



XỬ LÝ NGÔN NGỮ TỰ NHIÊN

CHƯƠNG 5: Sequential labeling

ThS. Lưu Thanh Sơn



NỘI DUNG

- 1. Định nghĩa bài toán.**
2. Part of speech tagging
3. Named entity recognition
4. Markov chains
5. HMM Tagger
6. CRF
7. Độ đo đánh giá



1. Định nghĩa bài toán



Định nghĩa bài toán

- Input: Một chuỗi các từ X có độ dài là n. Ký hiệu:

$$X = \langle w_1 \ w_2 \ w_3 \dots \ w_n \rangle$$

- Output: Nhãn ứng với từ phần tử trong chuỗi X. Ký hiệu:

$$y = \langle y_1 \ y_2 \ y_3 \dots \ y_n \rangle$$

Việc dự đoán và gán thẻ cho từng từ trong chuỗi đầu vào được gọi là “tagging”

Các bài toán liên quan đến Gán nhãn chuỗi



- Bài toán 1: Part of speech tagging.
- Bài toán 2: Named entity recognition.



NỘI DUNG

1. Định nghĩa bài toán.
- 2. Part of speech tagging**
3. Named entity recognition
4. Markov chains
5. HMM Tagger
6. CRF
7. Độ đo đánh giá



2. Part of Speech Tagging



Part of speech

- Định nghĩa **Part-of speech** (tiếng Việt: *tùy loại*): Tùy loại là những lớp từ có cùng bản chất ngữ pháp, được phân chia theo ý nghĩa khái quát, theo khả năng kết hợp với các từ ngữ khác trong ngữ lưu và thực hiện những chức năng ngữ pháp nhất định ở trong câu (**Đinh Văn Đức. Ngữ pháp tiếng Việt – Từ loại**).
- Các tiêu chí phân định tùy loại:
 - Ý nghĩa khái quát của từ: sự vật, hành động, tính chất...
 - Khả năng kết hợp với các từ ngữ khác.
 - Chức năng ngữ pháp (chức vụ ngữ pháp, chức năng thành phần câu)



Từ loại trong tiếng Anh

- Trong tiếng Anh, từ loại thường rơi vào 2 dạng sau:
 - **Open word classes** gồm có: **danh từ, động từ, tính từ, phó từ**. Số lượng mỗi từ loại thuộc nhóm này có thể từ một vài nghìn đến cả trăm nghìn từ. Nhóm này bao gồm các **content words**, là những từ mang nghĩa nội dung hay nghĩa từ điển (lexical meaning) như: home (nhà ở, quê hương), bridge (cây cầu), slowly (chậm chạp).
 - **Closed word classes** gồm **mạo từ, định từ, đại từ, giới từ, liên từ và thán từ**. Số lượng mỗi từ loại thuộc nhóm này chỉ từ vài từ đến vài mươi từ và rất ít khi nhận thêm từ mới. Nhóm này bao gồm các **function words**, là những từ ít mang nghĩa nội dung nhưng lại đóng vai trò quan trọng trong quan hệ cú pháp của câu, như on (ở trên), beside (bên cạnh), he (ông ấy), and (và).



Từ loại trong tiếng Anh

	Tag	Description	Example
Open Class	ADJ	Adjective: noun modifiers describing properties	<i>red, young, awesome</i>
	ADV	Adverb: verb modifiers of time, place, manner	<i>very, slowly, home, yesterday</i>
	NOUN	words for persons, places, things, etc.	<i>algorithm, cat, mango, beauty</i>
	VERB	words for actions and processes	<i>draw, provide, go</i>
	PROPN	Proper noun: name of a person, organization, place, etc..	<i>Regina, IBM, Colorado</i>
	INTJ	Interjection: exclamation, greeting, yes/no response, etc.	<i>oh, um, yes, hello</i>
Closed Class Words	ADP	Adposition (Preposition/Postposition): marks a noun's spacial, temporal, or other relation	<i>in, on, by under</i>
	AUX	Auxiliary: helping verb marking tense, aspect, mood, etc.,	<i>can, may, should, are</i>
	CCONJ	Coordinating Conjunction: joins two phrases/clauses	<i>and, or, but</i>
	DET	Determiner: marks noun phrase properties	<i>a, an, the, this</i>
	NUM	Numeral	<i>one, two, first, second</i>
	PART	Particle: a preposition-like form used together with a verb	<i>up, down, on, off, in, out, at, by</i>
	PRON	Pronoun: a shorthand for referring to an entity or event	<i>she, who, I, others</i>
Other	SCONJ	Subordinating Conjunction: joins a main clause with a subordinate clause such as a sentential complement	<i>that, which</i>
	PUNCT	Punctuation	<i>; , ()</i>
	SYM	Symbols like \$ or emoji	<i>\$, %</i>
	X	Other	<i>asdf, qwfg</i>

Universal Dependencies tagset (Nivre et al., 2016)



