ICA 9 N-Grams Revisit

(November 11th)

The goal of this ICA is for you to refamiliarize with N-Grams.

Question 1 (1.5 points): Consider the numeric expressions in the following sentence from the MedLine Corpus:

The corresponding free cortisol fractions in these sera were 4.53 +/- 0.15% and 8.16 +/- 0.20%, respectively.

Should we say that the numeric expression *8.16 +/- 0.20%* is three words?

Or should we say that it's a single compound word? Or should we say that it is actually *nine* words, since it's read "eight point one six, plus or minus zero point twenty percent"?

Or should we say that it's not a "real" word at all, since it wouldn't appear in any dictionary?

Discuss these different possibilities. Can you think of application domains that motivate at least two of these answers?

Question 2 (3 points) Recall that *<s>* and </s> are symbols that denote start of sentence and end of sentence respectively. We ignore punctuation here. Consider these expressions:

(a) The probability that someone asks `why do you' as a complete question.
(b) The probability that you hear `why do you' as a snippet of a conversation as you walk across the campus.
(c) The probability that you hear a question beginning `why do you'.

Match them with the probabilities below, and briefly explain.
1. P(why)P(do|why)P(you|do)

2. P(why|<s>)P(do|why)P(you|do)
3. P(why|<s>)P(do|why)P(you|do)P(</s>|you)

Consider the following corpus (adapted from Jurafsky and Martin)

<s> I am Sam </s>

<s> Sam I am </s>

<s> I do not like green eggs and ham </s>

Recall that the formula to estimate unigrams (also known as Maximum Likelihood Estimate) is given as below:

Formula for the MLE of Unigrams

The unsmoothed maximum likelihood estimate of the unigram probability of the word w_i is its count c_i normalized by the total number of word tokens N :

$$P(w_i) = \frac{c_i}{N}$$

Question 3 (0.5 point): According to MLE, find the value of P(I) for the above 3 sentence corpus.

Recall that the formula for MLE of bigrams is given as below:

Formula for the MLE of bigrams

$$P(w_i|w_{i-1}) = \frac{count(w_{i-1},w_i)}{count(w_{i-1})}$$

Question 4 (1 point): According to MLE, find the value of P(I|<s>) for the above 3 sentence corpus. Find the value of P(am|I).

Recall that MLE is problematic because it does not take into account the sparsity of training data. One technique to overcome this problem is smoothing. We discussed Add-one smoothing in class. Add-one smoothing is also known as Laplace smoothing.

Formula of Laplace for unigrams

$$P_{Laplace}(w_i) = \frac{c_i+1}{N+V}$$

Where V is size of vocabulary.

Question 5 (3 points): You come upon a new planet where the Blorks live. Blorks have only 2 words in their vocabulary, <<Ga>> and <<Bu>>. One day, you see the following corpus:

Ga Ga Ga

Given this. Can you associate each number to what it represents in the Laplace formula ?

1. $P_{Laplace}(Ga)$	a. $\frac{4}{5}$
2. $P_{Laplace}(Bu)$	b. 2
3. V	c. $\frac{1}{5}$
4. N	d. 3
5. c_{Ga}	e. 3
6. c_{Bu}	f. 0

Now you see another corpus:

Bu Bu Bu Ga

Formula of Laplace for Smoothing for bigrams

$$P_{Laplace}(w_i|w_{i-1}) = \frac{C(w_{i-1},w_i)+1}{C_{w_{i-1}}+V}$$

Question 6 (1 point): What is $P_{Laplace}$(Ga|Bu)?