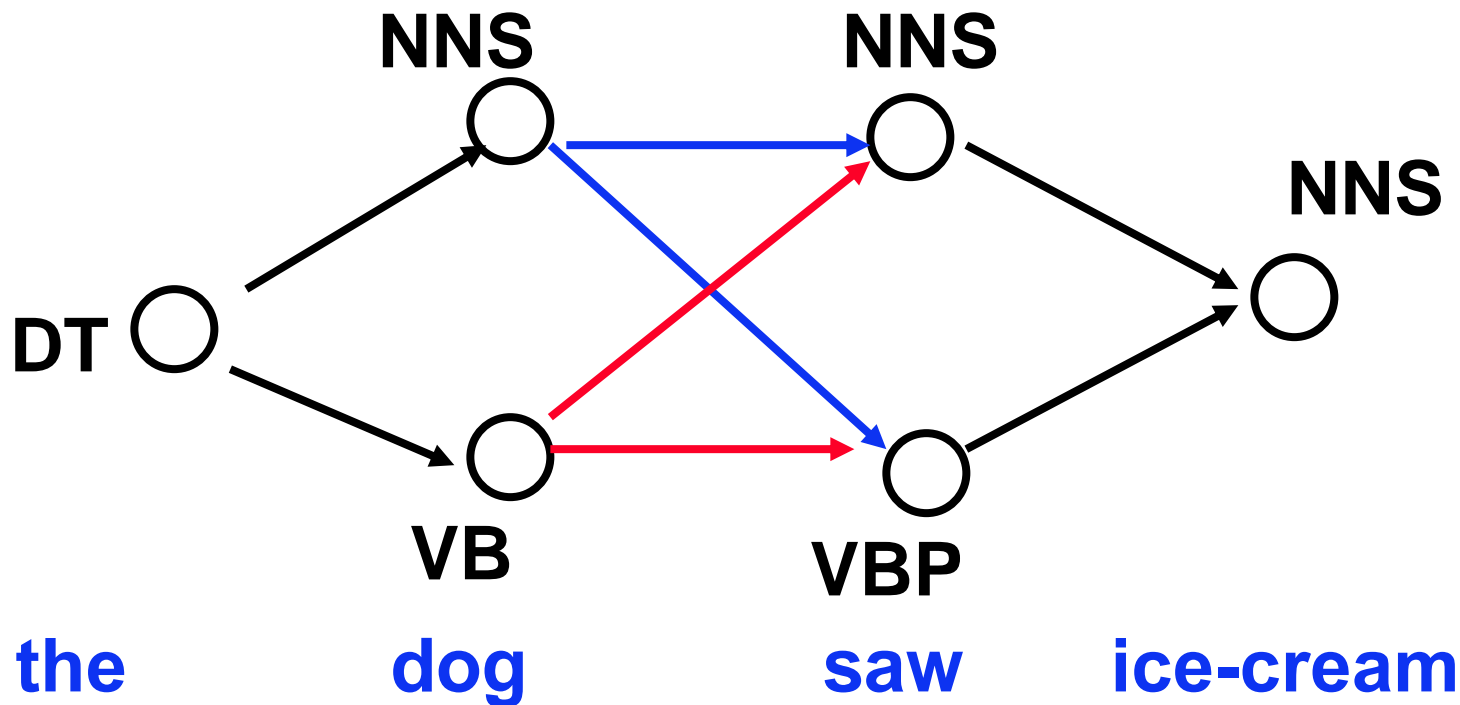
Bài toán

Cần giải quyết

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

Bây giờ ta có thể tính được tất cả các tích
 $P(T)P(W|T)$

Ví dụ



Tìm đường đi tốt nhất?

Cách tìm đường đi có điểm cao nhất

- Sử dụng tìm kiếm kiểu best-first (A^*)
 1. Tại mỗi bước, chọn k giá trị tốt nhất (\hat{T}) . Mỗi giá trị trong k giá trị này ứng với 1 khả năng kết hợp nhãn của tất cả các từ
 2. Khi gán từ tiếp theo, tính lại xác suất. Quay lại bước 1
- **Ưu:** nhanh (không cần kiểm tra tất cả các khả năng kết hợp, chỉ k cái tiềm năng nhất)
- **Nhược:** có thể không trả về kết quả tốt nhất mà chỉ chấp nhận được

Độ chính xác

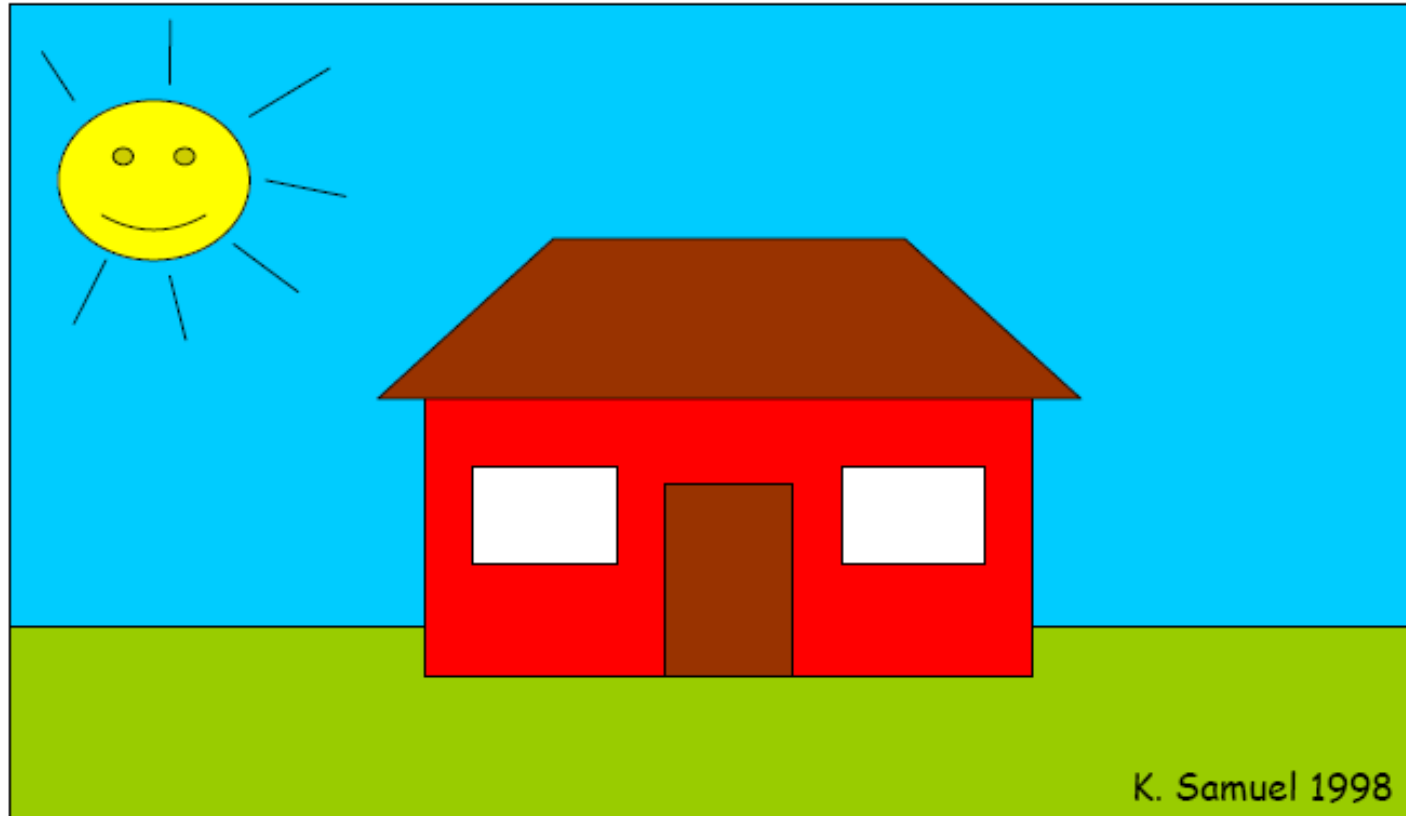
- > 96%
- Cách đơn giản nhất? 90%
 - Gán mỗi từ với từ loại thường xuyên nhất của nó
 - Gán từ chưa biết = danh từ
- Người: 97%+/- 3%; nếu có thảo luận: 100%

Cách tiếp cận thứ 2: gán nhãn dựa trên chuyển đổi

Transformation-based Learning (TBL):

- Kết hợp cách tiếp cận dựa trên luật và cách tiếp cận xác suất: sử dụng học máy để chỉnh lại thể thông qua vài lần duyệt
- Gán nhãn sử dụng tập luật tổng quát nhất, sau đó đến tập luật hẹp hơn, thay đổi một số nhãn, và tiếp tục

