NLP – Homework assignment 1

Jianqiang Li

1.3 QUESTIONS

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

There are 28250 word types.

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

There are 498474 tokens.

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)?

brown-test.txt
unitgram percentage of unknow tokens in test:0.08040825143104006 = 8.04%
unitgram percentage of unknow word types in test:0.3124370594159114 = 31.24%

learner-test.txt
unitgram percentage of unknow tokens in test:0.06074154852780807 = 6.07%
unitgram percentage of unknow word types in test:0.194621372965322 = 19.46%

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).

brown-test.txt
bigrams percentage of unknow tokens in test:0.3615056434627639 = 36.15%
bigrams percentage of unknow word types in test:0.8794058408862034 = 87.94%

learner-test.txt
bigrams percentage of unknow tokens in test:0.3597993238084851 = 35.97%
bigrams percentage of unknow word types in test:0.8499646142958245 = 84.99%

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 .

N = 28, V=21

1. **Unigram module:**

$logP_{unit}(S_j) = log\prod_{i=1}^{n} P_{unit}(w_i) = \sum_1^n logP_{unit}(w_i) = \sum_1^n log\frac{C(w_i)}{C(all)}$

1) $logP_{unit}(S_1) = \sum_1^n log\frac{C(w_i)}{C(all)} = log\frac{C(<s>)}{N} + log\frac{C(He)}{N} + log\frac{C(was)}{N} + log\frac{C(laughed)}{N} +$

$log\frac{C(off)}{N} + log\frac{C(the)}{N} + log\frac{C(screen)}{N} + log\frac{C(.)}{N} + log\frac{C(</s>)}{N}$

$= log\frac{3}{28} + log\frac{1}{28} + log\frac{2}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + + log\frac{3}{28} + log\frac{3}{28}$ = -37.52

2) $logP_{unit}(S_2) = \sum_1^n log\frac{C(w_i)}{C(all)} = log\frac{C(<s>)}{N} + log\frac{C(There)}{N} + log\frac{C(was)}{N} + log\frac{C(no)}{N} +$

$log\frac{C(compulsion)}{N} + log\frac{C(behind)}{N} + log\frac{C(them)}{N} + log\frac{C(.)}{N} + log\frac{C(</s>)}{N}$

$= log\frac{3}{28} + log\frac{1}{28} + log\frac{2}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{3}{28} + log\frac{3}{28}$ = -37.52

3) $logP_{unit}(S_3) = \sum_1^n log\frac{C(w_i)}{C(all)} = log\frac{C(<s>)}{N} + log\frac{C(I)}{N} + log\frac{C(look)}{N} + log\frac{C(forward)}{N} +$

$log\frac{C(to)}{N} + log\frac{C(hearing)}{N} + log\frac{C(your)}{N} + log\frac{C(reply)}{N} + log\frac{C(.)}{N} + log\frac{C(</s>)}{N}$

$= log\frac{3}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{1}{28} + log\frac{3}{28} + log\frac{3}{28}$

= -43.33

2. **Bigram module:**

3. $logP_{bi}(S_j) = log\prod_{i=1}^{n} P_{bi}(w_i) = \sum_1^n logP_{bi}(w_i) = \sum_1^n log\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

1) $P_{bi}(S_1) = \sum_1^n log\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

$= log\frac{C(He)}{C(<S>)} + log\frac{C(He\ was)}{C(He)} + log\frac{C(was\ laughed)}{C(was)} + log\frac{C(laughed\ off)}{C(laughed)} + log\frac{C(off\ the)}{C(off)} +$

$log\frac{C(the\ screen)}{C(the)} + log\frac{C(screen\ .)}{C(screen)} + log\frac{C(.\ </S>)}{C(.)}$

$= log\frac{1}{3} + log\frac{1}{1} + log\frac{1}{2} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{3}{3}$ = -2.58

Bigram module zero parameters in sentence one：

$log\frac{C(He\ was)}{C(He)}$, $log\frac{C(laughed\ off)}{C(laughed)}$, $log\frac{C(off\ the)}{C(off)}$, $log\frac{C(the\ screen)}{C(the)}$, $log\frac{C(screen\ .)}{C(screen)}$,

$log\frac{C(.\ </S>)}{C(.)}$

2) $logP_{bi}(S_2) = \sum_1^n log\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

$=log\frac{C(There)}{C(<S>)} + log\frac{C(There\ was)}{C(There)} + log\frac{C(was\ no)}{C(was)} + log\frac{C(no\ compulsion)}{C(no)} +$

$log\frac{C(compulsion\ behind)}{C(compulsion)} + log\frac{C(behind\ them)}{C(behind)} + log\frac{C(them\ .)}{C(them)} + log\frac{C(.\ </S>)}{C(.)}$

$= log\frac{1}{3} + log\frac{1}{1} + log\frac{1}{2} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{3}{3}$ = -2.58

Bigram module zero parameters in sentence two：

$log\frac{C(There\ was)}{C(There)}$, $log\frac{C(no\ compulsion)}{C(no)}$, $log\frac{C(compulsion\ behind)}{C(compulsion)}$, $log\frac{C(behind\ them)}{C(behind)}$, $log\frac{C(them\ .)}{C(them)}$

, $log\frac{C(.\ </S>)}{C(.)}$

3) $logP_{bi}(S_3) = \sum_1^n log\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

