עיבוד שפה טבעית - תרגיל 2

שיר שבח 322407701, אליה חסון 208845032

8 בפברואר 2024

.1

כדי למצוא את סדרת המצבים האופטימלית (בעלת ההיסטברות הגדולה ביותר) נבנה את תרשים זרימה, שבו התגיות הם המצבים H או L, כאשר המעברים בין מצבי התגיות מוגדרים להיות מ־

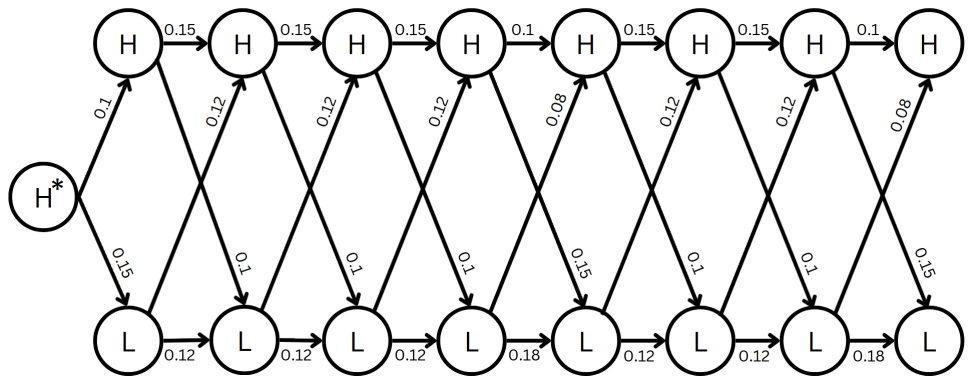
$$q\left(t \mid H\right) = \begin{cases} 0.5 & t = H \\ 0.5 & t = L \end{cases}$$

$$q\left(t \mid L\right) = \begin{cases} 0.4 & t = H \\ 0.6 & t = L \end{cases}$$

ונגדיר את ההסתברויות של ה־emission באופן הבא:

$$e\left(x \mid H\right) = \begin{cases} 0.2 & x = A \\ 0.3 & x = C \\ 0.3 & x = G \\ 0.2 & x = T \end{cases}$$

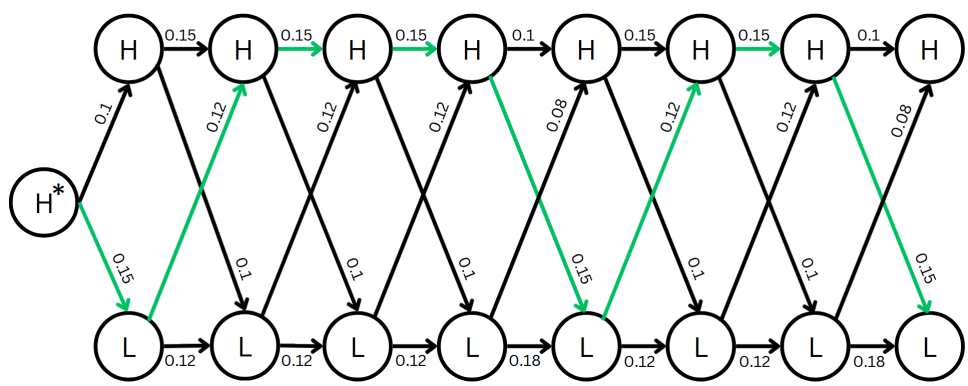
$$e\left(x \mid L\right) = \begin{cases} 0.3 & x = A \\ 0.2 & x = C \\ 0.2 & x = G \\ 0.3 & x = T \end{cases}$$



נשם לב שמדובר על תרשים זרימה עבור אלגוריתם viterbi מבוסס מודל ביגרם.

