

עיבוד שפה טבעית - תרגיל 2

ערן תורגימן (208484147) ודור דהוקי (315145490)

30 בנובמבר 2022

1 שאלה 1

$$P(L|H) = 0.5$$

$$P(H|L) = 0.4$$

עבור $k = 0$ מתקיים:

$$\begin{bmatrix} H & 1 \\ L & 0 \end{bmatrix}$$

נחשב עבור $k = 1$:

$$\pi(1, H) = \max \{ \pi(0, H) \cdot q(H|H) \cdot e(A|H), \pi(0, L) \cdot q(H|L) \cdot e(A|H) \}$$

$$= \max \{ 1 \cdot 0.5 \cdot 0.2, 0 \cdot 0.4 \cdot 0.2 \}$$

$$= 0.1$$

$$\pi(1, L) = \max \{ \pi(0, H) \cdot q(L|H) \cdot e(A|L), \pi(0, L) \cdot q(L|L) \cdot e(A|L) \}$$

$$= \{ 1 \cdot 0.5 \cdot 0.3, 0 \cdot 0.6 \cdot 0.3 \}$$

$$= 0.15$$

Backpointer – H

$$\begin{bmatrix} H & 0.1 \\ L & 0.15 \end{bmatrix}$$

עבור $k = 2$:

$$\begin{aligned}
\pi(2, H) &= \max \{ \pi(1, H) \cdot q(H|H) \cdot e(C|H), \pi(1, L) \cdot q(H|L) \cdot e(C|H) \} \\
&= \max \{ 0.1 \cdot 0.5 \cdot 0.3, 0.15 \cdot 0.4 \cdot 0.3 \} \\
&= 0.018 \\
&\text{Backpointer} - L \\
\pi(2, L) &= \max \{ \pi(1, H) \cdot q(L|H) \cdot e(C|L), \pi(1, L) \cdot q(L|L) \cdot e(C|L) \} \\
&= \max \{ 0.1 \cdot 0.5 \cdot 0.2, 0.15 \cdot 0.6 \cdot 0.2 \} \\
&= 0.018 \\
&\begin{bmatrix} H & 0.018 \\ L & 0.018 \end{bmatrix}
\end{aligned}$$

עבור $k = 3$:

$$\begin{aligned}
\pi(3, H) &= \max \{ \pi(2, H) \cdot q(H|H) \cdot e(C|H), \pi(2, L) \cdot q(H|L) \cdot e(C|H) \} \\
&= \max \{ 0.018 \cdot 0.5 \cdot 0.3, 0.018 \cdot 0.4 \cdot 0.3 \} \\
&= 0.0027 \\
&\text{Backpointer} - H \\
\pi(3, L) &= \max \{ \pi(2, H) \cdot q(L|H) \cdot e(C|L), \pi(2, L) \cdot q(L|L) \cdot e(C|L) \} \\
&= \max \{ 0.018 \cdot 0.5 \cdot 0.2, 0.018 \cdot 0.6 \cdot 0.2 \} \\
&= 0.00216 \\
&\begin{bmatrix} H & 0.0027 \\ L & 0.00216 \end{bmatrix}
\end{aligned}$$

עבור $k = 4$:

$$\begin{aligned}
\pi(4, H) &= \max \{ \pi(3, H) \cdot q(H|H) \cdot e(G|H), \pi(3, L) \cdot q(H|L) \cdot e(G|H) \} \\
&= \max \{ 0.0027 \cdot 0.5 \cdot 0.3, 0.00216 \cdot 0.4 \cdot 0.3 \} \\
&= 0.000405 \\
&\text{Backpointer} - H \\
\pi(4, L) &= \max \{ \pi(3, H) \cdot q(L|H) \cdot e(G|L), \pi(3, L) \cdot q(L|L) \cdot e(G|L) \} \\
&= \max \{ 0.0027 \cdot 0.5 \cdot 0.2, 0.00216 \cdot 0.6 \cdot 0.2 \} \\
&= 0.00027 \\
&\begin{bmatrix} H & 0.000405 \\ L & 0.00027 \end{bmatrix}
\end{aligned}$$

עבור $k = 5$:

$$\begin{aligned}
\pi(5, H) &= \max \{ \pi(4, H) \cdot q(H|H) \cdot e(T|H), \pi(4, L) \cdot q(H|L) \cdot e(T|H) \} \\
&= \max \{ 0.000405 \cdot 0.5 \cdot 0.2, 0.00027 \cdot 0.4 \cdot 0.2 \} \\
&= 0.0000405 \\
\pi(5, L) &= \max \{ \pi(4, H) \cdot q(L|H) \cdot e(T|L), \pi(4, L) \cdot q(L|L) \cdot e(T|L) \} \\
&= \max \{ 0.000405 \cdot 0.5 \cdot 0.3, 0.00027 \cdot 0.6 \cdot 0.3 \} \\
&= 0.00006075 \\
&\text{Backpointer} - H \\
&\begin{bmatrix} H & 0.0000405 \\ L & 0.00006075 \end{bmatrix}
\end{aligned}$$

