Project 1 Write Up

Moshe Weiss

**Preprocessing**

Several preprocessing steps were imposed upon the training and two testing sets of the data. In all three, each sentence was padded with a <s> prior to the first token in the sentence and a </s> prior to the sentence terminating symbol. The possible sentence terminating symbols in the corpus were ., ! !, and ? ?, as per the documents’ pre-imposed format. Line break characters were removed from the document, and casing was forced to lower. In the training set, tokens only present once in the document were mapped to <unk>. In the testing sets, tokens that did not appear in the processed training set were mapped to <unk>.

**Word Types in Training**

Word types in the Brown Train corpus were calculated after preprocessing by splitting the document on spaces. Note that while ! ! and ? ? were considered terminating symbols, they were considered two instances of the ! and ? tokens respectively because of the separating space. By this definition, there were 15032 token types in the training corpus. To accommodate for reasonable formatting error, let us say there were approximately 15000 token types.

**Word Tokens in Training**

Following the previous constraints, ! ! and ? ? were each considered two tokens, because of the space between them. There were a total of 990,189 tokens in the training corpus, which can be approximated to 990,000 to accommodate for error.

**Test Exclusive Tokens**

Prior to the <unk> mapping in both train and test, the percent of words exclusive to the testing corpora were calculated.

22.76% of the word types in the Brown Test corpus did not occur in the Brown Train corpus.

6.02% of the word tokens in the Brown Test corpus did not occur in the Brown Train corpus.

16.35% of the word types in the Learner Test corpus did not occur in the Brown Train corpus.

5.18% of the word tokens in the Learner Test corpus did not occur in the Brown Train corpus.

**Test Exclusive Bigrams**

The percent of bigrams exclusive to the testing corpora were calculated after the <unk> mapping in both train and test.

100% of the bigram types in the Brown Test corpus did not occur in the Brown Train corpus.

99.00% of the bigram tokens in the Brown Test corpus did not occur in the Brown Train corpus.

100% of the bigram types in the Learner Test corpus did not occur in the Brown Train corpus.

98.99% of the bigram tokens in the Learner Test corpus did not occur in the Brown Train corpus.

This cannot be right. During debugging, I discovered my bigram model was creating an empty token and logging all the conditional probabilities in the condition on that empty token. I do not know why.

**Log and Zero Probabilities**

For the sample sentences, the log and zero probabilities were recorded for the unigram MLE, bigram MLE, and bigram Smoothed models. For our intents and purposes, log refers to log base 2.

