CSCI 780 NLP Fall 2016: Homework 1 – Language Modeling

Nicolas Stoian

Due Date: October 6, 2016

1. How many word types (unique words) are there in the training corpus? Please include the padding symbols and the unknown token.

The number of word types in training corpus is 15031

2. How many word tokens are there in the training corpus?

The number of word tokens in training corpus is 498474

3. What percentage of word tokens and word types in each of the test corpora did not occur in training (before you mapped the unknown words to <unk> in training and test data)?

For brown-test.txt:
Total number of words in file = 18518
Number of words not appearing in training data = 1110
Percentage of words not appearing in training data = 5.994167836699427 %

For learner-test.txt:
Total number of words in file = 9170
Number of words not appearing in training data = 463
Percentage of words not appearing in training data = 5.04907306434024 %

4. What percentage of bigrams (bigram types and bigram tokens) in each of the test corpora that did not occur in training (treat <unk> as a token that has been observed).

For brown-test.txt
Total number of bigrams in file = 17694
Number of bigrams not appearing in training data = 4682
Percentage of bigrams not appearing in training data = 26.460947213744774 %

For learner-test.txt
Total number of bigrams in file = 8670
Number of bigrams not appearing in training data = 2310
Percentage of bigrams not appearing in training data = 26.643598615916954 %

5. Compute the log probabilities of the following sentences under the three models (ignore capitalization and pad each sentence as described above). Please list all of the parameters required to compute the probabilities and show the complete calculation. Which of the parameters have zero values under each model?

- **He was laughed off the screen .**
- **There was no compulsion behind them .**
- **I look forward to hearing your reply .**

6. Compute the perplexities of each of the sentences above under each of the models.

Sentence = <s> He was laughed off the screen . </s>

p(he) = 5957/498474
p(was) = 5149/498474
p(laughed) = 43/498474
p(off) = 402/498474
p(the) = 24657/498474
p(screen) = 15/498474
p(.) = 22238/498474
p(</s>) = 26000/498474

The unigram log probability = -64.86594292562941
The unigram perplexity = 147.78202494498612

p(he|<s>) = 2133/26000
p(was|he) = 692/5957
p(laughed|was) = 0/5149
p(off|laughed) = 0/43
p(the|off) = 75/402
p(screen|the) = 3/24657
p(.|screen) = 0/15
p(</s>|.) = 22238/22238

The bigram log probability = 0
The bigram perplexity = Undefined

p(he|<s>) = 2134/41031
p(was|he) = 693/20988
p(laughed|was) = 1/15074
p(off|laughed) = 1/15433
p(the|off) = 76/15433
p(screen|the) = 4/39688
p(.|screen) = 1/37269
p(</s>|.) = 22239/37269

The bigram with plus one smoothing log probability = -72.93025620609369
The bigram with plus one smoothing perplexity = 275.0141044324117

Sentence = <s> There was no compulsion behind them . </s>

p(there) = 1243/498474
p(was) = 5149/498474
p(no) = 998/498474
p(<unk>) = 13219/498474
p(behind) = 166/498474
p(them) = 812/498474
p(.) = 22238/498474
p(</s>) = 26000/498474

The unigram log probability = -59.00702251167198
The unigram perplexity = 94.11389507774473

p(there|<s>) = 379/26000
p(was|there) = 381/1243
p(no|was) = 120/5149
p(<unk>|no) = 27/998
p(behind|<unk>) = 5/13219
p(them|behind) = 10/166
p(.|them) = 137/812
p(</s>|.) = 22238/22238

The bigram log probability = -36.42614031888745
The bigram perplexity = 16.533828523548188

p(there|<s>) = 380/41031
p(was|there) = 382/16274
p(no|was) = 121/20180
p(<unk>|no) = 28/16029
p(behind|<unk>) = 6/28250
p(them|behind) = 11/15197
p(.|them) = 138/15843
p(</s>|.) = 22239/37269

The bigram with plus one smoothing log probability = -58.93122645395324
The bigram with plus one smoothing perplexity = 93.56610226591656

Sentence = <s> I look forward to hearing your reply . </s>

p(i) = 3235/498474
p(look) = 231/498474
p(forward) = 47/498474
p(to) = 9789/498474
p(hearing) = 30/498474
p(your) = 367/498474

p(reply) = 29/498474
p(.) = 22238/498474
p(</s>) = 26000/498474

The unigram log probability = -84.63012364878018
The unigram perplexity = 352.8747435254266

p(i|<s>) = 916/26000
p(look|i) = 1/3235
p(forward|look) = 4/231
p(to|forward) = 13/47
p(hearing|to) = 0/9789
p(your|hearing) = 0/30
p(reply|your) = 1/367
p(.|reply) = 6/29
p(</s>|.) = 22238/22238

The bigram log probability = 0
The bigram perplexity = Undefined

p(i|<s>) = 917/41031
p(look|i) = 2/18266
p(forward|look) = 5/15262
p(to|forward) = 14/15078
p(hearing|to) = 1/15061
p(your|hearing) = 1/15398
p(reply|your) = 2/15398
p(.|reply) = 7/15060
p(</s>|.) = 22239/37269

The bigram with plus one smoothing log probability = -93.4932169533023
The bigram with plus one smoothing perplexity = 652.268295379401

7. Compute the perplexities of the entire test corpora, separately for the brown-test.txt and learner-test.txt under each of the models. Discuss the differences in the results you obtained.

Test corpus name = brown-test-padded.txt

The total unigram perplexity = 320.1785803202881

The total bigram perplexity = 21.70397754661539
728 of 824 sentences had zero probability and were discarded

The total bigram with plus one smoothing perplexity = 668.709516762931

Test corpus name = learner-test-padded.txt

The total unigram perplexity = 348.7097569950318

The total bigram perplexity = 35.03097322384013
464 of 500 sentences had zero probability and were discarded

The total bigram with plus one smoothing perplexity = 845.4916726844544

For the bigram perplexity I completely discarded any sentences in the testing data that had zero probability. I wanted to see how the model performed on those sentences that had some probability. The bigram model far outperformed the other two models, with the smoothed bigram model performing the worst. I believe that this shows that plus one smoothing is a very poor smoothing technique as it is out performed by the unigram model.