Example of the Baum-Welch Algorithm

Larry Moss

Q520, Spring 2008

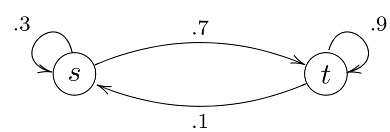
1 Our corpus c

We start with a very simple corpus. We take the set Y of unanalyzed words to be $\{ABBA, BAB\}$, and c to be given by $c(ABBA) = 10$, $c(BAB) = 20$.

Note that the total value of the corpus is $\sum_{u \in Y} c(u) = 10 + 20 = 30$.

2 Our first HMM h_1

The first HMM h_1 is arbitrary. To have definite numbers around, we select some.



Starting probability of s is .85, of t is .15. In s, $\Pr(A) = .4$, $\Pr(B) = .6$. In t, $\Pr(A) = .5$, $\Pr(B) = .5$.

3 $\alpha(y, j, s)$

Let $y \in Y$, and let n be the length of y. For $1 \leq j \leq n$ and s one of our states, we define $\alpha(y, j, s)$ to be the probability in the space of *analyzed words* that the first j symbols match those of y, and the ending state is s.

This is related to the computations in the Forward Algorithm because the overall probability of y in the HMM h is $\sum_{u \in S} \alpha(y, n, u)$. This is number is written as $\Pr_h(y)$.

Writing y as $A_1 A_2 \cdots A_n$, we have

$$\begin{array}{lcl} \alpha(y, 1, s) & = & \text{start}(s)\text{out}(s, A_1) \\ \alpha(y, j+1, s) & = & \sum_{t \in S} \alpha(y, j, t)\text{go}(t, s)\text{out}(s, A_{j+1}) \end{array}$$

$ABBA$

$\alpha(ABBA, 1, s) = (.85)(.4) = 0.34$.
$\alpha(ABBA, 1, t) = (.15)(.5) = 0.08$.
$\alpha(ABBA, 2, s) = (0.34)(.3)(.6) + (0.08)(.1)(.6) = 0.06120 + 0.00480 = 0.06600$.
$\alpha(ABBA, 2, t) = (0.34)(.7)(.5) + (0.08)(.9)(.5) = 0.11900 + 0.03600 = 0.15500$.
$\alpha(ABBA, 3, s) = (0.06600)(.3)(.6) + (0.15500)(.1)(.6) = 0.01188 + 0.00930 = 0.02118$.

$\alpha(ABBA, 3, t) = (0.06600)(.7)(.5) + (0.15500)(.9)(.5) = 0.02310 + 0.06975 = 0.09285.$
$\alpha(ABBA, 4, s) = (0.02118)(.3)(.4) + (0.09285)(.1)(.4) = 0.00254 + 0.00371 = 0.00625.$
$\alpha(ABBA, 4, t) = (0.02118)(.7)(.5) + (0.09285)(.9)(.5) = 0.00741 + 0.04178 = 0.04919.$
Total probability of $ABBA$ is $0.00625 + 0.04919 = 0.05544$.

BAB

$\alpha(BAB, 1, s) = (.85)(.6) = 0.51.$
$\alpha(BAB, 1, t) = (.15)(.5) = 0.08.$
$\alpha(BAB, 2, s) = (0.51)(.3)(.4) + (0.08)(.1)(.4) = 0.0612 + 0.0032 = 0.0644.$
$\alpha(BAB, 2, t) = (0.51)(.7)(.5) + (0.08)(.9)(.5) = 0.1785 + 0.0360 = 0.2145.$
$\alpha(BAB, 3, s) = (0.06600)(.3)(.6) + (0.15500)(.1)(.6) = 0.01188 + 0.00930 = 0.0209.$
$\alpha(BAB, 3, t) = (0.0644)(.7)(.5) + (0.2145)(.9)(.5) = 0.0225 + 0.0965 = 0.1190.$
Total probability of BAB is $0.0209 + 0.1190 = 0.1399$.

