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

6. Conditional Random Field (CRF)

Conditional Random Fields (CRF)

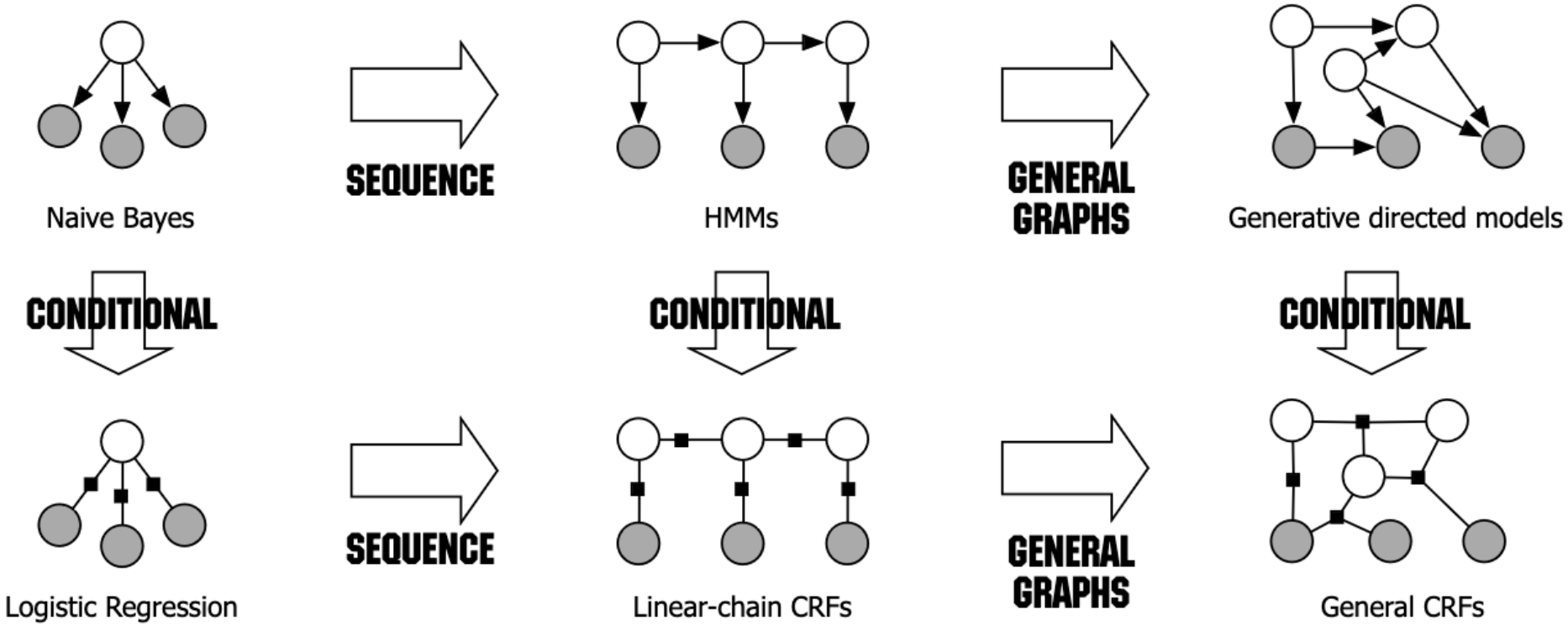
- Input: chuỗi văn bản đầu vào $X = x_1 .... x_n$ có độ dài là n.
- Output: chuỗi đầu ra $Y = y_1 ... y_n$.

- Xác suất của đầu ra Y dựa vào X là: $\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X)$

$$\begin{aligned}\hat{Y} &= \underset{Y}{\operatorname{argmax}}\, p(Y|X) \\ &= \underset{Y}{\operatorname{argmax}}\, p(X|Y)p(Y) \\ &= \underset{Y}{\operatorname{argmax}} \prod_i p(x_i|y_i) \prod_i p(y_i|y_{i-1})\end{aligned}$$

Trong HMM, ta tính P(Y|X) dựa vào công thức likelihood

CRF and HMM



Conditional Random Field là một phiên bản mở rộng của Logistic Regression

Xác suất đầu ra Y-hat

$$
\begin{aligned}
\hat{Y} &= \underset{Y \in \mathscr{Y}}{\operatorname{argmax}} P(Y|X) \\
&= \underset{Y \in \mathscr{Y}}{\operatorname{argmax}} \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right) \\
&= \underset{Y \in \mathscr{Y}}{\operatorname{argmax}} \exp\left(\sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)\right) \\
&= \underset{Y \in \mathscr{Y}}{\operatorname{argmax}} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \\
&= \underset{Y \in \mathscr{Y}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, X, i)
\end{aligned}
$$

*CRF không tính xác suất đầu ra cho từng nhãn như HMM, mà CRF sẽ **ước lượng** trên một tập tập các "đặc trưng". Các "đặc trưng" này được kết hợp lại và chuẩn hoá để tạo thành một xác suất **trên toàn chuỗi**.*