עבור $k = 6$:

$$\begin{aligned}
\pi(6, H) &= \max \{ \pi(5, H) \cdot q(H|H) \cdot e(G|H), \pi(5, L) \cdot q(H|L) \cdot e(G|H) \} \\
&= \max \{ 0.0000405 \cdot 0.5 \cdot 0.3, 0.00006075 \cdot 0.4 \cdot 0.3 \} \\
&= 0.00000729 \\
&\text{Backpointer} - L \\
\pi(6, L) &= \max \{ \pi(5, H) \cdot q(L|H) \cdot e(G|L), \pi(5, L) \cdot q(L|L) \cdot e(G|L) \} \\
&= \max \{ 0.0000405 \cdot 0.5 \cdot 0.2, 0.00006075 \cdot 0.6 \cdot 0.2 \} \\
&= 0.00000729 \\
&\begin{bmatrix} H & 0.00000729 \\ L & 0.00000729 \end{bmatrix}
\end{aligned}$$

עבור $k = 7$:

$$\begin{aligned}
\pi(7, H) &= \max \{ \pi(6, H) \cdot q(H|H) \cdot e(C|H), \pi(6, L) \cdot q(H|L) \cdot e(C|H) \} \\
&= \max \{ 0.00000729 \cdot 0.5 \cdot 0.3, 0.00000729 \cdot 0.4 \cdot 0.3 \} \\
&= 0.000001093 \\
&\text{Backpointer} - H \\
\pi(7, L) &= \max \{ \pi(6, H) \cdot q(L|H) \cdot e(C|L), \pi(6, L) \cdot q(L|L) \cdot e(C|L) \} \\
&= \max \{ 0.00000729 \cdot 0.5 \cdot 0.2, 0.00000729 \cdot 0.6 \cdot 0.2 \} \\
&= 0.000000875 \\
&\begin{bmatrix} H & 0.000001093 \\ L & 0.000000875 \end{bmatrix}
\end{aligned}$$

עבור $k = 8$:

$$\begin{aligned}\pi(8, H) &= \max \{ \pi(7, H) \cdot q(H|H) \cdot e(A|H), \pi(7, L) \cdot q(H|L) \cdot e(A|H) \} \\ &= \max \{ 0.000010935 \cdot 0.5 \cdot 0.2, 0.0000008748 \cdot 0.4 \cdot 0.2 \} \\ &= 0.000000109 \\ \pi(8, L) &= \max \{ \pi(7, H) \cdot q(L|H) \cdot e(A|L), \pi(7, L) \cdot q(L|L) \cdot e(A|L) \} \\ &= \max \{ 0.000010935 \cdot 0.5 \cdot 0.3, 0.0000008748 \cdot 0.6 \cdot 0.3 \} \\ &= 0.000000164 \\ \text{Backpointer} &= H \\ &\begin{bmatrix} H & 0.000000109 \\ L & 0.000000164 \end{bmatrix}\end{aligned}$$

רצף המצבים הצפוי הוא:

$H, L, H, H, H, L, H, H, L$

וההסתברות לקבל אותו היא:

0.000000164

2 שאלה 2

Init: $\text{Set}(0, *, *, *) = 1$
 Algorithm:
 for $k = 1 \dots n$
 for $t_{-2} \in S_{k-2}, t_{-1} \in S_{k-1}, t \in S_k$
 $\pi(k, t_{-2}, t_{-1}, t, w) = \max_{w \in V, t_{-3} \in S_{k-3}} (\pi(k-1, t_{-3}, t_{-2}, t_{-1}) \cdot q(t|t_{-3}, t_{-2}, t_{-1}) \cdot e(w|t))$
 return $\max_{\substack{t_{-2} \in S_{k-2} \\ t_{-1} \in S_{k-1} \\ t \in S_k}} (\pi(n, t_{-2}, t_{-1}, t) \cdot q(\text{stop}|t_{-2}, t_{-1}, t))$

```

B. MLE error rate:
total error rate: 0.14811123293132666
known words error rate: 0.0704399684933048
unknown words error rate: 0.75043630017452

C. Bigram HMM tagger error rate
total error rate: 0.23183494468254762
known words error rate: 0.1610217171148869
unknown words error rate: 0.7809773123909249

D. Bigram HMM tagger with add-one smoothing error rate
total error rate: 0.21249875411143226
known words error rate: 0.14549341735118715
unknown words error rate: 0.7321116928446771

E.ii Bigram HMM tagger with PseudoWords
total error rate: 0.21259842519685035
known words error rate: 0.1949737357091359
unknown words error rate: 0.7407407407407407

E.iii Bigram HMM tagger with PseudoWords with Add-One smoothing
total error rate: 0.22186783614073557
known words error rate: 0.2041404882068184
unknown words error rate: 0.7530864197530864

```

לאחר סקירת *Confusion Matrix* ראינו כי הרוב המוחלט של הערכים שאינם על האלכסון הראשי מאופסים, כלומר המודל חזה נכון את רוב התגיות. אבל שמנו לב לכך שהתבצע *over fitting* ל-*Train Set*. הקטגוריות שהיו הכי נפוצות וקיבלו בחישוב ההסתברויות את ההסתברויות הכי גדולות נחזו פעמים רבות למרות שהן לא היו הקטגוריות האמיתיות. הקטגוריות היו: *AT, NN, JJ, NP* שהופיעו: 8100, 14384, 4582, 7282 בהתאמה.