3.1 The likelihood of the corpus using h_1

$L(c, h_1) = \Pr(ABBA)^{c(ABBA)} \cdot \Pr(BAB)^{c(BAB)} = 0.05544^{10} 0.1399^{20}$

It is easier to work with the log of this, and then
$\log L(c, h_1) = (10 * \log 0.05544) + (20 * \log 0.1399) = -68.2611$

4 The $\beta(y, j, s)$ values

Define $\beta(y, j, s)$ to be the following conditional probability:

Given that the jth state is s, the $(j+1)$st symbol will be A_{j+1}, the $(j+2)$nd will be A_{j+2}, ..., the nth will be A_n.

Writing y as $A_1 A_2 \cdots A_n$, our equations go backward:

$$\begin{aligned} \beta(y, n, s) &= 1 \\ \beta(y, j, s) &= \sum_{u \in S} \mathrm{go}(s, u)\mathrm{out}(u, A_{j+1})\beta(y, j+1, u) \end{aligned}$$

4.1 $\beta(ABBA, j, s)$ for $1 \le j \le 4$

$\beta(ABBA, 4, s) = 1.$
$\beta(ABBA, 4, t) = 1.$
$\beta(ABBA, 3, s) = \sum_{u \in S} \mathrm{go}(s, u)\mathrm{out}(u, A)\beta(ABBA, 4, u) = (.3)(.4)(1) + (.7)(.5)(1) = 0.47000$
$\beta(ABBA, 3, t) = \sum_{u \in S} \mathrm{go}(t, u)\mathrm{out}(u, A)\beta(ABBA, 4, u) = (.1)(.4)(1) + (.9)(.5)(1) = 0.49000$
$\beta(ABBA, 2, s) = \sum_{u \in S} \mathrm{go}(s, u)\mathrm{out}(u, B)\beta(ABBA, 3, u) = (.3)(.6)(0.47000) + (.7)(.5)(0.49000) = 0.25610$
$\beta(ABBA, 2, t) = \sum_{u \in S} \mathrm{go}(t, u)\mathrm{out}(u, B)\beta(ABBA, 3, u) = (.1)(.6)(0.47000) + (.9)(.5)(0.49000) = 0.24870$
$\beta(ABBA, 1, s) = \sum_{u \in S} \mathrm{go}(s, u)\mathrm{out}(u, B)\beta(ABBA, 2, u) = (.3)(.6)(0.25610) + (.7)(.5)(0.24870) = 0.13315$
$\beta(ABBA, 1, t) = \sum_{u \in S} \mathrm{go}(t, u)\mathrm{out}(u, B)\beta(ABBA, 2, u) = (.1)(.6)(0.25610) + (.9)(.5)(0.24870) = 0.12729$

4.2 $\beta(BAB, j, s)$ for $1 \leq j \leq 3$

$\beta(BAB, 3, s) = 1.$
$\beta(BAB, 3, t) = 1.$
$\beta(BAB, 2, s) = \sum_{u \in S} \text{go}(s, u)\text{out}(u, B)\beta(BAB, 3, u) = (.3)(.4)(1) + (.7)(.5)(1) = 0.53000$
$\beta(BAB, 2, t) = \sum_{u \in S} \text{go}(t, u)\text{out}(u, B)\beta(BAB, 3, u) = (.1)(.4)(1) + (.9)(.5)(1) = 0.51000$
$\beta(BAB, 1, s) = \sum_{u \in S} \text{go}(s, u)\text{out}(u, A)\beta(BAB, 2, u) = (.3)(.6)(0.53000) + (.7)(.5)(0.51000) = 0.24210$
$\beta(BAB, 1, t) = \sum_{u \in S} \text{go}(t, u)\text{out}(u, A)\beta(BAB, 2, u) = (.1)(.6)(0.53000) + (.9)(.5)(0.51000) = 0.25070$

5 $\gamma(y, j, s, t)$

Let $y \in Y$, and write y as $A_1 \cdots A_n$. We want the probability in the subspace $A(y)$ that an analyzed word has s as its jth state, (A_{j+1} as its $(j+1)$st symbol), and t as its $(j+1)$st state. (This only makes sense when $1 \leq j < n$.)