Penn Tree bank

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coord. conj.	<i>and, but, or</i>	NNP	proper noun, sing.	<i>IBM</i>	TO	"to"	<i>to</i>
CD	cardinal number	<i>one, two</i>	NNPS	proper noun, plu.	<i>Carolinas</i>	UH	interjection	<i>ah, oops</i>
DT	determiner	<i>a, the</i>	NNS	noun, plural	<i>llamas</i>	VB	verb base	<i>eat</i>
EX	existential 'there'	<i>there</i>	PDT	predeterminer	<i>all, both</i>	VBD	verb past tense	<i>ate</i>
FW	foreign word	<i>mea culpa</i>	POS	possessive ending	<i>'s</i>	VBG	verb gerund	<i>eating</i>
IN	preposition/ subordin-conj	<i>of, in, by</i>	PRP	personal pronoun	<i>I, you, he</i>	VBN	verb past participle	<i>eaten</i>
JJ	adjective	<i>yellow</i>	PRP\$	possess. pronoun	<i>your, one's</i>	VBP	verb non-3sg-pr	<i>eat</i>
JJR	comparative adj	<i>bigger</i>	RB	adverb	<i>quickly</i>	VBZ	verb 3sg pres	<i>eats</i>
JJS	superlative adj	<i>wildest</i>	RBR	comparative adv	<i>faster</i>	WDT	wh-determ.	<i>which, that</i>
LS	list item marker	<i>1, 2, One</i>	RBS	superlatv. adv	<i>fastest</i>	WP	wh-pronoun	<i>what, who</i>
MD	modal	<i>can, should</i>	RP	particle	<i>up, off</i>	WP\$	wh-possess.	<i>whose</i>
NN	sing or mass noun	<i>llama</i>	SYM	symbol	<i>+, %, &</i>	WRB	wh-adverb	<i>how, where</i>

Marcus, Mitchell, Beatrice Santorini, and Mary Ann Marcinkiewicz. "Building a large annotated corpus of English: The Penn Treebank." (1993).



Từ loại trong tiếng Việt

Định từ
Phó từ
Kết từ
tình thái từ

Danh từ	nhà, đất, người,
Đại từ	tao, nó, đây, đó,
Động từ	đi, bò,
Tính từ	đẹp, xấu, ...
Lượng từ	các, những,
Chỉ từ	này, kia, nọ,
Tiền phó từ	đã, sẽ, ...
Hậu phó từ	rồi, hết, ra,
Giới từ	của, ...
liên từ	và, với, ...
trợ từ	cũng, nhưng,
Tiểu từ	à, ôi,

Nguyễn Công Hồng, Về vấn đề phân định từ loại trong tiếng Việt. T/c Ngôn ngữ số 2, 2003



Part of speech tagging

- **Part-of speech tagging (POS-TAGGING)** là bài toán gán nhãn từ loại cho từng từ trong một văn bản.
- Tagging là một bài toán khử nhập nhằng nghĩa: một từ có thể có nhiều từ loại khác nhau. Mục tiêu bài toán là tìm ra từ loại đúng nhất cho một từ trong một ngữ cảnh cụ thể.

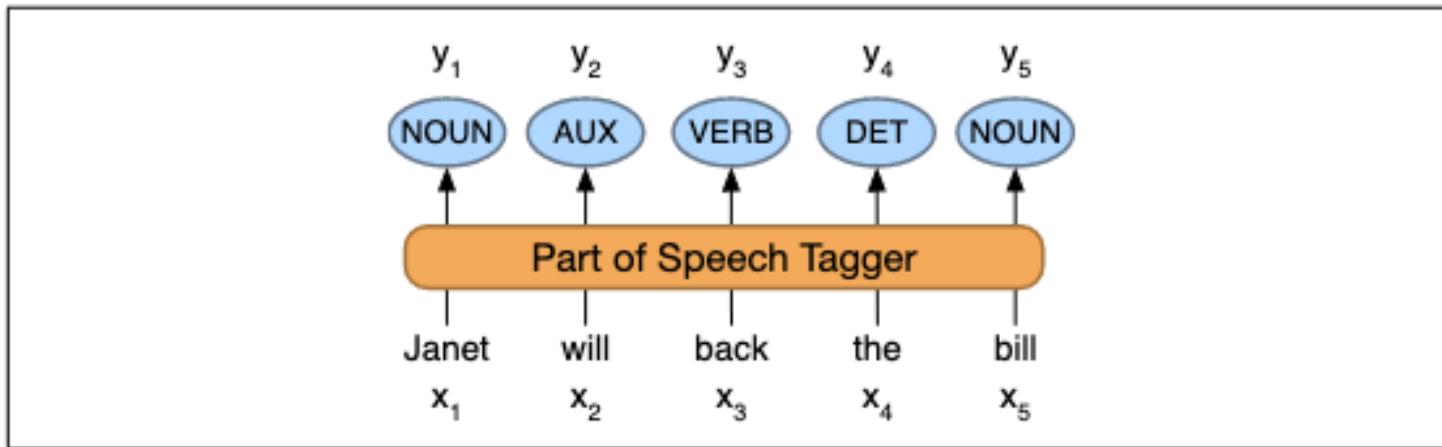
VD: từ loại cho từ “book”

book that flight → book được gán nhãn từ loại là động từ
hand me that book. → book được gán nhãn từ loại là danh từ



Part of speech tagging

“Janet will back the bill”





Vai trò của Part of speech tagging

- Bài toán phân tích ngữ pháp gồm:
 - *Phân tích từ pháp*: xác định từ loại của các từ trong câu.
 - *Phân tích cú pháp*: Xây dựng nên cây cú pháp cho câu, hoặc tìm ra mối quan hệ giữa các thành phần trong câu.



Một số bộ dữ liệu về Part of speech

- Tiếng Anh:
 - Universal Dependencies tagset (Nivre et al., 2016)
 - Penn Treebank P.O.S. Tags (Marcus et al., 1993)
- Tiếng Việt:
 - VLSP 2013 dataset for Word segmentation and POS Tagging (<https://vlsp.org.vn/resources-vlsp2013>)
 - Bộ dữ liệu NII-VTB (<https://github.com/mynlp/niivtb>)