# Kneser-Ney平滑

xuen

2013 年 12 月 30 日

假定2元文法为 $\mathcal{R} = \{(w_r w_s)|0 \leqslant r, s \leqslant N\}$其中 $N$为字典中词的数量,2元语法中词频排序表中第 $k$个数为:

$$n_k = \| \{(w_r w_s)|f(w_r w_s) = k, \forall 0 \leqslant r, s \leqslant N\} \| \tag{1}$$

那么KN平滑的概率如下计算:

$$P_{KN}(w_i|w_{i-1}) = \frac{\mathcal{A} + \mathcal{B} \cdot \mathcal{C}}{\sum\limits_{w_s} f(w_{i-1} w_s)} \tag{2}$$

其中:

$$\mathcal{A} = f(w_{i-1} w_i) - D(f(w_{i-1} w_i)) \tag{3}$$

$$\begin{aligned} \mathcal{B} = D_1 \cdot & \| \{(w_{i-1} w_s)|f(w_{i-1} w_s) = 1\} \| \\ & + D_2 \cdot \| \{(w_{i-1} w_s)|f(w_{i-1} w_s) = 2\} \| \\ & + D_{3+} \cdot \| \{(w_{i-1} w_s)|f(w_{i-1} w_s) \geq 3\} \| \end{aligned} \tag{4}$$

$$\mathcal{C} = \frac{\| \{(w_r w_i)|f(w_r w_i) > 0\} \|}{\| \{(w_r w_s)|f(w_r w_s) > 0\} \|} \tag{5}$$

这里 $Y = {}^{n_1}/_{n_1+2n_2}$,$D_1 = 1 - 2Y \cdot {}^{n_2}/_{n_1}$,$D_2 = 2 - 3Y \cdot {}^{n_3}/_{n_2}$,
$D_{3+} = 3 - 4Y \cdot {}^{n_4}/_{n_3}$,另外:

$$D(\alpha) = \begin{cases} 0 & \alpha = 0, \\ D_1 & \alpha = 1, \\ D_2 & \alpha = 2, \\ D_{3+} & \alpha \geq 3. \end{cases} \tag{6}$$