HW 1 - Brief Report

Elvis Fernandez

February 2017

1 Perplexity

As suspected the value for perplexity of the Interpolated Bigram is lower than both the Laplace smoothing for unigrams and bigrams. The naive approach of simply pretending to seeing every word or every bigram once more than seen in training is what is used in Laplace. Interpolation does something different, it takes advantages of different language models to implement its own smoothing. For example, suppose you have a unigram, a bigram, and a trigram on a training corpus. Interpolation allows you to take advantage of all three language models by combining them. In my assignment $Pr_{Int}(y|x) = \lambda Pr_{MLE}(y|x) + (1 - \lambda)Pr_L(y)$. Here interpolation uses both MLE probability and the Laplace smoothing for y, by combining these probabilities and lambda from dev.txt. In my calculations Laplace Unigram was smaller than Laplace Bigram, because perhaps the training data was not large enough and unigrams were actually able to perform better than high order n-grams. I was expecting the opposite, since the bigram model usually suggests that there is more context of a word, and in turn should be a better model. And in the case of unigram models, which takes zero to no care for its context can actually create nonsensical sentences.

Perplexities Calculated:

- Laplace Bigram: 1214.68641571
- Interpolated Bigram: 401.104384608
- Laplace Unigram: 515.72961315

2 Top 20 Bigrams for Laplace.

1. of the 0.00106450915247
2. in the 0.000824927743286
3. the fly 0.000381751185344

4. the child 0.000269471424949
5. the body 0.000269471424949
6. the most 0.000202103568712
7. the house 0.000202103568712
8. of insects 0.000189570123043
9. and the 0.000186279884599
10. to the 0.000183483066101
11. the wings 0.000179647616633
12. the home 0.000179647616633
13. of a 0.000174987805886
14. of these 0.000160405488728
15. the rural 0.000157191664554
16. the mouth 0.000157191664554
17. the mosquito 0.000134735712474
18. the maggots 0.000134735712474
19. the germs 0.000134735712474
20. the field 0.000134735712474