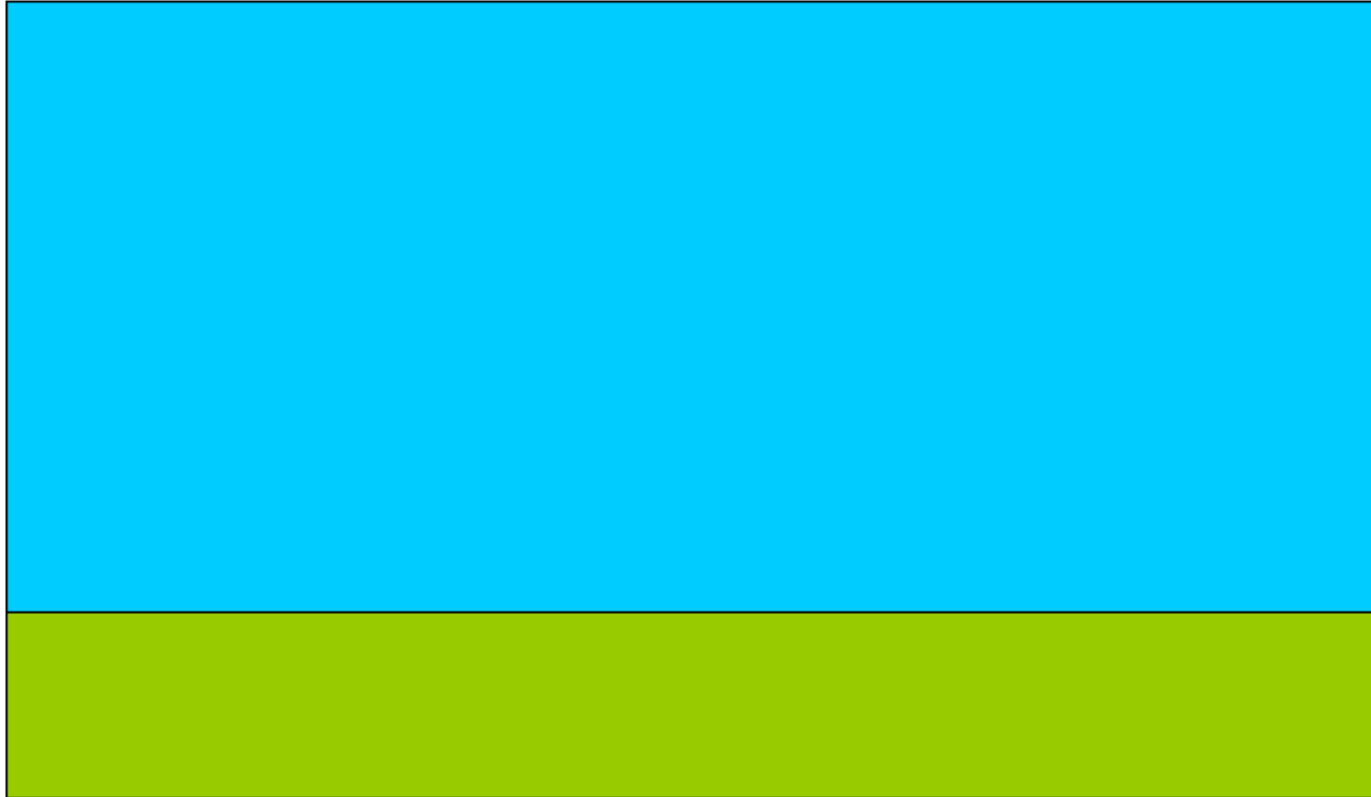
Transformation-based painting



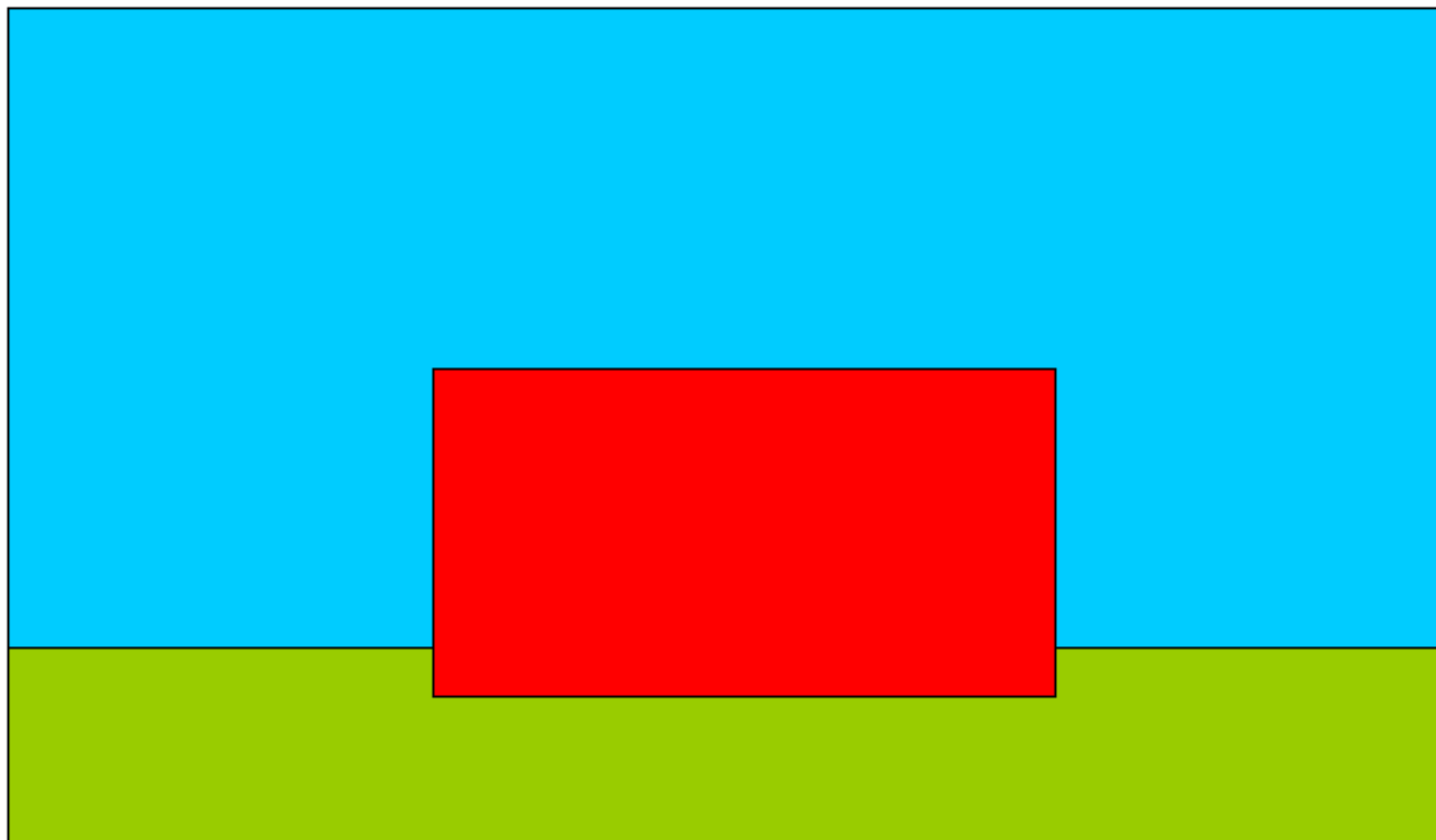
Transformation-based painting



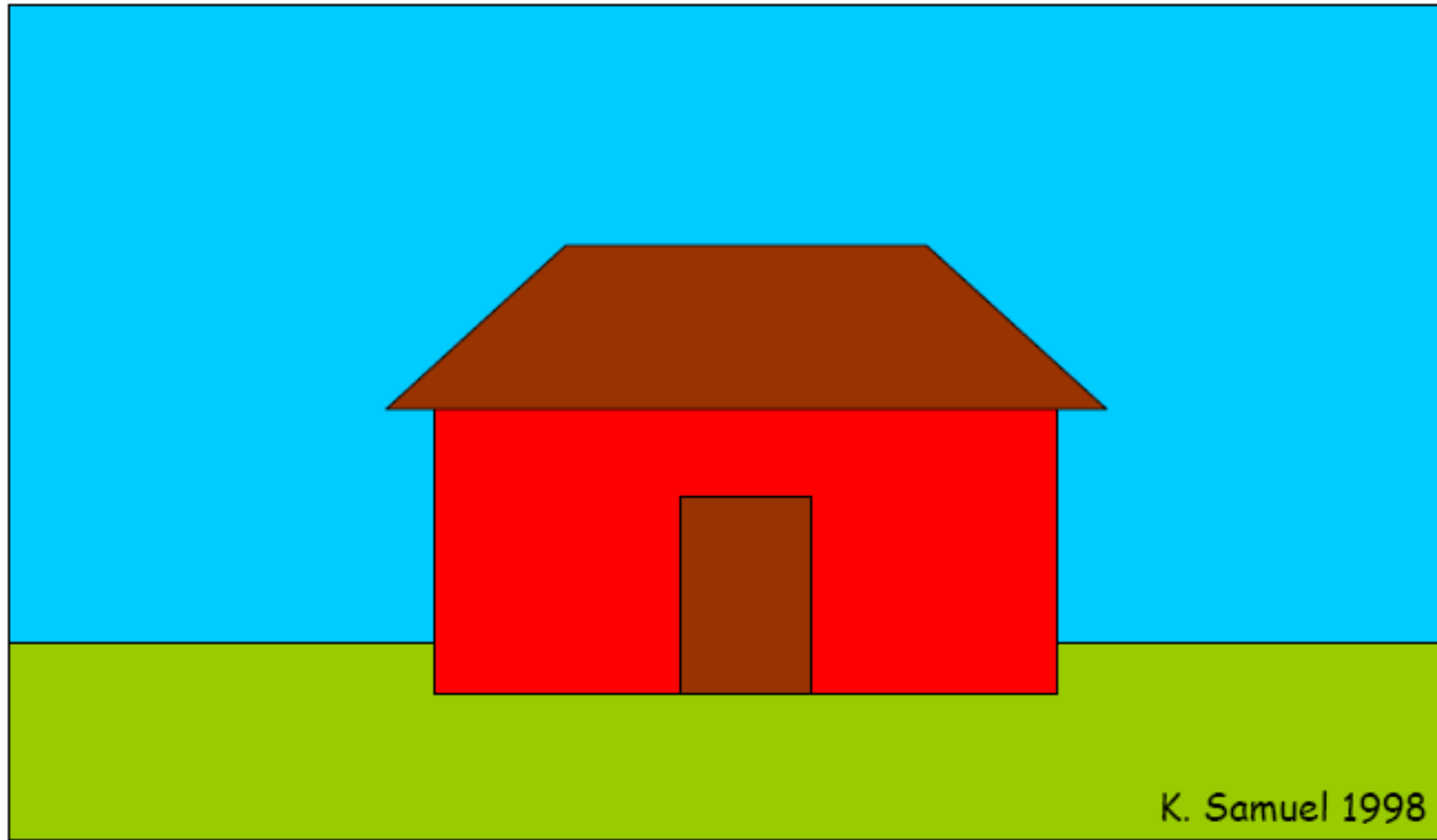
Transformation-based painting



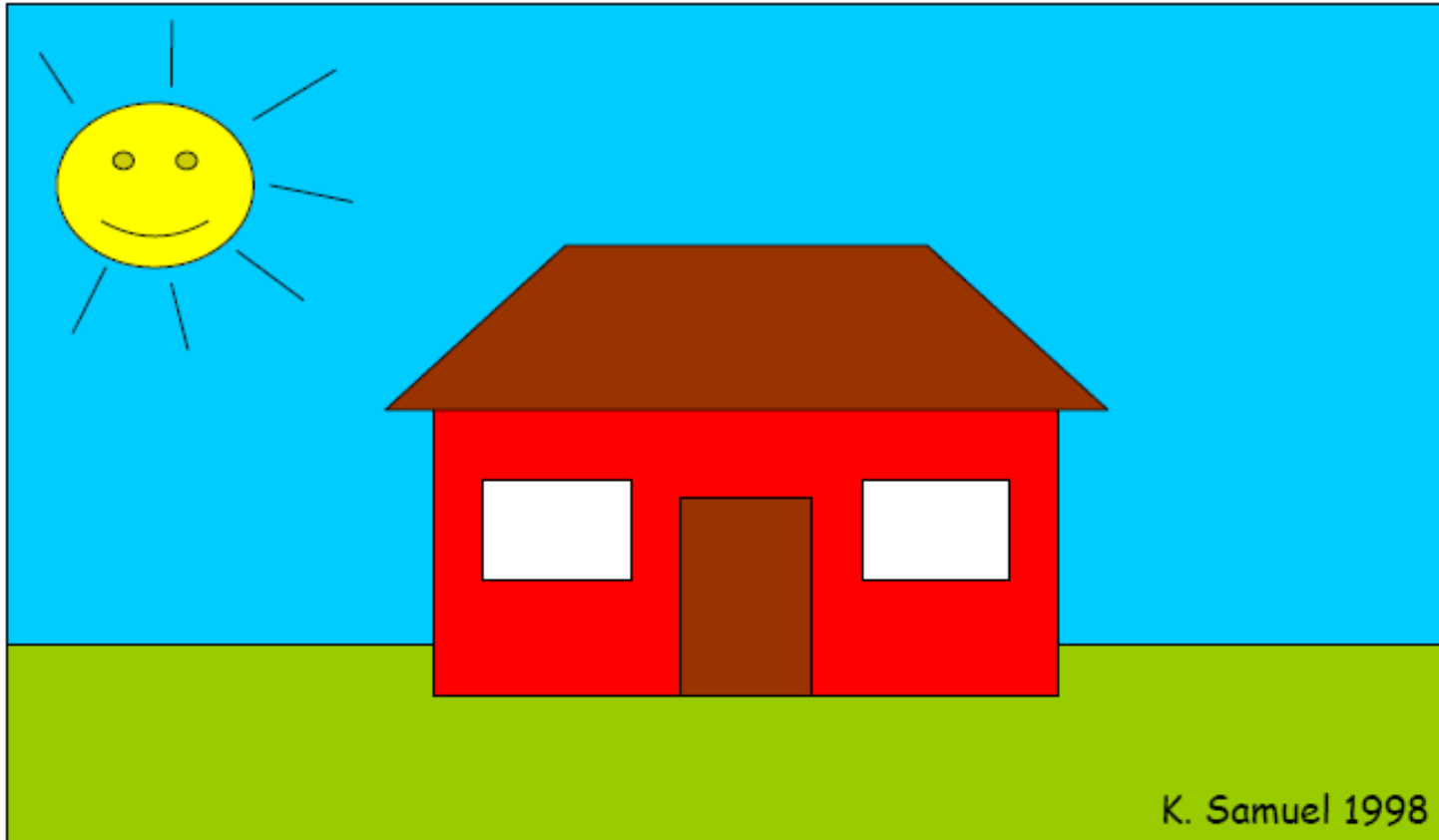
Transformation-based painting



Transformation-based painting



Transformation-based painting



Ví dụ với TBL

lexicon

data:NN
decided:VB
her:PN
she:PN N
table:NN VB
to:TO

rules

```
pos:NN>VB <- pos:TO@[-1] o
pos:VB>NN <- pos:DT@[-1] o
....
```

input

She	decided	to	table	her	data
NP	VB	TO	MB	PN	NN

Ví dụ với TBL

1. Gán mọi từ với nhãn thường xuất hiện nhất (thường độ chính xác khoảng 90%). Từ tập ngữ liệu Brown:

$$P(\text{NN}|\text{race}) = 0.98$$

$$P(\text{VB}|\text{race}) = 0.02$$

2. ...expected/VBZ to/ TO ~~race/NN~~ tomorrow/NN
...the/DT ~~race/NN~~ for/TO ~~race/VB~~ pace/NN