This probability is called $\gamma(y, j, s, t)$. It is given by

$$\gamma(y, j, s, t) = \frac{\alpha(y, j, s)\text{go}(s, t)\text{out}(t, A_{j+1})\beta(y, j+1, t)}{\Pr_h(y)}.$$

In other words, $\gamma(y, j, s, t)$ is the probability that a word in $A(y)$ has an s as its jth symbol and a t as its $(j+1)$st symbol.

It is important to see that for different unanalyzed words, say y and z, $\gamma(y, j, s, t)$ and $\gamma(z, j, s, t)$ are probabilities *in different spaces.*

For example,

$$\gamma(ABBA, 1, t, s) = \frac{\alpha(ABBA, 1, t)\text{go}(t, s)\text{out}(s, B)\beta(ABBA, 2, s)}{\Pr_h(ABBA)} = \frac{0.08 * .1 * .6 * 0.25610}{0.05544} = 0.02217.$$

The values are

$\gamma(ABBA, 1, s, s) = 0.28271$ $\quad \gamma(ABBA, 2, s, s) = 0.10071$ $\quad \gamma(ABBA, 3, s, s) = 0.04584$
$\gamma(ABBA, 1, s, t) = 0.53383$ $\quad \gamma(ABBA, 2, s, t) = 0.20417$ $\quad \gamma(ABBA, 3, s, t) = 0.13371$
$\gamma(ABBA, 1, t, s) = 0.02217$ $\quad \gamma(ABBA, 2, t, s) = 0.07884$ $\quad \gamma(ABBA, 3, t, s) = 0.06699$
$\gamma(ABBA, 1, t, t) = 0.16149$ $\quad \gamma(ABBA, 2, t, t) = 0.61648$ $\quad \gamma(ABBA, 3, t, t) = 0.75365$

$\gamma(BAB, 1, s, s) = 0.23185$ $\quad \gamma(BAB, 2, s, s) = 0.08286$
$\gamma(BAB, 1, s, t) = 0.65071$ $\quad \gamma(BAB, 2, s, t) = 0.16112$
$\gamma(BAB, 1, t, s) = 0.01212$ $\quad \gamma(BAB, 2, t, s) = 0.09199$
$\gamma(BAB, 1, t, t) = 0.13124$ $\quad \gamma(BAB, 2, t, t) = 0.68996$

6 $\delta(y, j, s)$

This is the probability of an analyzed word in $A(y)$ that the jth state is s. For $j <$ *length*(y), $\delta(y, j, s) = \sum_{u \in S} \gamma(y, j, s, u)$. Also, $\delta(y, n, s) = \alpha(y, n, s)/\Pr_h(y)$.

So here we have

$$\begin{aligned}
\delta(ABBA,1,s) &= 0.81654 & \delta(BAB,1,s) &= 0.88256\\
\delta(ABBA,1,t) &= 0.18366 & \delta(BAB,1,t) &= 0.14336\\
\delta(ABBA,2,s) &= 0.30488 & \delta(BAB,2,s) &= 0.24398\\
\delta(ABBA,2,t) &= 0.69532 & \delta(BAB,2,t) &= 0.78195\\
\delta(ABBA,3,s) &= 0.17955 & \delta(BAB,3,s) &= 0.14939\\
\delta(ABBA,3,t) &= 0.82064 & \delta(BAB,3,t) &= 0.85061\\
\delta(ABBA,4,s) &= 0.11273 & &\\
\delta(ABBA,4,t) &= 0.88727 & &
\end{aligned}$$

7 Our next HMM h_2

Recall that we start with a corpus c given by $c(ABBA) = 10$, $c(BBA) = 20$.

