CMPT-882: Statistical Learning of Natural Language

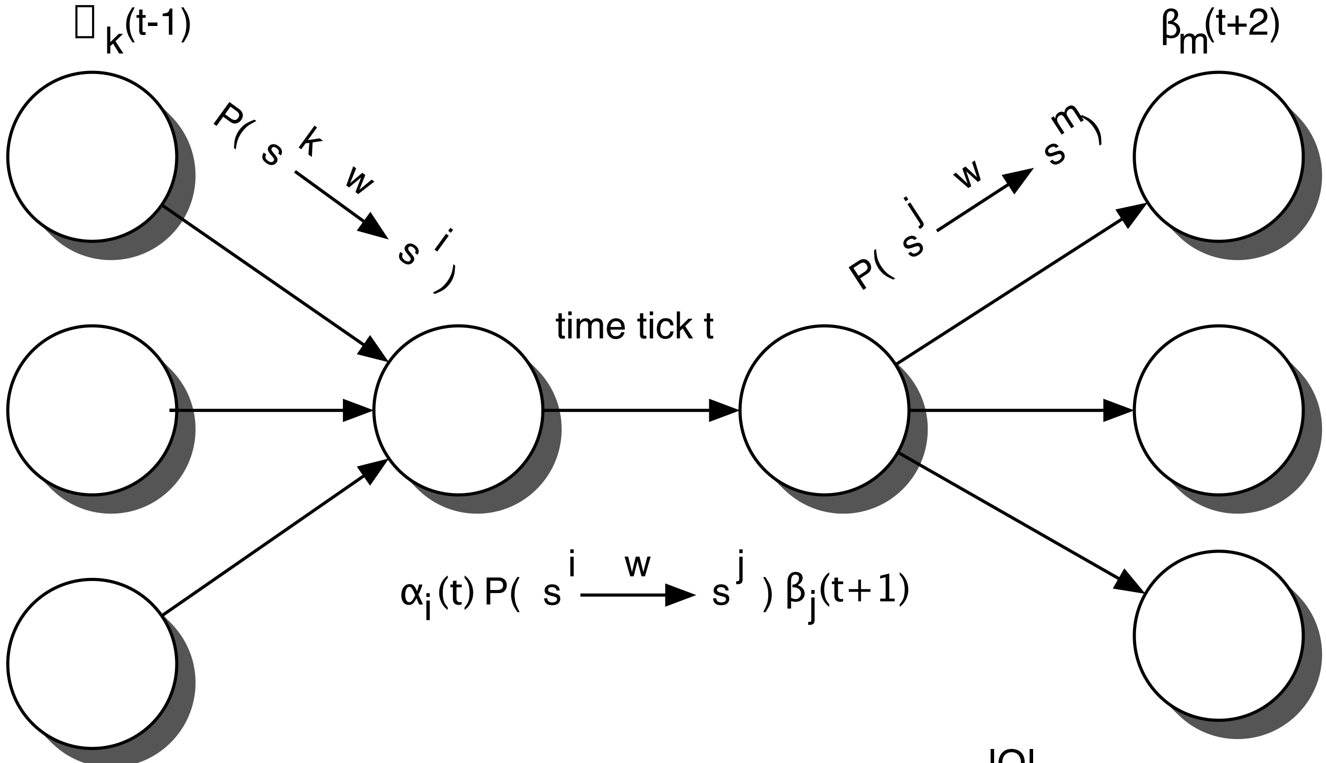
Lecture #5

Anoop Sarkar
anoop@cs.sfu.ca
http://www.sfu.ca/˜anoop

- Previous and current homework

- HMM review

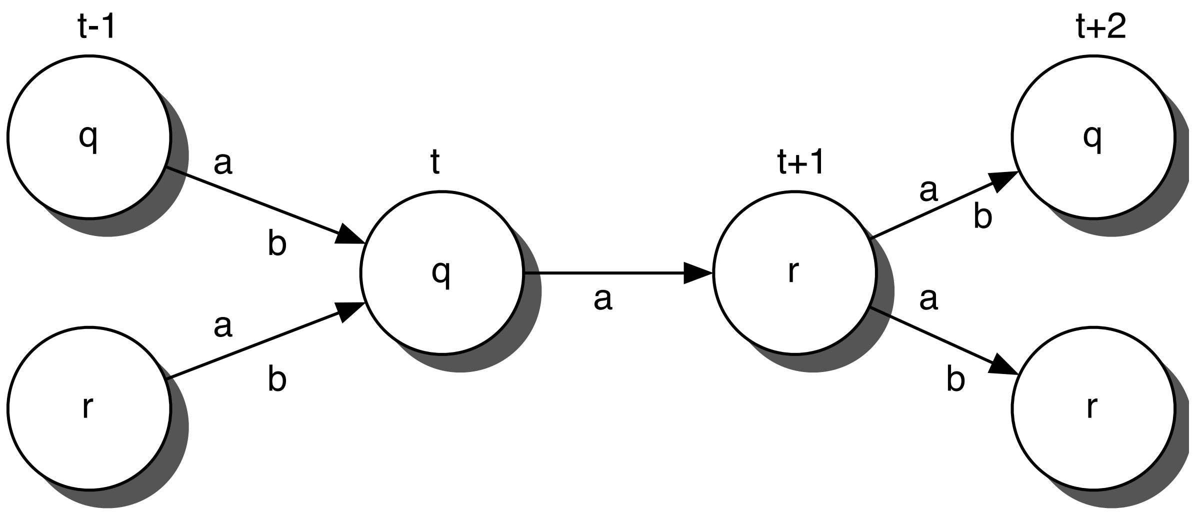
- Elworthy (1994) and Merialdo (1994)



$$\alpha_i(t) = \sum_{k=1}^{|Q|} \alpha_k(t-1)\,P(s^k \xrightarrow{w} s^i)$$

$$\beta_j(t+1) = \sum_{m=1}^{|Q|} P(s^j \xrightarrow{w} s^m)\,\beta_m(t+2)$$

$$C(s^i \xrightarrow{w} s^j) = \frac{1}{P(w_{1,n})} \sum_{t=1}^{n} \alpha_i(t)\,P(s^i \xrightarrow{w} s^j)\,\beta_j(t+1)$$



$$
\begin{aligned}
\alpha_q(t) &= \alpha_q(t-1)P(a,q \mid q) + \alpha_q(t-1)P(b,q \mid q) + \\
&\quad \alpha_r(t-1)P(a,q \mid r) + \alpha_r(t-1)P(b,q \mid r) \\
\beta_r(t+1) &= P(a,q \mid r)\beta_q(t+2) + P(b,q \mid r)\beta_q(t+2) + \\
&\quad P(a,r \mid r)\beta_r(t+2) + P(b,r \mid r)\beta_r(t+2) \\
C(q \xrightarrow{a} r) &= \frac{1}{P(w_{1,n})} \sum_{t=1}^{n} \alpha_q(t)P(a,r \mid q)\beta_r(t+1)
\end{aligned}
$$

Forward-Backward Algorithm

- Set initial transition probabilities to appropriate values (usually random)

- Compute $C(s^i \stackrel{w}{\rightarrow} s^j)$ for each state i and then
$P_e(s^i \stackrel{w}{\rightarrow} s^j) = \frac{C(s^i \stackrel{w}{\rightarrow} s^j)}{\sum_{k,w'} C(s^i \stackrel{w'}{\rightarrow} s^k)}$