3. Sử dụng luật chuyển đổi:

Thay **NN** bằng **VB** khi thẻ trước đó là **TO**

pos: 'NN' > 'VB' ← pos: 'TO' @[-1] o

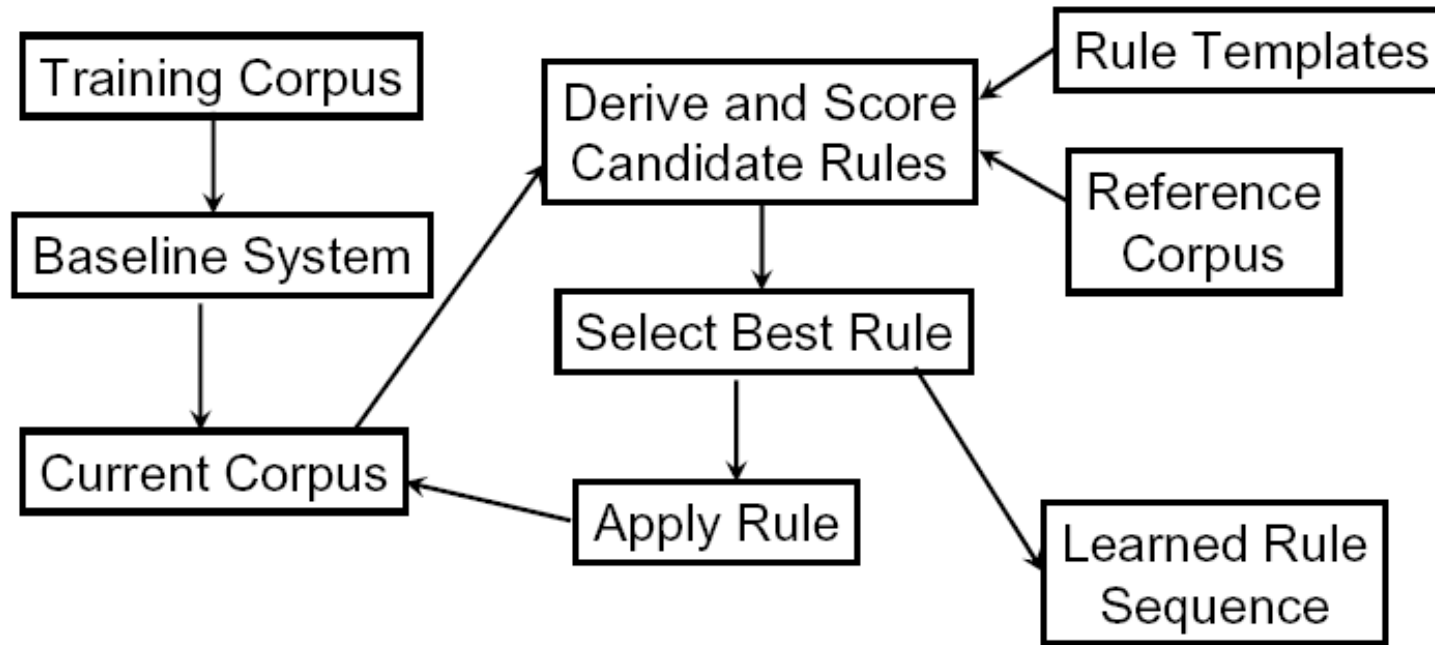
Luật gán nhãn từ loại

```
pos: 'NN' > 'VB' <- pos: 'TO' @ [-1] o
pos: 'VBP' > 'VB' <- pos: 'MD' @ [-1, -2, -3] o
pos: 'NN' > 'VB' <- pos: 'MD' @ [-1, -2] o
pos: 'VB' > 'NN' <- pos: 'DT' @ [-1, -2] o
pos: 'VBD' > 'VBN' <- pos: 'VBZ' @ [-1, -2, -3] o
pos: 'VBN' > 'VBD' <- pos: 'PRP' @ [-1] o
pos: 'POS' > 'VBZ' <- pos: 'PRP' @ [-1] o
pos: 'VB' > 'VBP' <- pos: 'NNS' @ [-1] o
pos: 'IN' > 'RB' <- wd:as@[0] & wd:as@[2] o
pos: 'IN' > 'WDT' <- pos: 'VB' @ [1, 2] o
pos: 'VB' > 'VBP' <- pos: 'PRP' @ [-1] o
pos: 'IN' > 'WDT' <- pos: 'VBZ' @ [1] o
```

Luật gán nhãn từ loại

NN VB PREVTAG TO
VB VBP PREVTAG PRP
VBD VBN PREV1OR2TAG VBD
VBN VBD PREVTAG PRP
NN VB PREV1OR2TAG MD
VB VBP PREVTAG NNS
VB NN PREV1OR2TAG DT
VBN VBD PREVTAG NNP
VBD VBN PREV1OR2OR3TAG VBZ
IN DT PREVTAG IN
VBP VB PREV1OR2OR3TAG MD
IN RB WDAND2AFT as as
VBD VBN PREV1OR2TAG VB
RB JJ NEXTTAG NN
VBP VB PREV1OR2OR3TAG TO
POS VBZ PREVTAG PRP
NN VBP PREVTAG PRP
DT PDT NEXTTAG DT

Học luật TB trong hệ thống TBL



Stop when score of best rule falls below threshold.

Các tập ngữ liệu

- Tập huấn luyện

w0 w1 w2 w3 w4 w5 w6 w7 w8 w9 w10

- Tập ngữ liệu hiện tại (CC 1)

dt vb nn dt vb kn dt vb ab dt vb

- Tập ngữ liệu tham khảo

dt nn vb dt nn kn dt jj kn dt nn

Khuôn dạng cho luật gán nhãn từ loại

- Trong TBL, chỉ các luật thỏa khuôn dạng mới được học.

- Ví dụ: các luật

tag:'VB'>'NN' \leftarrow tag:'DT'@[-1].

tag:'NN'>'VB' \leftarrow tag:'DT'@[-1].

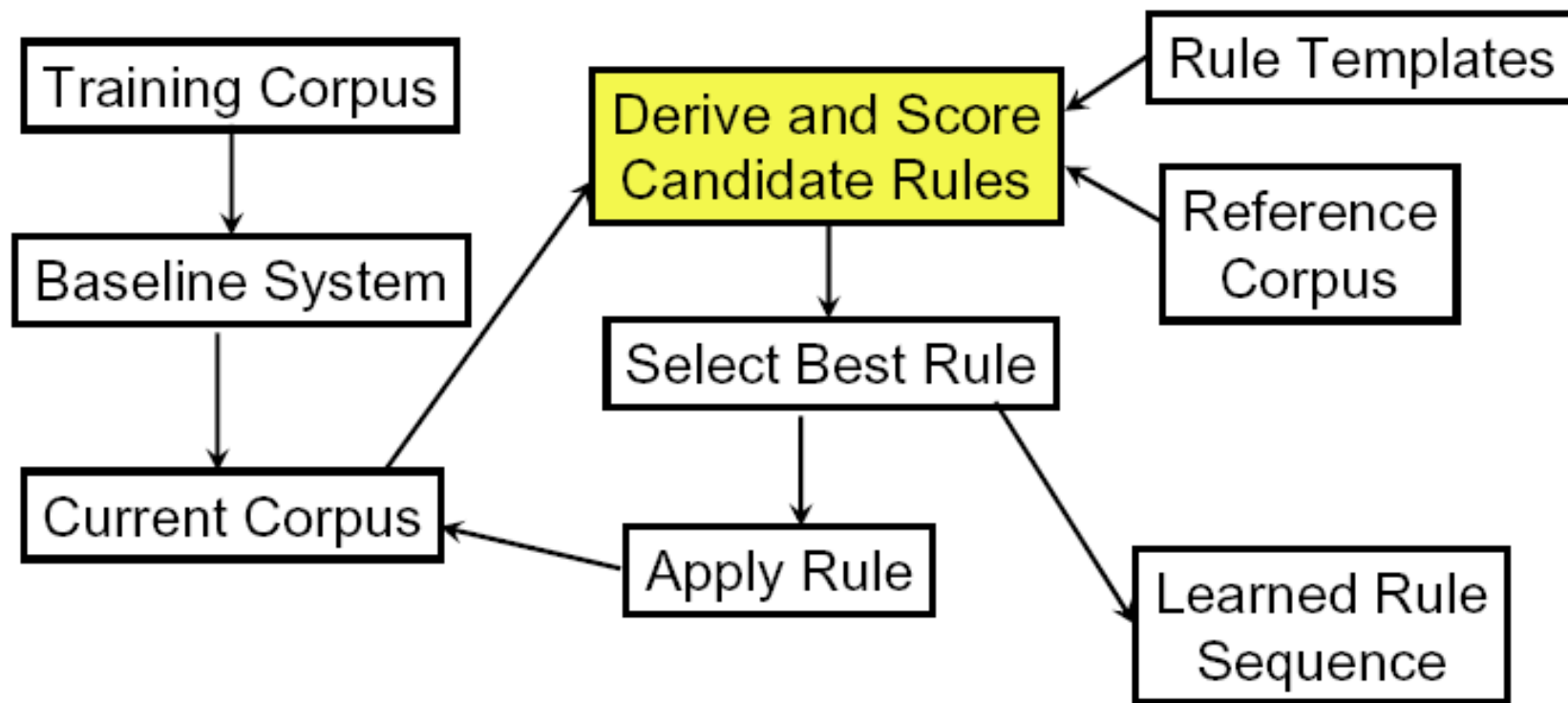
thỏa khuôn dạng

tag:A>B \leftarrow tag:C@[-1].

- Có thể tạo khuôn dạng sử dụng các biến vô danh

tag:_>_ \leftarrow tag:_@[-1].

