Report Assignment 3

Arushi Dogra
201302084

Laplace Smoothing (add one smoothing)

$$p(w_i|w_{i-1}) = (1 + c(w_{i-1}w_i)) / (|V| + c(w_{i-1}))$$

In Laplace Smoothing we simply add one in the numerator and length of the distinct unigrams in denominator to account for unseen words and phrases.

When calculating linear interpolation, we calculate Laplace smoothing for trigrams, bigrams and unigrams and assign them weights 0.5, 0.3 and 0.2 respectively.

Good Turing Smoothing

$$r^* = (r + 1)(n_{r+1} / n_r)$$

In Good Turing Smoothing, we relocate the mass of n grams that occur r+1 times in the training data to the n grams that occur r times. Hence we calculate the new r (or r^*) as shown in above equation.

Now if n_{r+1} is missing, we find the next one that exists and use it's value rather than applying a complex form fitting function, as specified. If the token doesn't exist in the dictionary, we take the frequency of bin with r=1. If r is the maximum possible, we take $n_{r+1} = n_r$ so that r^* becomes r+1.

Similar to Laplace, when calculating linear interpolation, we calculate Good Turing smoothing for trigrams, bigrams and unigrams and assign them weights 0.5, 0.3 and 0.2 respectively.

Pipeline

First I have run my tokenizer to generate the unigram, bigram and trigram dictionaries, along with their lengths. Then for a dataset, I read a line, tokenize it using a simple split, and then apply smoothing and calculate and print probability of that line occurring given our training data. This is done for both Hindi as well as English, applying all the four smoothings on each line of the test data.

Likelihood of sentences given in ToyTestData.txt

1. English

a. Laplace Smoothing

line 1 : 7.04255961319e 41
line 2 : 6.43265057625e 51
line 3 : 3.63526330496e 115
line 4 : 9.5628591494e 126
line 5 : 9.48174971683e 86
line 6 : 8.52419371581e 161
line 7 : 1.97889534635e 158
line 8 : 1.56588866745e 44
line 9 : 5.51473727396e 138
line 10 : 1.34714912093e 45
line 11 : 1.34720069225e 45

b. Laplace Smoothing with interpolation

line 1 : 4.64143778148e 41
line 2 : 2.01210058492e 44
line 3 : 4.46573665755e 78
line 4 : 1.05882298789e 66
line 5 : 2.35319887935e 47
line 6 : 5.77435095464e 87
line 7 : 1.25622109982e 77
line 8 : 5.00813385761e 25
line 9 : 1.72507833421e 68
line 10 : 5.7013843976e 30
line 11 : 2.96315950188e 33

c. Good Turing Smoothing

line 1 : 0.381977652089
line 2 : 0.300294545371
line 3 : 4.29476270845e 07
line 4 : 1.09622081875e 28
line 5 : 1.35547663478e 08

line 6 : 2.21617841554e 36
line 7 : 2.28863871417e 61
line 8 : 9.25457536999e 13
line 9 : 6.751246388e 32
line 10 : 3.54857575402e 08
line 11 : 3.54857575402e 08

d. Good Turing Smoothing with Interpolation

line 1 : 0.190988837161
line 2 : 0.150147272687
line 3 : 2.14738135422e 07
line 4 : 5.48110409377e 29
line 5 : 6.7773831739e 09
line 6 : 1.10808920777e 36
line 7 : 5.49560549891e 48
line 8 : 4.6272876867e 13
line 9 : 3.375623194e 32
line 10 : 1.77428787701e 08
line 11 : 1.77428787701e 08

2. Hindi

a. Laplace Smoothing

line 1 : 1.61328659223e 221
line 2 : 8.45677233892e 67
line 3 : 1.64168779638e 121
line 4 : 2.72731355739e 228
line 5 : 2.26018327784e 98
line 6 : 4.59016426242e 112
line 7 : 6.99687239052e 72
line 8 : 9.16934756543e 121
line 9 : 7.06582578297e 105
line 10 : 3.68957085121e 107

b. Laplace Smoothing with Interpolation

line 1 : 3.66530502438e 114
line 2 : 1.00316508212e 47
line 3 : 3.85361858025e 67
line 4 : 3.35635598125e 126
line 5 : 9.99691571356e 59
line 6 : 1.46274501078e 68
line 7 : 1.4567434039e 47
line 8 : 1.00543854706e 75
line 9 : 2.63473480287e 58
line 10 : 1.78651417028e 64

c. Good Turing Smoothing

line 1 : 9.08117636957e 91
line 2 : 0.163927562298
line 3 : 2.63966531851e 35
line 4 : 6.83460637512e 72
line 5 : 8.66465343105e 25
line 6 : 2.4065346224e 08
line 7 : 0.142639758412
line 8 : 8.04101322252e 21
line 9 : 1.02273947485e 48
line 10 : 6.7264173733e 09

d. Good Turing Smoothing with Interpolation

line 1 : 1.01595512593e 84
line 2 : 0.0819637811488
line 3 : 1.31983265925e 35
line 4 : 3.41730318756e 72
line 5 : 4.33232671553e 25
line 6 : 1.2032673112e 08
line 7 : 0.0713198792059
line 8 : 4.02050661126e 21
line 9 : 2.44387434201e 42
line 10 : 3.36320868665e 09