- Compute likelihood $P(w_{1,n}) = \beta_{s^1}(1)$; iterate until likelihood is maximized (or entropy is minimized)

- Here we considered the case for one training sentence $w_{1,n}$. For a whole corpus, $\prod_k P(w^k_{1,n})$ is the likelihood of the entire corpus with k sentences

Elworthy (1994)

- Using the Forward-Backward Algorithm to decrease human supervision

- Does Baum-Welch Re-estimation help taggers? (1994). David Elworthy. *Proceedings of 4th ACL Conf on ANLP*, Stuttgart. pp. 53-58.

Elworthy (1994)

Lexicon	Transitions
D0 : Fully Supervised $\frac{f(t^i,w)}{f(t^i)}$ D1 : $w \mid t$ and $\mathsf{order}(w \mid t)$ D2 : $p(w \mid t) = p(t)$ D3 : $p(w \mid t) = \frac{1}{N_t}$	T0 : Fully Supervised $\frac{f(t^i,t^j)}{f(t^i)}$ T1 : $\frac{1}{N_q}$

Elworthy (1994)

- Combinations (e.g. D0+T0) and their performance — Table 1
- Patterns of Re-estimation — Fig 1 and Table 2–3

Merialdo (1994)

- Viterbi tagging vs. ML tagging: best tag per word in a sequence as opposed to best tag sequence)

$$\Phi(W)_i = \underset{t}{\text{arg max}}\; p(t_i = t \mid w) = \underset{t}{\text{arg max}} \sum_{T:t_i=t} p(W,T)$$

- Table 2 — HMM training from various initial starting conditions

- Constrained HMM training — tw constraint and t constraint