Học luật TB trong hệ thống TBL



Điểm, độ chính xác, ngưỡng

- Điểm của 1 luật:

$$\text{score}(R) = |\text{pos}(R)| - |\text{neg}(R)|$$

- Độ chính xác:

$$\text{accuracy}(R) = \frac{|\text{pos}(R)|}{|\text{pos}(R)| + |\text{neg}(R)|}$$

- Threshold*: ngưỡng mà độ chính xác của 1 luật cần vượt qua để có thể được lựa chọn.
- Trong TBL, ngưỡng của độ chính xác thường < 0.5 .

Sinh và tính điểm cho luật ứng viên 1

- Template = tag:_>_ \leftarrow tag:_@[-1]
- R1 = tag:vb>nn \leftarrow tag:dt@[-1]

CC i	dt	vb	nn	dt	vb	kn	dt	vb	ab	dt	vb
CC i+1	dt	nn	nn	dt	nn	kn	dt	nn	ab	dt	nn

Ref. C	dt	nn	vb	dt	nn	kn	dt	jj	kn	dt	nn
--------	----	----	----	----	----	----	----	----	----	----	----

- $\text{pos}(R1) = 3$
- $\text{neg}(R1) = 1$
- $\text{score}(R1) = \text{pos}(R1) - \text{neg}(R1) = 3 - 1 = 2$

Sinh và tính điểm cho luật ứng viên 2

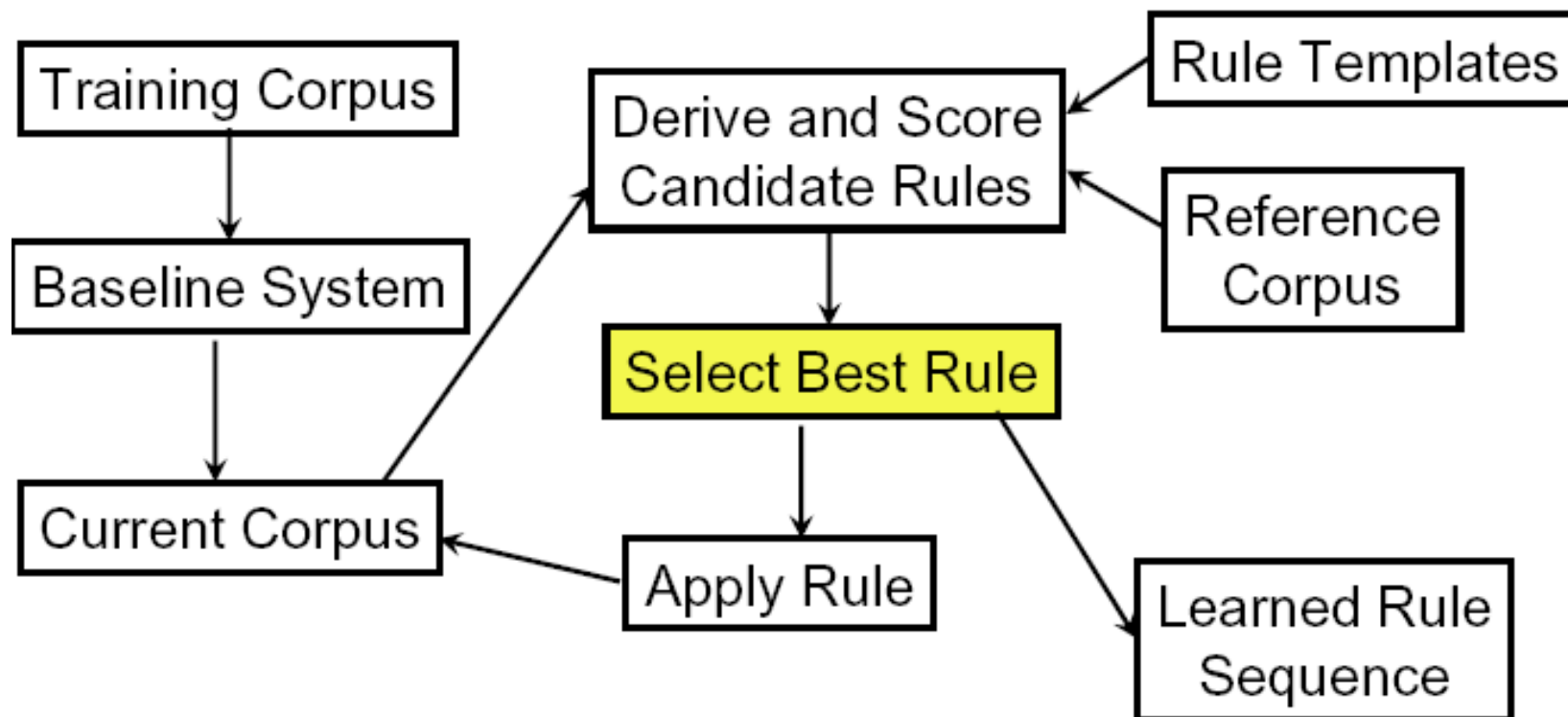
- Template = tag:_>_ \leftarrow tag:_@[-1]
- R2 = tag:nn>vb \leftarrow tag:vb@[-1]

CC i	dt	vb	nn	dt	vb	kn	dt	vb	ab	dt	vb
CC i+1	dt	vb	vb	dt	vb	kn	dt	vb	ab	dt	vb

Ref. C	dt	nn	vb	dt	nn	kn	dt	nn	kn	dt	nn
--------	----	----	----	----	----	----	----	----	----	----	----

- $\text{pos}(R2) = 1$
- $\text{neg}(R2) = 0$
- $\text{score}(R2) = \text{pos}(R2) - \text{neg}(R2) = 1 - 0 = 1$

Học luật TB trong hệ thống TBL



Stop when score of best rule falls below threshold.

Chọn luật tốt nhất

- Thứ hạng hiện tại của luật ứng viên

$R1 = \text{tag:vb} > \text{nn} \leftarrow \text{tag:dt} @ [-1] \text{ Score} = 2$

$R2 = \text{tag:nn} > \text{vb} \leftarrow \text{tag:vb} @ [-1] \text{ Score} = 1$

...

- Nếu score threshold ≤ 2 thì chọn R1
- ngược lại nếu score threshold > 2 , dừng

Tối ưu hóa việc chọn luật tốt nhất

- **Giảm dư thừa luật:** chỉ sinh các luật ứng viên phù hợp ít nhất với 1 dữ liệu trong tập luyện.
- **Đánh giá tăng cường:**
 - Lưu vết của các luật ứng viên tốt nhất
 - Bỏ qua các luật phù hợp với số lượng mẫu $<$ score của luật tốt nhất

Ưu điểm của TBL

- Luật có thể được tạo thủ công
- Luật dễ hiểu và logic
- Dễ cài đặt
- Có thể chạy rất nhanh (nhưng cài đặt thì phức tạp)

Phân tích lỗi: khó khăn đối với bộ gán nhãn từ loại

Các lỗi thông thường (> 4%)

- NN (common noun) vs .NNP (proper noun) vs. JJ (adjective): khó phân biệt, sự phân biệt này là quan trọng đặc biệt trong trích rút thông tin
- RP (particle) vs. RB (adverb) vs. IN (preposition): tất cả các loại này có thể xuất hiện tuần tự sau động từ
- VBD vs. VBN vs. JJ: phân biệt thời quá khứ, phân từ 2, tính từ (*raced* vs. *was raced* vs. *the out raced horse*)

Cách tốt nhất phát hiện các từ chưa biết

- Dựa trên 3 dạng đuôi biến tố (*-ed, -s, -ing*); 32 đuôi phái sinh (*-ion, etc.*); chữ hoa; gạch nối
- Tổng quát hơn:
 - Phân tích hình thái từ
 - Các cách tiếp cận học máy

