

# עיבוד שפה טבעיות ש9:

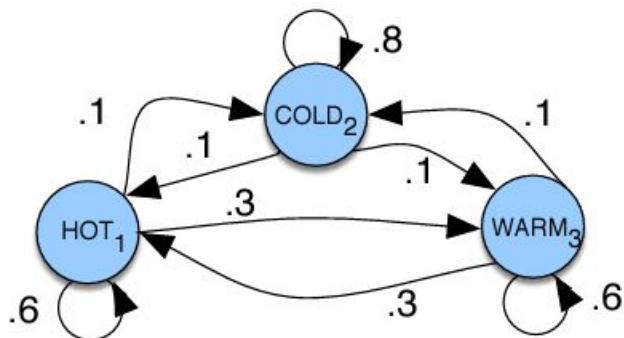
## תיוג רצפים - HMM, CRF

פרק 8: Eisenstein 7-8

## זיכרון - הגדרת הבעה

- קלט: מסמך D בעל כ- מילים
- פלט: כ- תגים, כל אחד מהם בעל ערך מתוך קבוצה משותפת (או לעיתים, שונה)
- אוסף התגים ה"אמיתי", ה-Y, תלוי למעשה באורך הטקסט, למשל (m)Y
- מה שמנגדיר בעיית חיפוש (search)

# שרשרת מרקוב (Markov Chain)



$$\pi(\text{HOT}) = 0.3$$

**0.0001296**

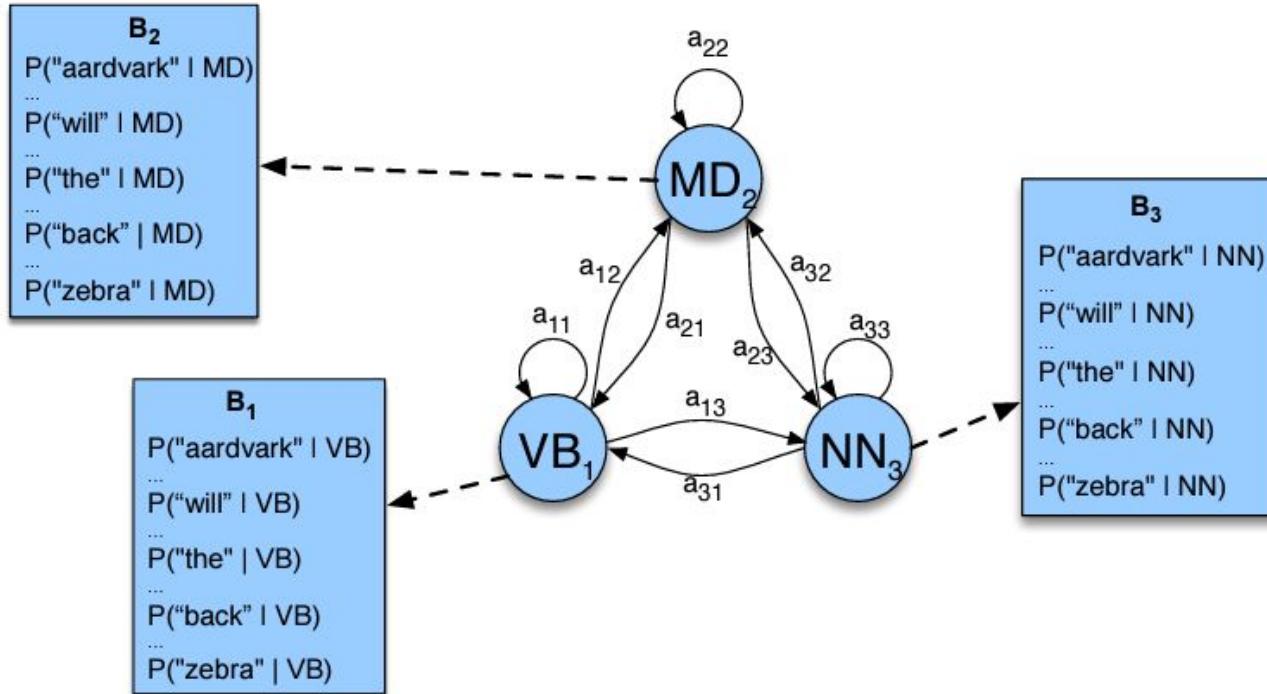
- מודל לתיאור מעבר בשלבים דרך **מצבים**  $q_i$
- בכל נקודה בזמן, השלב מתקיים ע"י **מעבר** מהמצב הקודם
- הנחה היסוד המרковית:** כל מצב תלוי רק במצב שקדם לו
$$P(q_i=a|q_{i-1}=a_1 \dots q_1) = P(q_i=a|q_{i-1})$$
- מה ההסתברות של רצף התוצאות הבא? (והאם יש לנו מספיק מידע?)