We want to use the δ values along with the corpus to get a new HMM, defined by relative frequency estimates of the expected analyzed corpus c^*.

The starting probability of state s is $I/(I + J)$, and that of t is $J/(I + J)$, where

$$\begin{aligned}
I &= \delta(ABBA,1,s)(c(ABBA)) + \delta(BAB,1,s)(c(BAB)) &= (0.81654 * 10) + (0.88256 * 20) &= 25.816600\\
J &= \delta(ABBA,1,t)(c(ABBA)) + \delta(BAB,1,t)(c(BAB)) &= (0.18366 * 10) + (0.14336 * 20) &= 4.703800
\end{aligned}$$

So we get that the start of s is 0.846, and the start of t is 0.154.

The probability of going from state s to state s will be $K/(K + L)$, where

$$\begin{aligned}
K &= (\gamma(ABBA,1,s,s) + \gamma(ABBA,2,s,s) + \gamma(ABBA,3,s,s)) * c(ABBA)\\
&\qquad +(\gamma(BAB,1,s,s) + \gamma(BAB,2,s,s)) * c(BAB)\\
&= (0.28271 + 0.10071 + 0.04584) * (10) + (0.23185 + 0.08286) * (20)\\
&= 10.58680\\
L &= (\gamma(ABBA,1,s,t) + \gamma(ABBA,2,s,t) + \gamma(ABBA,3,s,t)) * c(ABBA)\\
&\qquad +(\gamma(BAB,1,s,t) + \gamma(BAB,2,s,t)) * c(BAB)\\
&= (0.53383 + 0.20417 + 0.13371) * (10) + (0.65071 + 0.16112) * (20)\\
&= 24.95370
\end{aligned}$$

So the new value of go(s,s) is 0.298. Similarly, the new value of go(s,t) is 0.702. The probability of going from state t to state s will be $M/(M + N)$, where

$$\begin{aligned}
M &= (\gamma(ABBA,1,t,s) + \gamma(ABBA,2,t,s) + \gamma(ABBA,3,t,s)) * c(ABBA)\\
&\qquad +(\gamma(BAB,1,t,s) + \gamma(BAB,2,t,s)) * c(BAB)\\
&= (0.02217 + 0.07884 + 0.06699) * (10) + (0.01212 + 0.09199) * (20)\\
&= 3.76220\\
N &= (\gamma(ABBA,1,t,t) + \gamma(ABBA,2,t,t) + \gamma(ABBA,3,t,t)) * c(ABBA)\\
&\qquad +(\gamma(BAB,1,t,t) + \gamma(BAB,2,t,t)) * c(BAB)\\
&= (0.16149 + 0.61648 + 0.75365) * (10) + (0.13124 + 0.68996) * (20)\\
&= 31.74020
\end{aligned}$$

So the new value of go(t,s) is 0.106. Similarly, the new value of go(t,t) is 0.894.

Turning to the outputs, the probability that in state s we output A is $K/(K+L)$, where

$$\begin{aligned}
K &= (\delta(ABBA,1,s)+\delta(ABBA,4,s))*c(ABBA))+(\delta(BAB,2,s)*c(BAB)) \\
&= ((0.81654+0.11273)*10)+(0.24398*20) \\
&= 14.17230 \\
L &= (\delta(ABBA,2,s)+\delta(ABBA,3,s))*c(ABBA))+((\delta(BAB,1,s)+\delta(BAB,3,s))*c(BAB)) \\
&= ((0.30488+0.17955)*10)+(0.88256+0.14939)*20) \\
&= 25.48330
\end{aligned}$$

Thus the probability is 0.357. Similarly, the probability that we output B in state s is 0.643.