Gán nhãn từ loại tiếng Việt

Câu tiếng Việt đã tách từ	Qua những lần từ Sài_Gòn về Quảng_Ngãi kiểm_tra công_việc , Sophie và Jane thường trò_chuyện với Mai , cảm_nhận ngọn_lửa_sống và niềm_tin mãnh_liệt từ người phụ_nữ VN này .
Câu tiếng Việt đã được gán nhãn từ loại	Qua những lần từ Sài_Gòn về Quảng_Ngãi kiểm_tra công_việc , Sophie và Jane thường trò_chuyện với Mai , cảm_nhận ngọn_lửa_sống và niềm_tin mãnh_liệt từ người phụ_nữ VN này .
Chú thích từ loại	<div> DANH TỪ ■ DANH TỪ ■ DANH TỪ ■ ĐỘNG TỪ ■ ĐỘNG TỪ ■ ĐỘNG TỪ ■ TÍNH TỪ ■ TÍNH TỪ ■ TÍNH TỪ ■ ĐẠI TỪ ■ ĐẠI TỪ ■ ĐẠI TỪ ■ ĐỊNH TỪ ■ ĐỊNH TỪ ■ ĐỊNH TỪ ■ </div> <div> SỐ TỪ ■ SỐ TỪ ■ SỐ TỪ ■ PHỤ TỪ ■ PHỤ TỪ ■ PHỤ TỪ ■ GIỚI TỪ ■ GIỚI TỪ ■ GIỚI TỪ ■ CẢM TỪ ■ CẢM TỪ ■ CẢM TỪ ■ LIÊN TỪ ■ LIÊN TỪ ■ LIÊN TỪ ■ </div> <div> THÁN TỪ ■ THÁN TỪ ■ THÁN TỪ ■ TRỢ TỪ ■ TRỢ TỪ ■ TRỢ TỪ ■ TỪ ĐƠN LẺ ■ TỪ ĐƠN LẺ ■ TỪ ĐƠN LẺ ■ TỪ VIẾT TẮT ■ TỪ VIẾT TẮT ■ TỪ VIẾT TẮT ■ KHÔNG XÁC ĐỊNH ■ KHÔNG XÁC ĐỊNH ■ KHÔNG XÁC ĐỊNH ■ </div>

Các bước thực hiện

- Gán nhãn cơ sở
 - Gán nhãn tiên nghiệm (gán mỗi từ với tất cả các nhãn từ loại mà nó có thể có).
 - Với một từ mới, dùng một nhãn ngầm định hoặc gán cho nó tập tất cả các nhãn. Với ngôn ngữ biến đổi hình thái → dựa vào hình thái từ
- Quyết định kết quả gán nhãn (loại bỏ nhập nhằng)
 - dựa vào quy tắc ngữ pháp
 - dựa vào xác suất
 - sử dụng mạng nơ-ron
 - các hệ thống lai sử dụng kết hợp tính toán xác suất và ràng buộc ngữ pháp

Dữ liệu phục vụ gán nhãn

- Ngữ liệu:
 - Từ điển từ vựng
 - Kho văn bản đã gán nhãn, có thể kèm theo các quy tắc ngữ pháp xây dựng bằng tay
 - Kho văn bản chưa gán nhãn, có kèm theo các thông tin ngôn ngữ như là tập từ loại
 - Kho văn bản chưa gán nhãn, với tập từ loại được xây dựng tự động nhờ các tính toán thống kê

A Penn Treebank tree

```
( (S (NP-SBJ The move)
    (VP followed
      (NP (NP a round)
        (PP of
          (NP (NP similar increases)
            (PP by
              (NP other lenders))
            (PP against
              (NP Arizona real estate loans))))))
    ,
    (S|ADV (NP-SBJ *)
      (VP reflecting
        (NP (NP a continuing decline)
          (PP-LOC in
            (NP that market))))))
  .))
```


Khó khăn trong gán nhãn từ loại tiếng Việt

- đặc trưng riêng về ngôn ngữ
 - thiếu các kho dữ liệu chuẩn như Brown hay Penn Treebank
- khó khăn trong đánh giá kết quả

Cách tiếp cận 1

[Đình Điền] Dien Dinh and Kiem Hoang, POS-tagger for English-Vietnamese bilingual corpus. HLTNAACL Workshop on Building and using parallel texts: data driven machine translation and beyond, 2003.

- chuyển đổi và ánh xạ từ thông tin từ loại từ tiếng Anh do
 - gán nhãn từ loại trong tiếng Anh đã đạt độ chính xác cao (>97%)
 - những thành công gần đây của các phương pháp giống hàng từ (word alignment methods) giữa các cặp ngôn ngữ.

[Đinh Điền]

- Xây dựng một tập ngữ liệu song ngữ Anh – Việt ~ 5 triệu từ (cả Anh lẫn Việt).
- gán nhãn từ loại cho tiếng Anh dựa trên Transformation-based Learning – TBL [Brill 1995]
- giống hàng giữa hai ngôn ngữ (độ chính xác khoảng 87%) để chuyển nhãn từ loại sang tiếng Việt.
- kết quả được hiệu chỉnh bằng tay để làm dữ liệu huấn luyện cho bộ gán nhãn từ loại tiếng Việt.

[Định Điền]

- Ưu điểm:
 - tránh được việc gán nhãn từ loại bằng tay nhờ tận dụng thông tin từ loại ở một ngôn ngữ khác.
- Nhược:
 - Tiếng Anh và tiếng Việt khác nhau: về cấu tạo từ, trật tự và chức năng ngữ pháp của từ trong câu → khó khăn trong giống hàng
 - Lỗi tích lũy qua hai giai đoạn: (a) gán nhãn từ loại cho tiếng Anh và (b) giống hàng giữa hai ngôn ngữ
 - Tập nhãn được chuyển đổi trực tiếp từ tiếng Anh sang tiếng Việt không diễn hình cho từ loại tiếng Việt

Cách tiếp cận 2

- [Nguyen Huyen, Vu Luong] Thi Minh Huyen Nguyen, Laurent Romary, and Xuan Luong Vu, A Case Study in POS Tagging of Vietnamese Texts. The 10th annual conference TALN 2003.
- dựa trên nền tảng và tính chất ngôn ngữ của tiếng Việt.
- xây dựng tập từ loại (tagset) cho tiếng Việt dựa trên chuẩn mô tả khá tổng quát của các ngôn ngữ Tây Âu, nhằm mô đun hóa tập nhãn ở hai mức:
 - mức cơ bản/cốt lõi (kernel layer): đặc tả chung nhất cho các ngôn ngữ
 - mức tính chất riêng (private layer): mở rộng và chi tiết hóa cho một ngôn ngữ cụ thể dựa trên tính chất của ngôn ngữ đó

[Nguyen Huyen, Vu Luong]

- mức cơ bản: danh từ (noun – N), động từ (verb – V), tính từ (adjective – A), đại từ (pronoun – P), mạo từ (determine – D), trạng từ (adverb – R), tiền-hậu giới từ (adposition – S), liên từ (conjunction – C), số từ (numeral – M), tình thái từ (interjection – I), và từ ngoại Việt (residual – X, như foreign words, ...).
- mức tính chất riêng: được triển khai tùy theo các dạng từ loại trên như danh từ đếm được/không đếm được đối với danh từ, giống đực/cái đối với đại từ, .v.v.

Cách tiếp cận 3

- [Phuong] Nguyễn Thị Minh Huyền, Vũ Xuân Lương, Lê Hồng Phương . Sử dụng bộ gán nhãn từ loại xác suất QTAG cho văn bản tiếng Việt. Kỷ yếu Hội thảo ICT.rda'03
- làm việc trên một cửa sổ chứa 3 từ, sau khi đã bổ sung thêm 2 từ giả ở đầu và cuối văn bản.
- Nhãn được gán cho mỗi từ đã lọt ra ngoài cửa sổ là nhãn kết quả cuối cùng.

Thủ tục gán nhãn từ loại [Phương]

1. Đọc từ (token) tiếp theo
 2. Tìm từ đó trong từ điển
 3. Nếu không tìm thấy, gán cho từ đó tất cả các nhãn có thể
 4. Với mỗi nhãn có thể
 - a. tính $P_w = P(\text{tag}|\text{token})$
 - b. tính $P_c = P(\text{tag}|t_1, t_2)$, t_1, t_2 , là nhãn tương ứng của hai từ đứng trước từ token.
 - c. tính $P_{w,c} = P_w * P_c$, kết hợp hai xác suất trên.
 5. Lặp lại phép tính cho hai nhãn khác trong cửa sổ
- Sau mỗi lần tính lại (3 lần cho mỗi từ), các xác suất kết quả được kết hợp để cho ra xác suất toàn thể của nhãn được gán cho từ.

[Phương]

- Chia kho văn bản đã gán nhãn làm 2 tập: tập huấn luyện và tập thử nghiệm
- Tự động gán nhãn cho các phần văn bản
- So sánh kết quả thu được với dữ liệu mẫu.
- Thời gian huấn luyện với 32000 từ: ~ 30s

[Phương]

- Câu đã gán nhãn:

<w pos="Nc"> **hồi**</w> <w pos="Vto"> **lên** </w> < w pos="Nn">
sáu </w> <w pos=","> , </w> <w pos="Vs"> **có** </w> <w
pos="Nu"> **lần** </w> <w pos="Pp"> **tôi** </w> <w pos="Jt"> **đã** </w>
<w pos="Vt"> **nhìn** </w> <w pos="Vt"> **thấy** </w> <w pos="Nn">
một </w> <w pos="Nt"> **bức** </w> <w pos="Nc"> **tranh** </w> <w
pos="Jd"> **tuyệt** </w> <w pos="Aa"> **đẹp** </w>

Nc - danh từ đơn thể, Vto - ngoại động từ chỉ hướng, Nn - danh từ số lượng, Vs - động từ tồn tại, Nu - danh từ đơn vị, Pp - đại từ nhân xưng, Jt - phụ từ thời gian, Vt - ngoại động từ, Nt - danh từ loại thể, Jd - phụ từ chỉ mức độ, Aa - tính từ hàm chất.