HOT WARM WARM COLD COLD WARM HOT

# מודל מركוב נסתיר (Hidden Markov Model - HMM)

- המשתנה הנצפה הוא **פלט** (output) של  **מצב** (state)
- מה שנסתיר/חבוי/סמי זה **המצבים**, במקורה שלנו התגים (או עם מזג האויר - "חזית קרה")
- הנחה יסוד: פלט תלוי רק במצב שלו  $P(o_1, \dots, o_T | q_1, \dots, q_i) = P(o_1 | q_i) \dots P(o_T | q_i)$

# מודל מركוב נסתן (Hidden Markov Model - HMM)



# מודל מרקוב נסתר (Hidden Markov Model - HMM)

- המשתנה הנצפה הוא **פלט** (output) של  **מצב** (state)
- מה שנסתור/חבוי/סמי זה המצבים, במקרה שלנו התגים
- הנחה יסוד ||: פלט תלוי רק במצב שלו  $P(o|q_i, q_1, \dots, q_{i-1}, o_1, \dots, o_{i-1}) = P(o_i|q_i)$
- ניתן ללמד הסתברויות מעבר (transition) ופליטה (emission) יחסית בקלות, בהינתן קורפו מתייג
- תהליך התיאוג הוא המסביר כאן, כבעיית **חיפוש / פענוח decoding** - בהינתן התצפויות, מצאו את המצבים
- קירוב בייזיאני

$$\hat{t}_{1:n} = \underset{t_1 \dots t_n}{\operatorname{argmax}} P(t_1 \dots t_n | w_1 \dots w_n) = \underset{t_1 \dots t_n}{\operatorname{argmax}} \frac{P(w_1 \dots w_n | t_1 \dots t_n) P(t_1 \dots t_n)}{P(w_1 \dots w_n)}$$

$$= \underset{t_1 \dots t_n}{\operatorname{argmax}} P(w_1 \dots w_n | t_1 \dots t_n) P(t_1 \dots t_n)$$

$$\approx \underset{t_1 \dots t_n}{\operatorname{argmax}} \prod_{i=1}^n \overbrace{P(w_i | t_i)}^{\text{emission transition}} \overbrace{P(t_i | t_{i-1})}^{\text{transition}}$$

# מרכיבי מודל מרקוב

$Q = q_1 q_2 \dots q_N$  a set of  $N$  states

$A = a_{11} \dots a_{ij} \dots a_{NN}$  a **transition probability matrix**  $A$ , each  $a_{ij}$  representing the probability of moving from state  $i$  to state  $j$ , s.t.  $\sum_{j=1}^N a_{ij} = 1 \quad \forall i$

$O = o_1 o_2 \dots o_T$  a sequence of  $T$  **observations**, each one drawn from a vocabulary  $V = v_1, v_2, \dots, v_V$

$B = b_i(o_t)$  a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation  $o_t$  being generated from a state  $q_i$

$\pi = \pi_1, \pi_2, \dots, \pi_N$  an **initial probability distribution** over states.  $\pi_i$  is the probability that the Markov chain will start in state  $i$ . Some states  $j$  may have  $\pi_j = 0$ , meaning that they cannot be initial states. Also,  $\sum_{i=1}^n \pi_i = 1$