לפי אלגוריתם viterbi, המסלול האופטימלי ביותר עבור $S = ACCGTGCA$ הוא:

$$
\begin{aligned}
&P(L \mid H) \cdot P(A \mid L) + P(H \mid L) \cdot P(C \mid L) + P(H \mid H) \cdot P(C \mid H) \\
&+P(H \mid H) \cdot P(G \mid H) + P(L \mid H) \cdot P(T \mid L) + P(H \mid L) \cdot P(G \mid H) \\
&+P(H \mid H) \cdot P(C \mid H) + P(L \mid H) \cdot P(A \mid L) \\
&= 0.15 \cdot 0.12 \cdot 0.15 \cdot 0.15 \cdot 0.15 \cdot 0.15 \cdot 0.12 \cdot 0.15 \cdot 0.15
\end{aligned}
$$



לכן קיבלנו כי המסלול האופטימלי ביותר הוא $H^*LHHHLHHL$ כאשר H^* הוא המצב ההתחלתי.

.2

קלט: מספר טבעי n, פרמטרים $q(w \mid t, u, v)$ ו־ $e(x, s)$.

הגדרות: נגדיר $\mathcal{K}$ להיות קבוצה של תגיות אפשריות. נגדיר $\mathcal{K}_{-2} = \mathcal{K}_{-1} = \mathcal{K}_0 = \{*\}$, ו־ $\mathcal{K}_k = \mathcal{K}$ לכל $k = 1...n$. נגדיר $\mathcal{V}$ להיות קבוצה של מילים אפשריות.

אתחול: נגדיר $\pi\left(0,^*,^*,^*\right)=1$ ומילון d שישמור את כל ההסתברויות המקסימליות של כל תת קבוצה של $\mathcal{V}$ באורך n.

אלגוריתם:

לכל תת קבוצה $V\subset\mathcal{V}$ כך ש־ $|V|=n$ כך ש־ $V=x_1...x_n$:

לכל $k=1...n$:

לכל $t\in\mathcal{K}_{k-2},u\in\mathcal{K}_{k-1},v\in\mathcal{K}_k$:

$$\pi\left(k,t,u,v\right)=\max_{w\in\mathcal{K}_{k-3}}\left(\pi\left(k-1,w,t,u\right)\times q\left(v\mid w,t,u\right)\times e\left(x_k\mid v\right)\right)$$

$$d\left(V\right)=\max_{t\in\mathcal{K}_{n-2},u\in\mathcal{K}_{k-1},v\in\mathcal{K}_k}\left(\pi\left(n,t,u,v\right)\times q\left(\mathrm{STOP}\mid t,u,v\right)\right)$$

נאתחל את π מחדש ונגדיר $\pi\left(0,^*,^*,^*\right)=1$.

נחזיר: $\max\left(d\right)$

.3

להלן התוצאות השגיאה שקיבלנו בהרצת האלגוריתם:

```
B2
Known word error rate: 0.07044634806131655
Unknown word error rate: 0.743006993006993
Total error rate: 0.14726437699680506
C
Known word error rate: 0.03922452660054099
Unknown word error rate: 0.743006993006993
Total error rate: 0.1196086261980831
D
Known word error rate: 0.07044634806131655
Unknown word error rate: 0.743006993006993
Total error rate: 0.14726437699680506
E2
Known word error rate: 0.11018577834721333
Unknown word error rate: 0.7076923076923076
Total error rate: 0.14896166134185307
E3
Known word error rate: 0.21739617149893653
Unknown word error rate: 0.6971375807940905
Total error rate: 0.2692691693290735
```

להלן השגיאות הכי נפוצות שקיבלנו על פי מטריצת הבלבול:

```
for NNS, viterbi predicted NN, 316 times
for NP, viterbi predicted NN, 168 times
for JJ, viterbi predicted NN, 153 times
for VB, viterbi predicted NN, 85 times
for NN, viterbi predicted AT, 66 times
for IN, viterbi predicted AT, 59 times
for TO, viterbi predicted IN, 57 times
for NP, viterbi predicted AT, 56 times
for VBN, viterbi predicted NN, 50 times
for RB, viterbi predicted NN, 49 times
for JJ, viterbi predicted AT, 47 times
for VBD, viterbi predicted NN, 46 times
for NN, viterbi predicted NP, 45 times
for CD, viterbi predicted NN, 44 times
for VBG, viterbi predicted NN, 44 times
```