- **Precision** = số từ gán nhãn đúng/ tổng số từ đã gán nhãn
 - **Recall** = số từ gán nhãn đúng/ tổng số từ đúng
- # [Phương]

- Câu từ tập ngữ liệu mẫu

<w pos="Nc"> **hỏi**</w> <w pos="Vto"> **lên** </w> < w pos="Nn">
sáu </w> <w pos=","> , </w> <w pos="Vs"> **có** </w> <w
pos="Nu"> **lần** </w> <w pos="Pp"> **tôi** </w> <w pos="Jt"> **đã** </w>
<w pos="Vt"> **nhìn** </w> <w pos="Vt"> **thấy** </w> <w pos="Nn">
một </w> <w pos="Nt"> **bức** </w> <w pos="Nc"> **tranh** </w> <w
pos="Jd"> **tuyệt** </w> <w pos="Aa"> **đẹp** </w>

Câu do chương trình gán nhãn

<w pos="Nc"> **hỏi**</w> <w pos="Adv"> **lên** </w> < w pos="Nn">
sáu </w> <w pos=","> , </w> <w pos="Vs"> **có** </w> <w
pos="Nu"> **lần** </w> <w pos="Pp"> **tôi** </w> <w pos="JJ"> **đã**
</w> <w pos="Vt"> **nhìn** </w> <w pos="Vt"> **thấy** </w> <w
pos="Nn"> **một** </w> <w pos="Nt"> **bức** </w> <w pos="Nc">
tranh </w> <w pos="Jd"> **tuyệt** </w> <w pos="Aa"> **đẹp** </w>

Mẫu: (30)

(E Ở)(N số)(M 10)(N phố)(Np Hàng Mạnh)(Np Hà Nội)(, ,)
(N vợ chồng) (Np Dương Tuấn) (- -) (Np Đặng Hải Lý)(, ,)
(M 26) (N tuổi)(, ,)(V mở)(N lớp) (V dạy)(V viết)(N chữ) (A
đẹp)(. .)

(N Lớp học)(E của)(P họ)(X ngày càng)(V thu hút)
(L nhiều)(N học viên)(. .)

Chương trình gán: (30)

(R Ở)(N số)(M 10)(N phố)(Np Hàng Mạnh)(Np Hà Nội)(, ,)
(N vợ chồng) (Np Dương Tuấn) (- -) (Np Đặng Hải Lý)(, ,)
(M 26) (N tuổi)(, ,)(V mở)(N lớp) (V dạy)(V viết)(N chữ) (A
đẹp)(. .)

(N Lớp học)(C của)(P họ)(R ngày càng)(A thu hút)
(A nhiều)(N học viên)(. .)

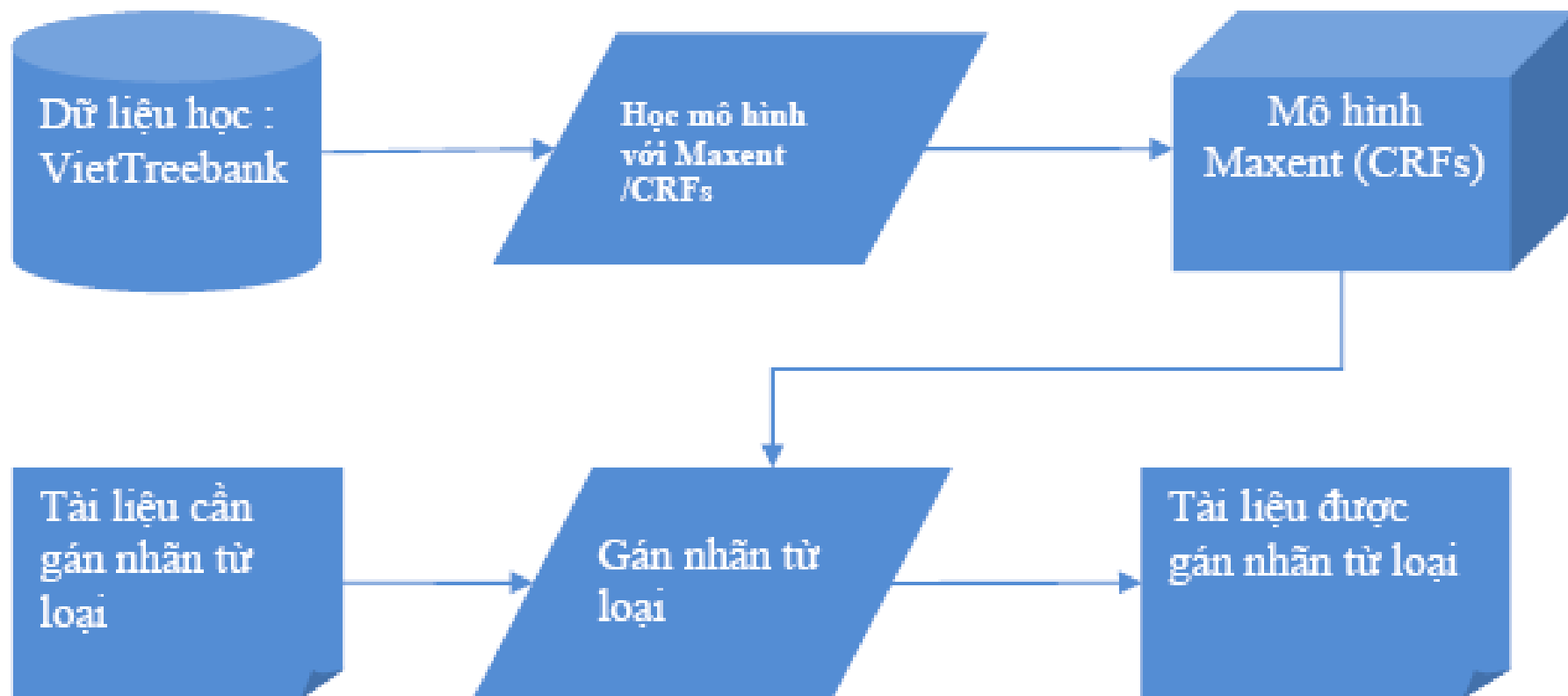
[Phương]

- Kết quả:
 - ~94% (9 nhãn từ vựng và 10 nhãn cho các loại kí hiệu)
 - ~85% (48 nhãn từ vựng và 10 nhãn cho các loại kí hiệu)
- Nếu không dùng đến từ điển từ vựng (chỉ sử dụng kho văn bản đã gán nhãn mẫu) thì các kết quả chỉ đạt được tương ứng là ~80% và ~60%.

Cách tiếp cận 4

- Phan Xuân Hiếu (2009). Công cụ gán nhãn từ loại tiếng Việt dựa trên Conditional Random Fields và Maximum Entropy JvnTagger.
- Dựa trên phương pháp Maximum Entropy (MaxEnt) và Conditional Random Fields (CRFs) - ứng dụng rất nhiều cho các bài toán gán nhãn cho các thành phần trong dữ liệu chuỗi.
- Dữ liệu huấn luyện: là tập ngữ liệu Viet Treebank bao gồm hơn 10.000 câu tiếng Việt được gán nhãn từ loại bởi các chuyên gia ngôn ngữ.

[Hiếu]



Học mô hình gán nhãn từ loại

Trích chọn đặc trưng

- ... thường trò_chuyện với Mai ...
- Cần xác định từ loại cho từ “trò_chuyện”, các đặc trưng:
 - Chính bản thân từ “trò_chuyện” thường xuất hiện với từ loại nào trong tập dữ liệu Viet Treebank?
 - Từ “trò_chuyện” thường có nhãn từ loại là gì trong từ điển? Là động từ chẳng?
 - Từ “thường” đi ngay trước từ “trò_chuyện” thường có gợi ý gì?
 - Từ “với” đi sau từ “trò_chuyện” có gợi ý gì? Có phải nó gợi ý là ngay trước nó là một động từ hay không?
 - Kết hợp của hai từ “với Mai” gợi ý điều gì, chắc từ trước đó (“trò_chuyện”) nên là một động từ?