Unigram Zero Probabilities: None

Sentence 1

log2(P(<s>)) = -5.348082124024594

log2(P(he)) = -7.376974153197119

log2(P(was)) = -7.587267842143228

log2(P(laughed)) = -14.491079642238587

log2(P(off)) = -11.266292705761757

log2(P(the)) = -5.327634738672769

log2(P(screen)) = -16.010453801332165

log2(P(</s>)) = -5.348082124024594

log2(P(.)) = -5.476604973926412

Log probability of sentence 1 = -78.23247210532124

Sentence 2

log2(P(<s>)) = -5.348082124024594

log2(P(there)) = -9.6377338158882

log2(P(was)) = -7.587267842143228

log2(P(no)) = -9.954448391603425

log2(P(<unk>)) = -6.227018974404066

log2(P(behind)) = -12.542304965593761

log2(P(them)) = -10.25200847975551

log2(P(</s>)) = -5.348082124024594

log2(P(.)) = -5.476604973926412

Log probability of sentence 2 = -72.37355169136379

Sentence 3

log2(P(<s>)) = -5.348082124024594

log2(P(i)) = -8.257794400060815

log2(P(look)) = -12.065595355524628

log2(P(forward)) = -14.362755545263049

log2(P(to)) = -6.660398624127665

log2(P(hearing)) = -15.010453801332167

log2(P(your)) = -11.397708144097473

log2(P(reply)) = -15.059363401813114

log2(P(</s>)) = -5.348082124024594

log2(P(.)) = -5.476604973926412

Log probability of sentence 3 = -98.98683849419453

Bigram MLE Model

Bigram MLE Zero Probabilities: All except sentence 3

Sentence 1

log2(P(he | <s>)) = -inf

log2(P(was | he)) = -inf

log2(P(laughed | was)) = -inf

log2(P(off | laughed)) = -inf

log2(P(the | off)) = -inf

log2(P(screen | the)) = -inf

log2(P(</s> | screen)) = -inf

log2(P(. | </s>)) = -inf

Log probability of sentence 1 = -inf

Sentence 2

log2(P(there | <s>)) = -inf

log2(P(was | there)) = -inf

log2(P(no | was)) = -inf

log2(P(<unk> | no)) = -inf

log2(P(behind | <unk>)) = -inf

log2(P(them | behind)) = -inf

log2(P(</s> | them)) = -inf

log2(P(. | </s>)) = -inf

Log probability of sentence 2 = -inf

Sentence 3

log2(P(<s>)) = -5.348082124024594

log2(P(i)) = -8.257794400060815

log2(P(look)) = -12.065595355524628

log2(P(forward)) = -14.362755545263049

log2(P(to)) = -6.660398624127665

log2(P(hearing)) = -15.010453801332167

log2(P(your)) = -11.397708144097473

log2(P(reply)) = -15.059363401813114

log2(P(</s>)) = -5.348082124024594

log2(P(.)) = -5.476604973926412

Log probability of sentence 3 = -98.98683849419453

Bigram Smooth Model

Bigram Smooth Zero Probabilities: None

Sentence 1

log2(P(he | <s>)) = -15.263782680245184

log2(P(was | he)) = -14.357345812048587

log2(P(laughed | was)) = -14.300710043552511

log2(P(off | laughed)) = -13.87987038167481

log2(P(the | off)) = -13.913824390408063

log2(P(screen | the)) = -15.276451592347872

log2(P(</s> | screen)) = -13.877188257461297

log2(P(. | </s>)) = -15.263782680245184

Log probability of sentence 1 = -116.13295583798352

Sentence 2

log2(P(there | <s>)) = -15.263782680245184

log2(P(was | there)) = -13.99036992294036

log2(P(no | was)) = -14.300710043552511

log2(P(<unk> | no)) = -13.96848680504191

log2(P(behind | <unk>)) = -14.786014315024428

log2(P(them | behind)) = -13.891593862433657

log2(P(</s> | them)) = -13.951648985904619

log2(P(. | </s>)) = -15.263782680245184

Log probability of sentence 2 = -115.41638929538786

Sentence 3

log2(P(i | <s>)) = -15.263782680245184

log2(P(look | i)) = -14.156952098325734

log2(P(forward | look)) = -13.897750936724783

log2(P(to | forward)) = -13.880253135551403

log2(P(hearing | to)) = -14.599273620157868

log2(P(your | hearing)) = -13.878625729805357

log2(P(reply | your)) = -13.910549045931726

log2(P(</s> | reply)) = -13.878529942862396

log2(P(. | </s>)) = -15.263782680245184

Log probability of sentence 3 = -128.72949986984966

**Perplexities of Unseen Sentences**

The perplexities of the three new sentences were computed for each model. The lower the perplexity, the better fit a model is to the sentence.

Unigram MLE Perplexity

Sentence 1 Perplexity: 0.0024171171272496864

Sentence 2 Perplexity: 0.0037954699813357927

Sentence 3 Perplexity: 0.0010476090553378189

Bigram MLE Perplexity

Sentence 1 Perplexity: 0.0

Sentence 2 Perplexity: 0.0

Sentence 3 Perplexity: 0.0

Bigram Smooth Perplexity

Sentence 1 Perplexity: 0.00013050007353855155

Sentence 2 Perplexity: 0.0001379044607997218

Sentence 3 Perplexity: 0.00013330791049822184

**Perplexities of Brown and Learner Test Corpora**

The perplexities of both test corpora were computed. The Brown Test corpus had a lower perplexity, and so the Brown Train data was better fit to the Learner Test corpus.

Brown Test Perplexities

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Unigram: 0.0013538118188978902

Bigram MLE: 0.0

Bigram Smooth: 4.6764554695938516e-05

Learner Test Perplexities

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Unigram: 0.03720716032525958

Bigram MLE: 0.0

Bigram Smooth: 0.008077995984558427