# אלגוריתם ויטרבי (Viterbi) (פרק 8 עמ' 12 ב-SLP)

- מטרה: למצוא את המסלול הסביר ביותר בין המצבים שיפיקו את סדרת התצפויות

$$v_t(j) = \max_{q_1, \dots, q_{t-1}} P(q_1 \dots q_{t-1}, o_1, o_2 \dots o_t, q_t = j | \lambda)$$

$$v_t(j) = \max_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t)$$

מצבים קודמים אפשריים

הסיכוי שהיינו במצב הקודם

הסיכוי שעברנו מהמצב הקודם

הסיכוי שפלטנו את המילה הנכנית

- مبוסס על **תכנות דינמי**
  - מה מתאר מצב בינים?
  - מה כלל המעבר?

# אלגוריתם ויטרבי (Viterbi) (פרק 8 עמ' 12 ב-SLP)

- מטרה: למצוא את המסלול הסביר ביותר בין המצבים שיפיקו את סדרת התצפויות

$$v_t(j) = \max_{q_1, \dots, q_{t-1}} P(q_1 \dots q_{t-1}, o_1, o_2 \dots o_t, q_t = j | \lambda)$$

$$v_t(j) = \max_{i=1}^N v_{t-1}(i) a_{ij} b_j(o_t)$$

מצבים קודמים אפשריים

הסיכוי שהיינו במצב הקודם

הסיכוי שעברנו מהמצב הקודם

הסיכוי שפלטנו את המילה הנכנית

מבוסס על **תכנות דינמי**

- מה מתאר מצב בינים?
- מה כלל המעבר?
- מה תנאי הפתיחה?

ניתן לעבור ללוג-הסתברויות

אם ויטרבי תמיד עובר דרך המקסימום

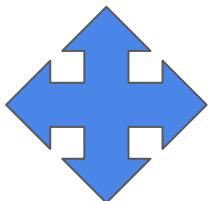
של כל השלבים?

## עוד אלגוריתמים בסביבת HMM

- "קדימה" (Forward) - לחישוב הסתברות כוללת של משפט
- "קדימה-אחורה" (Forward-backward / Baum-Welch) - לאימון בלתי-מופיע
  - בינטיים לא רלוונטי

# הרצאה (ЛОЧ)

Transitions	-> V	-> N	-> M	-> P	-> J			Emissions	שרה	שיר	שם
V ->	0.05	0.35	0.25	0.2	0.15			V	0.02	0.002	0.001
N ->	0.6	0.05	0.05	0.1	0.2			N	0.005	0.01	0
M ->	0.6	0.1	0.1	0.15	0.05			M	0.03	0.02	0.001
P ->	0.05	0.5	0.3	0.01	0.14			P	0	0.00001	0
J ->	0.5	0.05	0.1	0.2	0.15			J	0	0	0.04
START ->	0.15	0.25	0.25	0.3	0.05						
Transitions (log)	-> V	-> N	-> M	-> P	-> J			Emissions (log)	שרה	שיר	שם
V ->	-2.996	-1.050	-1.386	-1.609	-1.897			V	-3.912	-6.215	-6.908
N ->	-0.511	-2.996	-2.996	-2.303	-1.609			N	-5.298	-4.605	
M ->	-0.511	-2.303	-2.303	-1.897	-2.996			M	-3.507	-3.912	-6.908
P ->	-2.996	-0.693	-1.204	-4.605	-1.966			P		-11.513	
J ->	-0.693	-2.996	-2.303	-1.609	-1.897			J			-3.219
	-1.897	-1.386	-1.386	-1.204	-2.996						