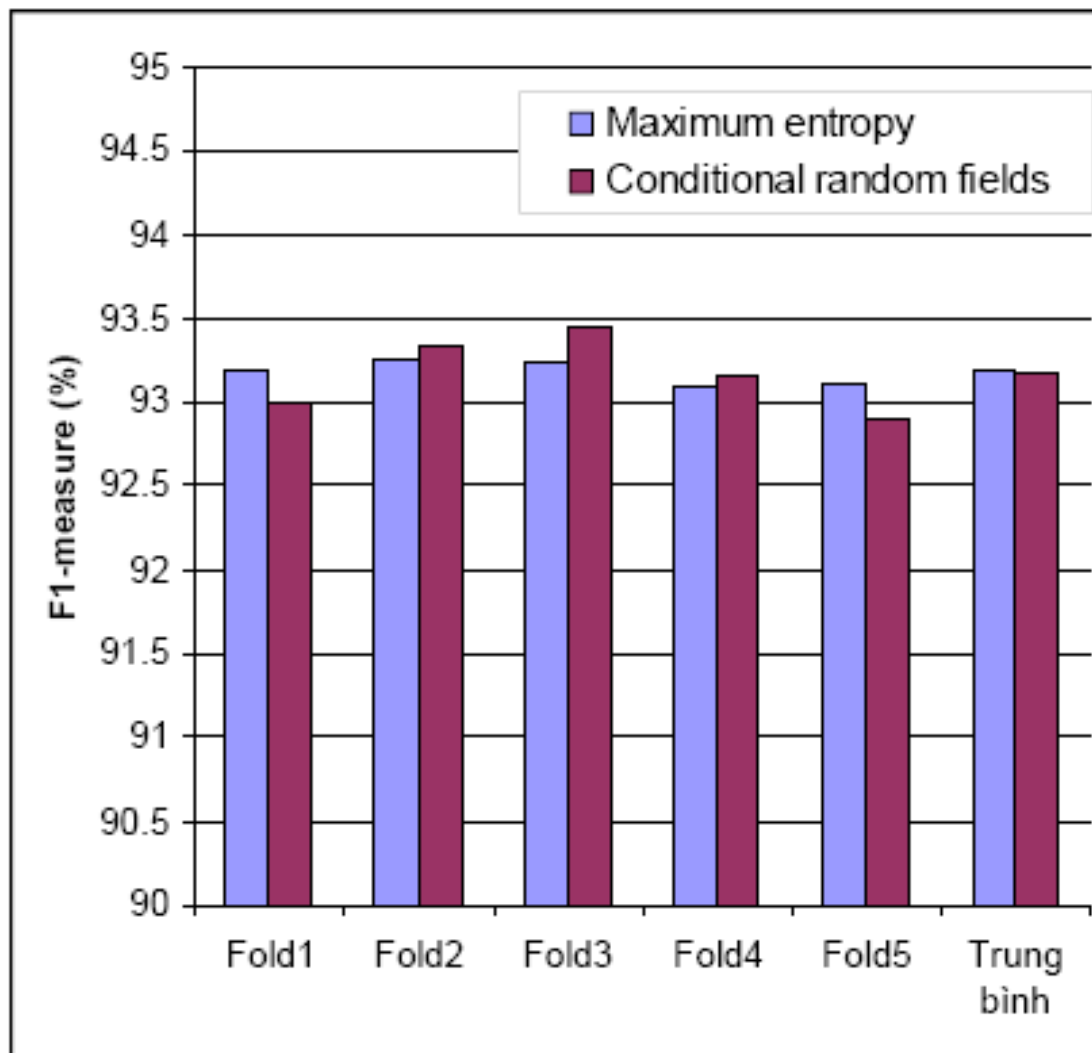
Ngữ cảnh cho trích xuất đặc trưng

Loại	Ngữ cảnh	Giải thích
<i>Mẫu ngữ cảnh cho cả Maxent và CRFs</i>		
Mẫu ngữ cảnh cơ bản (loại 1)	$w_{i-2}; w_{i-1}; w_i; w_{i+1}; w_{i+2}$	w_i cho biết từ tại vị trí thứ i trong chuỗi đầu vào (nằm trong cửa sổ trượt với kích cỡ 5)
	$w_{j-1}; w_j; w_{j+1}$	$w_{j-1}; w_j; w_{j+1}$ kết hợp từ thứ $j-1$ và từ thứ $j+1$ trong chuỗi đầu vào
	is_all_capitalized(i) ($i=0;1$); is_initial_capitalized(i) ($i=0;1$); is_number(i) ($i=-1;0;1$); contain_numbers(i) (i), contain_hyphen, contain_comma, is_marks	Kiểm tra một số thuộc tính của từ thứ i trong cửa sổ hiện tại như: từ có phải là toàn chữ viết hoa hay có kí tự đầu viết hoa hay không, có chứa số, v.v...
Mẫu ngữ cảnh từ điển (loại 2)	dict(i) ($i=0,1$)	Các từ loại có thể gán cho từ thứ i trong cửa sổ hiện tại (V, N, A, ...)

Ngữ cảnh cho trích xuất đặc trưng


Loại	Ngữ cảnh	Giải thích
<i>Mẫu ngữ cảnh cho cả Maxent và CRFs</i>		
Mẫu ngữ cảnh từ điển (loại 2)	dict(i) (i=0,1)	Các từ loại có thể gán cho từ thứ i trong cửa sổ hiện tại (V, N, A, ...)
Mẫu ngữ cảnh đặc trưng tiếng Việt (loại 3)	is_full_repretative(0), is_partial_repretative(0)	Kiểm tra xem một từ có phải từ láy toàn bộ hay một phần không
Mẫu ngữ cảnh dựa vào suffix (loại 4)	prf(0), sff(0)	Âm tiết đầu tiên (ví dụ “sự” trong “sự hướng dẫn”), cuối cùng trong từ hiện tại (“hóa” trong “công nghiệp hóa”)
<i>Mẫu cho đặc trưng cạnh của CRFs</i>		
$t_{-1} t_0$	Nhãn của từ trước đó và nhãn của từ hiện tại. Đặc trưng này được trích chọn trực tiếp từ dữ liệu bởi FlexCrfs	

Kết quả gán nhãn sử dụng MaxEnt và CRFs



Tập từ loại tiếng Việt

idPOS	symbolPOS	vnPOS	enPOS
1	N	danh từ	noun
2	V	động từ	verb
3	A	tính từ	adjective
4	M	số từ	numeral
5	P	đại từ	pronoun
6	R	phụ từ	adverb
7	O	giới từ	preposition
8	C	liên từ	conjunction
9	I	trợ từ	auxiliary word
10	E	cảm từ	emotivity word
11	Xy*	từ tắt	abbreviation
12	S	yếu tố từ (bắt, vô...)	component stem
13	U	không xác định	undetermined


 Từ tắt mang nhãn kép: X = từ loại của từ tắt ;
 y = kí hiệu từ tắt. Ví dụ: GDP-Ny; HIV – Ny.

Tập tiêu từ loại tiếng Việt

idPOS	idSub POS	symbol POS	vnPOS	enPOS
1	1	Np	danh từ riêng	proper noun
1	2	Nc	danh từ đơn thể	countable noun
1	3	Ng	danh từ tổng thể	collective Noun
1	4	Na	danh từ trừu tượng	abstract noun
1	5	Ns	danh từ chỉ loại	classifier noun
1	6	Nu	danh từ đơn vị	unit noun
1	7	Nq	danh từ chỉ lượng	quantity noun
2	8	Vi	động từ nội động	intransitive verb
2	9	Vt	động từ ngoại động	transitive verb
2	10	Vs	động từ trạng thái	state verb
2	11	Vm	động từ tình thái	modal verb
2	12	Vr	động từ quan hệ	relative verb
3	13	Ap	tính từ tính chất	property adjective
3	14	Ar	tính từ quan hệ	relative adjective
3	15	Ao	tính từ tượng thanh	onomatopoetic adjective
3	16	Ai	tính từ tượng hình	pictographic adjective

Tập tiểu từ loại tiếng Việt

idPOS	idSub POS	symbol POS	vnPOS	enPOS
4	17	Mc	số từ số lượng	cardinal numeral
4	18	Mo	số từ thứ tự	ordinal numeral
5	19	Pp	đại từ xưng hô	personal pronoun
5	20	Pd	đại từ chỉ định	demonstrative pronoun
5	21	Pq	đại từ số lượng	quality pronoun
5	22	Pi	đại từ nghi vấn	interrogative pronoun
6	23	R	phụ từ	adverb
7	24	O	giới từ	preposition
8	25	C	liên từ	conjunction
9	26	I	trợ từ	auxiliary word
10	27	E	cảm từ	emotivity word
11	28	Xy	từ tắt	abbreviation
12	29	S	yếu tố từ (bắt, vô...)	component stem
13	30	U	không xác định	undetermined