• Previous and current homework

• HMM review

• Elworthy (1994) and Merialdo (1994)
\[ \alpha_i(t) = \sum_{k=1}^{\left|Q\right|} \alpha_k(t-1) P(s^k w \rightarrow s^i) \]

\[ \beta_j(t+1) = \sum_{m=1}^{\left|Q\right|} P(s^j w \rightarrow s^m) \beta_m(t+2) \]

\[ C(s^i w \rightarrow s^j) = \frac{1}{P(w_{1,n})} \sum_{t=1}^{n} \alpha_i(t) P(s^i w \rightarrow s^j) \beta_j(t+1) \]
\[ \alpha_q(t) = \alpha_q(t-1)P(a,q \mid q) + \alpha_q(t-1)P(b,q \mid q) + \alpha_r(t-1)P(a,q \mid r) + \alpha_r(t-1)P(b,q \mid r) \]

\[ \beta_r(t+1) = P(a,q \mid r)\beta_q(t+2) + P(b,q \mid r)\beta_q(t+2) + P(a,r \mid r)\beta_r(t+2) + P(b,r \mid r)\beta_r(t+2) \]

\[ C(q \xrightarrow{a} r) = \frac{1}{P(w_{1,n})} \sum_{t=1}^{n} \alpha_q(t)P(a,r \mid q)\beta_r(t+1) \]
Forward-Backward Algorithm

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

- Compute $C(s^i \xrightarrow{w} s^j)$ for each state $i$ and then

$$P_e(s^i \xrightarrow{w} s^j) = \frac{C(s^i \xrightarrow{w} s^j)}{\sum_{k,w'} C(s^i \xrightarrow{w'} s^k)}$$

- Compute likelihood $P(w^1_{1,n}) = \beta_{s1}(1)$; iterate until likelihood is maximized (or entropy is minimized)

- Here we considered the case for one training sentence $w^1_{1,n}$. 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

Elworthy (1994)

<table>
<thead>
<tr>
<th>Lexicon</th>
<th>Transitions</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>D0</strong> : Fully Supervised ( \frac{f(t_i, w)}{f(t_i)} )</td>
<td><strong>T0</strong> : Fully Supervised ( \frac{f(t_i, t_j)}{f(t_i)} )</td>
</tr>
<tr>
<td><strong>D1</strong> : ( w \mid t ) and ( \text{order}(w \mid t) )</td>
<td><strong>T1</strong> : ( \frac{1}{N_q} )</td>
</tr>
<tr>
<td><strong>D2</strong> : ( p(w \mid t) = p(t) )</td>
<td></td>
</tr>
<tr>
<td><strong>D3</strong> : ( p(w \mid t) = \frac{1}{N_t} )</td>
<td></td>
</tr>
</tbody>
</table>
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

\[ \Phi(W)_i = \arg \max_t p(t_i = t \mid w) = \arg \max_t \sum_{T : t_i = t} p(W, T) \]

- Table 2 — HMM training from various initial starting conditions

- Constrained HMM training — \( tw \) constraint and \( t \) constraint