$\Rightarrow J$	$\cap NL$	$\Rightarrow N$	$\cap \emptyset$	$\Rightarrow V$	$\cap \emptyset$	$\Rightarrow M$	$\cap \emptyset$	
$M$	-22.033	$M$	-17.427	$V$	$\max \{-5.809 - 2.996 - 3.912, -6.685 - 0.511 - 3.912, -4.893 - 0.511 - 3.912\}$ line 12 $= -9.316$	$M$	-1.897 - 3.912 $= -5.859$	$\vee$
$\emptyset$		$V$	-14.971	$V$	$\max \{-5.809 - 1.050 - 5.298, -6.685 - 2.496 - 5.298, -4.893 - 2.303 - 5.298\}$ equal as above $= -12.157$		-6.685	$N$
$M$	$\max \{-17.427 - 1.386 - 6.908, -14.971 - 2.496 - 6.908, -14.614 - 2.303 - 6.908, -22.438 - 1.204 - 6.908\}$ $= -23.824$	$V$	-14.614	$V$	$\max \{-5.809 - 2.586 - 3.507, -6.685 - 2.496 - 3.507, -4.893 - 2.303 - 3.507\}$ $= 10.702$		-4.893	$M$
$\emptyset$		$V$		$\emptyset$		$\emptyset$		$P$ on. $\nearrow \delta N$
$N$	-14.799	$\emptyset$		$\emptyset$		$\emptyset$	J	$\nearrow k \uparrow P$

# בעיות ב-HMM

- מילים לא ידועות
  - החלוקת מספקת?
- הנחת אי-התלוות בין מילים
- הנחת המרקבויות?

# שדות אקראיים מותנים (Conditional Random Fields - CRF)

משמעותו של שדה אקראי מותן הוא:  
שימו לב לשינוי הנטיצה הכללית -  $X$  במקומות  $d$   
או  $W$  (מילים, מסגר),  $W$  במקומות תטא

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

מה מצינים ה- $k$ ?

- עקרונית הוא תלוי בכל הקלט עבור כל הפלט:

(הgrsah המולטינומית של רגרסיה לוגיסטיבית,  
נפרט בשבוע הבא)

- מודל דיסקרימינטיבי

- ספציפיתナルם של שרשרות קוויות (Linear chain CRF)

# שדות אקראיים מותנים (Conditional Random Fields - CRF)

- מודל דיסקרימינטיבי
- ספציפית נלמד שא"מ של שרשרות קוויות (Linear chain CRF)
- עקרונית הוא תלוי בכל הקלט עברו כל הפלט
- מעשית, מפרקים לכל Tag כתלי בתג הקודם, **במקום עצמו, ובכל הקלט** (לצורך פיצרים)

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

## פתרונות אפשריים ובלתי-אפשריים ב-LC-CRF

- "המילה הנוכחית מתחילה ב-'הט'"
- "המילה הראשונה במשפט היא 'the'"
- "יש במשפט חמישה מופעים של המילה 'את'"
- "יש במשפט שלושה כינויי גוף"
- "התג הרביעי במשפט הוא שם-תואר"
- "המילה הבאה מתחילה ב-'ל-'"
- "אחד משלוש המילים האחרונות היא פועל"

# tabniot pivrim (Feature templates)

- לייצרת מרחב קומבינטורי של פיצרים אפשריים
  - תבניות מקובלות:
- $\langle y_i, x_i \rangle$
- $\langle y_i, y_{i-1} \rangle$
- $\langle y_i, x_i, x_{i-1} \rangle$
- כמה פיצרים נוצרים מכל אחת מהtabniot האלה?
- חלק מילים
  - צורת מילים
- אנגלית: אות קטנה ל-x, גדולה ל-p, ספרה ל-X, קר שלמשל UB-40-XX הופך ל-pp-XX
  - עברית: הכללה לאותיות איתן?

## (Inference) הסקה - CRF

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

אפשר עם ויטרבי:

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



# CRF - למידה

- קדימה-אחרה: מעדכנים פרמטרים להסתבריות מעבר ופליטה עם כל דוגמא באופן איטרטיבי
  - עבור HMM רגיל - נועד לעדכן פרמטרים בעיקר **כשאין ידע על המצביעים**
  - ל-CRF - מקל על חישובי הגרדיאנטים
- (נספח A.5 ב-SLP למי שרצה לדעת יותר)

## התאמות לזיהוי ישוויות

- רשימת כלליים עבור HMM ו-CRF שמוגדים שנקבל ישוויות הולמות: (לוח)