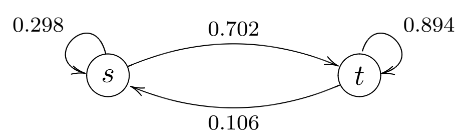
The probability that in state t we output A is $M/(M+N)$, where

$$\begin{aligned}
M &= (\delta(ABBA,1,t)+\delta(ABBA,4,t))*c(ABBA))+(\delta(BAB,2,t)*c(BAB)) \\
&= ((0.18366+0.88727)*10)+(0.78195*20) \\
&= 26.34830 \\
N &= (\delta(ABBA,2,s)+\delta(ABBA,3,s))*c(ABBA))+((\delta(BAB,1,s)+\delta(BAB,3,s))*c(BAB)) \\
&= ((0.69532+0.82064)*10)+(0.14336+0.85061)*20) \\
&= 35.03900
\end{aligned}$$

Thus the probability is 0.4292. Similarly, the probability that we output B in state s is $N/(M+N)$, 0.5708.

7.1 Another model

We have a new HMM which we call h_2:



Starting probability of s is 0.846, of t is 0.154.

In s, $\Pr(A) = 0.357$, $\Pr(B) = 0.643$. In t, $\Pr(A) = 0.4292$, $\Pr(B) = 0.5708$.

8 Again

At this point, we do all the calculations over again. I have hidden them, and only report the probabilities of the elements of Y and the log likelihood of the corpus.
Total probability of $ABBA$ is $0.00635 + 0.04690 = 0.05325$.
Total probability of BAB is $0.0223 + 0.1250 = 0.1473$.

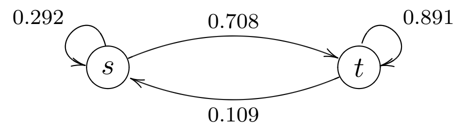
8.1 The likelihood of the corpus using h_2

$L(c, h_2) = \Pr(ABBA)^{c(ABBA)} \cdot \Pr(BAB)^{c(BAB)} = 0.05325^{10} 0.1473^{20}$

$\log L(c, h_2) = (10 * \log 0.05325) + (20 * \log 0.1473) = -67.6333$

8.2 Again a new model

Using h_2, we then do all the calculations and construct a new HMM which we call h_3:



Starting probability of s is 0.841 of t is 0.159.

In s, $\Pr(A) = 0.3624$, $\Pr(B) = 0.6376$. In t, $\Pr(A) = 0.4252$, $\Pr(B) = 0.5748$.

9 Again Again

Total probability of $ABBA$ is $0.00653 + 0.04672 = 0.05325$.
Total probability of BAB is $0.0223 + 0.1254 = 0.1477$.

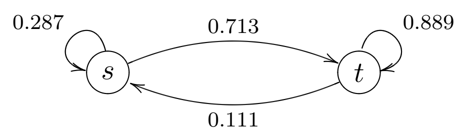
9.1 The likelihood of the corpus using h_3

$L(c, h_3) = \Pr(ABBA)^{c(ABBA)} \cdot \Pr(BAB)^{c(BAB)} = 0.05325^{10} 0.1477^{20}$

$\log L(c, h_3) = (10 * \log 0.05325) + (20 * \log 0.1477) = -67.5790$

9.2 Another model

After doing all the calculations once again, we have a new HMM which we call h_4:



Starting probability of s is 0.841, of t is 0.159. In s, $\Pr(A) = 0.3637$, $\Pr(B) = 0.6363$. In t, $\Pr(A) = 0.4243$, $\Pr(B) = 0.5757$.

10 The likelihoods

The likelihood of c in h_1 was -68.2611.

The likelihood of c in h_2 was -67.6333.

The likelihood of c in h_3 was -67.5790.

In playing around with different starting values, I found that the likelihood on h_3 sometimes was worse than that of h_2 (contrary to what we'll prove in class). I believe this is due to rounding errors in the calculations of the starting probabilities in the different states. I also noticed that most of the updating was actually to those starting probabilities, with the others changing only a little.