AIMS | African Institute for Mathematical Sciences GHANA

Topic:

NLP

By:

NLP-Group-Members

Book: Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognitio

January 17, 2020

Contents

1 Exercise 3.1

1.1 Problem

Write out the equation for trigram probability estimation (modifying Eq. 3.11). Now write out all the non-zero trigram probabilities for the I am Sam corpus on page 33.

1.2 Trying Answer

In the case of 3-gram, one has the equation (1):

$$\mathbb{P}\left(w_1|w_2^{i-1}\right) = \frac{C\left(w_{i-2}w_{1-1}w_i\right)}{C\left(w_{i-2}^{(i-1)}\right)} \tag{1}$$

Let us compute all the three grams

$$\mathbb{P}\left(am|<\!/s\!>I\right) = \frac{C(<\!/s\!>Iam)}{C(<\!/s\!>I)} = \frac{1}{2}.$$

$$\mathbb{P}\left(Sam|I \quad am\right) = \frac{1}{2}.$$

$$\mathbb{P}\left(<\!/s\!>|am \quad Sam\right) = 1.$$

$$\mathbb{P}\left(I|<\!/s\!>Sam\right) = 1.$$

$$\mathbb{P}(<\!/s\!>|I \quad am) = \frac{1}{2}.$$

$$\mathbb{P}(do|<\!/s\!> \quad I) = \frac{1}{2}.$$

$$\mathbb{P}(not|I \quad do) = \frac{1}{1}.$$

$$\mathbb{P}(like|do \quad not) = \frac{1}{1}.$$

$$\mathbb{P}(green|not \quad like) = \frac{1}{1}.$$

$$\mathbb{P}(eggs|like \quad green) = \frac{1}{1}.$$

$$\mathbb{P}(and|green \quad eggs) = \frac{1}{1}.$$

$$\mathbb{P}(ham|eggs \quad and) = \frac{1}{1}.$$

$$\mathbb{P}(< /s > |and \quad ham) = \frac{1}{1}.$$

The aim here was to compute all the non zero 3-grams.

2 Exerise 3.2

2.1 Statement

Calculate the probability of the sentence i want chinese food. Give two probabilities, one using Fig 3.2 and the useful probabilities just below it on page 35, and another using the add-1 smoothed table in Fig. 3.6. Assume the additional add-1 smoothed probabilities $P(i| < s >) = 0.19$ and $P(< /s > |food) = 0.40$.

2.2 Solution

$$\begin{aligned}\mathbb{P}\left(< /s > IwantChinesefood < /s >\right) =& \mathbb{P}(I| < /s >)\mathbb{P}(want|I)\mathbb{P}(Chinese|Want) \\ & \times \mathbb{P}(food|chinese)\mathbb{P}(< /s > |food) \\ =& \, 0.00008477\end{aligned}$$

In the case of Laplace that we are using Laplace Smoothing, it becomes:

$$\mathbb{P}_{Laplace}\left(< /s > IwantChinesefood < /s >\right) = 0.0000024$$

3 Exercise 3.3

3.1 Statement

Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why.

3.2 Answer

$\mathbb{P}_{Laplace} < \mathbb{P}$ since $N + V >> N$.