Tính P(Y|X)

- A CRF is a log-linear model that assigns a probability to an entire output (tag) sequence Y, out of all possible sequences Y, given the entire input (word) sequence X.
- In a CRF, the function F maps an entire input sequence X and an entire output sequence Y to a feature vector. Let's assume we have K features, with a weight w_k for each feature F_k.

$$p(Y|X) = \frac{\exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)}{\sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)}$$

Chuẩn hoá

$$p(Y|X) = \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right)$$

$$Z(X) = \sum_{Y' \in \mathcal{Y}} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y')\right)$$

Tìm giá trị tối ưu cho CRF

- Xác suất đầu ra của Y: $\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X)$

$$\begin{aligned}
\hat{Y} &= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} P(Y|X) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \frac{1}{Z(X)} \exp\left(\sum_{k=1}^{K} w_k F_k(X,Y)\right) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \exp\left(\sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)\right) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{k=1}^{K} w_k \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i) \\
&= \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{i=1}^{n} \sum_{k=1}^{K} w_k f_k(y_{i-1}, y_i, X, i)
\end{aligned}$$

Sử dụng giải thuật Viterbi dựa trên mô hình HMM:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, a_{ij}\, b_j(o_t); \quad 1 \le j \le N, 1 < t \le T$$

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i)\, P(s_j|s_i)\, P(o_t|s_j) \quad 1 \le j \le N, 1 < t \le T$$

Sử dụng giải thuật Viterbi dựa trên mô hình CRF:

$$v_t(j) = \max_{i=1}^{N} v_{t-1}(i) \sum_{k=1}^{K} w_k f_k(y_{t-1}, y_t, X, t) \quad 1 \le j \le N, 1 < t \le T$$

Tính F_k

- F_k được gọi là **Global features,** là tổng hợp của các local features ứng với mỗi vị trí i trong Y.

$$F_k(X,Y) = \sum_{i=1}^{n} f_k(y_{i-1}, y_i, X, i)$$

- Each of these local features f_k in a linear-chain CRF is allowed to make use of the **current output token y_i**, the **previous output token y_{i-1}**, the **entire input string X** (or any subpart of it), and the current position i.

Xây dựng features cho bài toán POS Tagging

- Feature templates:

$$\langle y_i, x_i \rangle, \langle y_i, y_{i-1} \rangle, \langle y_i, x_{i-1}, x_{i+2} \rangle$$

VD: Janet/NNP will/MD back/VB the/DT bill/NN, xi = “back”

Features templates:

f_{3743}: y_i = VB and x_i = back
f_{156}: y_i = VB and y_{i-1} = MD
f_{99732}: y_i = VB and x_{i-1} = will and x_{i+2} = bill

Xây dựng features cho bài toán POS Tagging (tt)

- Word shapes: Map words to simplified representation that encodes attributes such as **length**, **capitalization**, **numerals**, *Greek letters*, **internal punctuation**, etc.

x_i contains a particular prefix (perhaps from all prefixes of length ≤ 2)
x_i contains a particular suffix (perhaps from all suffixes of length ≤ 2)
x_i's word shape
x_i's short word shape

prefix(x_i) = `w`
prefix(x_i) = `we`
suffix(x_i) = `ed`
suffix(x_i) = `d`
word-shape(x_i) = `xxxx-xxxxxxx`
short-word-shape(x_i) = `x-x`

Xây dựng features cho bài toán NER

- Sử dụng **gazetteer** và **name-list**.
  - Gazetteer gồm một danh sách các từ điển liên quan đến vị trí địa lý (LOC and GPE).
  - Name-list là từ điển các tên riêng.

identity of w_i, identity of neighboring words
embeddings for w_i, embeddings for neighboring words
part of speech of w_i, part of speech of neighboring words
presence of w_i in a **gazetteer**
w_i contains a particular prefix (from all prefixes of length ≤ 4)
w_i contains a particular suffix (from all suffixes of length ≤ 4)
word shape of w_i, word shape of neighboring words
short word shape of w_i, short word shape of neighboring words
gazetteer features

prefix(x_i) = L	suffix(x_i) = tane
prefix(x_i) = L'	suffix(x_i) = ane
prefix(x_i) = L'0	suffix(x_i) = ne
prefix(x_i) = L'0c	suffix(x_i) = e
word-shape(x_i) = X'Xxxxxxxx	short-word-shape(x_i) = X'Xx

Xây dựng features cho bài toán NER

Words	POS	Short shape	Gazetteer	BIO Label
Jane	NNP	Xx	0	B-PER
Villanueva	NNP	Xx	1	I-PER
of	IN	x	0	O
United	NNP	Xx	0	B-ORG
Airlines	NNP	Xx	0	I-ORG
Holding	NNP	Xx	0	I-ORG
discussed	VBD	x	0	O
the	DT	x	0	O
Chicago	NNP	Xx	1	B-LOC
route	NN	x	0	O
.	.	.	0	O

Chicago and Villanueva là "location" từ gazetteer