# (520|600).666 Information Extraction

## Homework # 3

Due February 27, 2024.

Review §2.7 in the Jelinek book before starting this homework.

1. Consider the HMM of Homework #2 again, with the parameter matrices $p(s'|s)$, $q(s'|s)$, $q(\mathbf{0}|s \to s')$ and $q(\mathbf{1}|s \to s')$ as previously specified.

   Perform the following calculations *by hand*, again retaining adequate numerical precision of the intermediate answers.

   (a) Draw a 4-stage trellis for this HMM, showing *only* the paths which could have resulted in the output **0110**. (You may copy your solution from Homework #2.)

   (b) Calculate the *a posteriori* probabilities $P(t^i = t \,|\, y_1y_2y_3y_4 = \mathbf{0110}, s_0 = 1)$ for each arc in the trellis. Show your answers on the trellis.

   (c) Based on your calculations in (b), compute the *expected* counts $c(y,t)$ of each non-null arc, and *reestimate* the emission probability matrices $q(\mathbf{0}|s \to s')$ and $q(\mathbf{1}|s \to s')$.

   (d) Based on your calculations in (b), compute the *expected* counts $c(t)$ of each transition, and *reestimate* the transition probability matrices $p(s'|s)$ and $q(s'|s)$.

   **Note**: The HMM of Figure 2.8 and your trellis have up to $|\mathcal{Y}|$ distinct non-null *arcs* from any $s$ to $s'$, one *arc* per output symbol, while Figure 2.11 and the matrix $p(s'|s)$ have only one non-null *transition* from $s$ to $s'$. Make sure to sum the *arc counts* $c^*(y,t)$ from the trellis over all $y$ to obtain the *transition count* $c^*(t)$ for computing $p(s'|s)$.

2. Consider the HMM of Homework #2 once again (!) and the Viterbi path you calculated in Homework #2 (f) for the output **0110**.

   (a) Compute the *Viterbi* or "hard" counts $\hat{c}(y,t)$ of each arc of the HMM. How many $\hat{c}(y,t)$ are nonzero?

   (b) Replace $c^*(y,t)$ in Equations (36) on p33 with these $\hat{c}(y,t)$ and *reestimate* the HMM emission probability matrices $q(\mathbf{0}|s \to s')$ and $q(\mathbf{1}|s \to s')$.

   (c) Compute the *Viterbi* or "hard" counts $\hat{c}(t)$ of each arc of the HMM. How many $\hat{c}(t)$ are nonzero?

   (d) Replace $c^*(t)$ in Equations (37) on p33 with these $\hat{c}(t)$ and *reestimate* the HMM transition probability matrices $p(s'|s)$ and $q(s'|s)$. (See the Note above.)

Review the remainder of Chapter 2 after finishing this homework, paying particular attention to §2.8, which will be useful for Project #1.