Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Part-of-Speech Tagging

1, Learn (by counting) the parameters of a HMM model
+ Implement a function count transitions to count tag-to-tag transitions.

+ Implement a function count emissions to count tag-to-word emissions C(wi, ti) and total tag counts C(ti) for each combination of word and POS-tag that occurs in the training corpus.

+ Describe an appropriate way to compute the initial state probabilities and implement it.
At initial state, we need to calculate emission probability $P(w_i|t_i)$ and transition probability $P(t_i|t_{i-1})$. The emission probability of initial state could be calculated as the same as other state. To calculate transition probability, there are two ways:
- 1: We could assume that the probability of tag t_i is equal to the probability of unigram: $P(t_i|t_{i-1}) = P(t_i)$
- 2: We could assume that the probability of tag t_i depends on the beginning of the sentences: $P(t_i|t_{i-1}) = P(t_i|"."")$

The current implementation follows the second approach.

+ Implement a function get transition probability to compute Maximum Likelihood Estimates for the transition probabilities.

2, OOV

+ We have:

$$\begin{aligned}\frac{C(w,t_i)+\epsilon}{C(t_i)+\epsilon V} &= \frac{C(w_i,t_i)}{C(t_i)+\epsilon V} + \frac{\epsilon}{C(t_i)+\epsilon V}\\ &= \frac{C(w_i,t_i)}{C(t_i)+\epsilon V} * \frac{C(t_i)}{C(t_i)} + \frac{\epsilon V}{(C(t_i)+\epsilon V)V}\\ &= \frac{C(t_i)}{C(t_i)+\epsilon V} * \frac{C(w_i,t_i)}{C(t_i)} + \frac{C(t_i)-C(t_i)+\epsilon V}{C(t_i)+\epsilon V} * \frac{1}{V}\\ &= \frac{C(t_i)}{C(t_i)+\epsilon V} * \frac{C(w_i,t_i)}{C(t_i)} + \left(\frac{C(t_i)+\epsilon V}{C(t_i)+\epsilon V} - \frac{C(t_i)}{C(t_i)+\epsilon V}\right) * \frac{1}{V}\\ &= \frac{C(t_i)}{C(t_i)+\epsilon V} * \frac{C(w_i,t_i)}{C(t_i)} + \left(1 - \frac{C(t_i)}{C(t_i)+\epsilon V}\right) * \frac{1}{V}\end{aligned}$$

With $\mu = \frac{C(t_i)}{C(t_i)+\epsilon V}(clearly, 0 \leq \mu \leq 1)$, then we have:

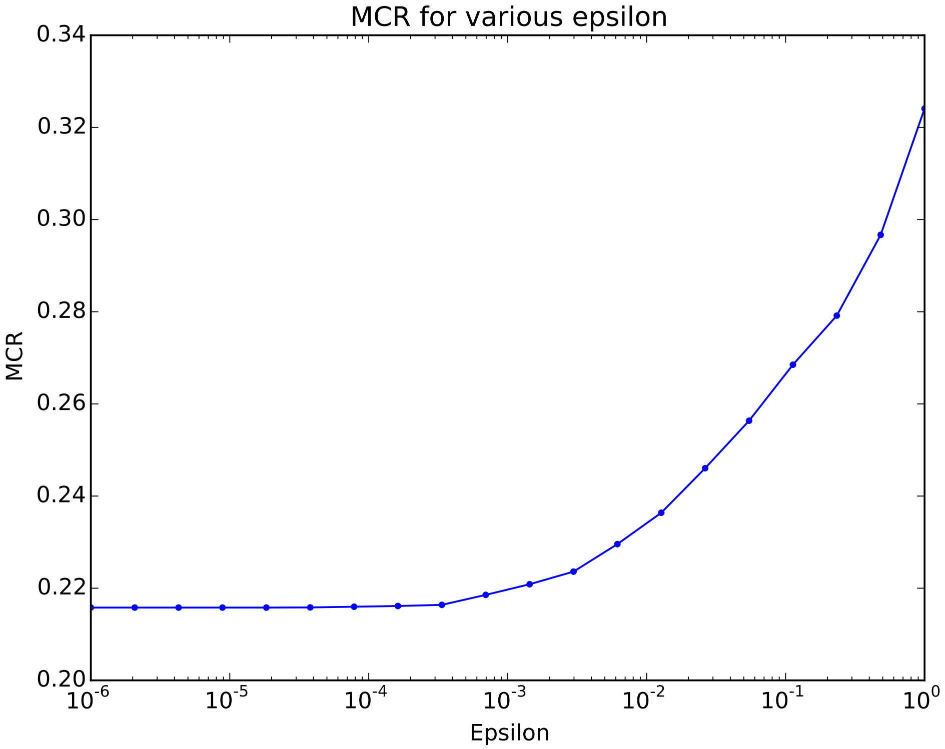
$$\frac{C(w_i,t_i)+\epsilon}{C(t_i)+\epsilon V} = \mu * \frac{C(w_i,t_i)}{C(t_i)} + (1-\mu) * \frac{1}{V}$$

This smoothing function will scale down the actual probability of the bigram (the first fraction) by a parameter μ, and then add into it a constant equals to the probability of zerogram (the second fraction), scaled up by $(1-\mu)$.

+ We got MCR for some epsilon values (run on brown-test set):
- $\epsilon = 1$: MCR = 0.3197
- $\epsilon = 10^{-4}$: MCR = 0.2153
- $\epsilon = 10^{-5}$: MCR = 0.2148

+ To calculate the suitable values for epsilon, we could try many alternative values of epsilon in range (0,1) such as they are spaced evenly in log scale, then choose the highest values of epsilon such as if we decrease its value, the MCR does not change significantly.

+ We got the following Misclassification Rate graph (MCR) (run on brown-development set):



Based on this graph, we conclude that the optimal values for epsilon is $\epsilon = 10^{-4}$

Ho Vinh Thinh (2562630)
Prabal Agarwal (2561419)
Nurzat Rakhmanberdieva (2561148)

Bonus

The mean of Poisson is equal to λ, which is the average number of times a word w_i appear on a single document.

$$\lambda = \frac{cf}{N}$$

- cf is the collection frequency of word w_i.
- N is the number of documents in the collection.

The variance of Poisson is also equal to λ, then we have:

$$\sigma^2 = \lambda$$

and standard deviation:

$$\sigma = \sqrt{\lambda}$$