=

$log\frac{C(I)}{C(<S>)} + log\frac{C(I\ look)}{C(I)} + log\frac{C(look\ forward)}{C(look)} + log\frac{C(forward\ to)}{C(forward)} + log\frac{C(to\ hearing)}{C(to)} +$

$log\frac{C(hearing\ your)}{C(hearing)} + log\frac{C(your\ reply)}{C(your)} + log\frac{C(reply\ .)}{C(reply)} + log\frac{C(.\ </S>)}{C(.)}$

$= log\frac{1}{3} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{1}{1} + log\frac{3}{3}$ = -1.58

Bigram module zero parameters in sentence three：

$log\frac{C(I\ look)}{C(I)}$, $log\frac{C(look\ forward)}{C(look)}$, $log\frac{C(forward\ to)}{C(forward)}$, $log\frac{C(to\ hearing)}{C(to)}$, $log\frac{C(hearing\ your)}{C(hearing)}$,

$log\frac{C(your\ reply)}{C(your)}$, $log\frac{C(reply\ .)}{C(reply)}$, $log\frac{C(.\ </S>)}{C(.)}$

4. **Add-one smooth Bigram module:** N = 28, V = 21

$logP_{add\_one\_bi}(S_j) = log\prod_{i=1}^n P_{add\_one\_bi}(w_i) = \sum_1^n logP_{add\_one\_bi}(w_i) = \sum_1^n log\frac{C(w_{i-1},w_i)+1}{C(w_{i-1})+V}$

1) $P_{add\_one\_bi}(S_1) = \sum_1^n log\frac{C(w_{i-1},w_i)+1}{C(w_{i-1})+V}$

$= log\frac{C(He)+1}{C(<S>)+V} + log\frac{C(He\ was)+1}{C(He)+V} + log\frac{C(was\ laughed)+1}{C(was)+V} + log\frac{C(laughed\ off)+1}{C(laughed)+V} +$

$log\frac{C(off\ the)+1}{C(off)+V} + log\frac{C(the\ screen)+1}{C(the)+V} + log\frac{C(screen\ .)+1}{C(screen)+V} + log\frac{C(.\ </S>)+1}{C(.)+V}$

$= log\frac{2}{3+21} + log\frac{2}{1+21} + log\frac{2}{2+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21}$

$+log\frac{4}{3+21}$ = -26.98

2) $logP_{add\_one\_bi}(S_2) = \sum_1^n log\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

$= log\frac{C(There)+1}{C(<S>)+V} + log\frac{C(There\ was)+1}{C(There)+V} + log\frac{C(was\ no)+1}{C(was)+V} + log\frac{C(no\ compulsion)+1}{C(no)+V} +$

$log\frac{C(compulsion\ behind)+1}{C(compulsion)+V} + log\frac{C(behind\ them)+1}{C(behind)+V} + log\frac{C(them\ .)+1}{C(them)+V} + log\frac{C(.\ </S>)+1}{C(.)+V}$

$= log\frac{2}{3+21} + log\frac{2}{1+21} + log\frac{2}{2+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} +$

$log\frac{4}{3+21}$ = -26.98

3) $logP_{add\_one\_bi}(S_3) = \sum_1^n log\frac{C(w_{i-1},w_i)}{C(w_{i-1})}$

$= log\frac{C(I)+1}{C(<S>)+V} + log\frac{C(I\ look)+1}{C(I)+V} + log\frac{C(look\ forward)+1}{C(look)+V} + log\frac{C(forward\ to)+1}{C(forward)+V} +$

$log\frac{C(to\ hearing)+1}{C(to)+V} + log\frac{C(hearing\ your)+1}{C(hearing)+V} + log\frac{C(your\ reply)+1}{C(your)+V} + log\frac{C(reply\ .)+1}{C(reply)+V} +$

$log\frac{C(.\ </S>)+1}{C(.)+V}$

$= log\frac{2}{3+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} + log\frac{2}{1+21} +$

$log\frac{2}{1+21} + log\frac{4}{3+21}$ = -30.38

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

Perplexity = 2^{-l}

$l = \frac{1}{M}logP(S_i)$, where M is the number of tokens in the sentence

Perplexity of sentence one:

Unitgram:

$l_1 = \frac{1}{M_1}logP_{unit}(S_1) = \frac{1}{9} * (-37.52) = -4.17$

Perplexity(S_1) = $2^{-l_1} = 2^{4.17} = 18.00$

$$l_2 = \frac{1}{M_2} logP_{unit}(S_2) = \frac{1}{9} * (-37.52) = -4.17$$

Perplexity(S_2) = $2^{-l_2} = 2^{4.17} = 18.00$

$$l_3 = \frac{1}{M_3} logP_{unit}(S_3) = \frac{1}{10} * (-43.33) = -4.33$$

Perplexity(S_3) = $2^{-l_3} = 2^{4.33} = 20.11$

Bigram:

$$l_1 = \frac{1}{M_1} logP_{bi}(S_1) = \frac{1}{9} * (-2.58) = -0.29$$

Perplexity(S_1) = $2^{-l_1} = 2^{0.29} = 1.22$

$$l_2 = \frac{1}{M_2} logP_{bi}(S_2) = \frac{1}{9} * (-2.58) = -0.29$$

Perplexity(S_2) = $2^{-l_2} = 2^{0.29} = 1.22$

$$l_3 = \frac{1}{M_3} logP_{bi}(S_3) = \frac{1}{10} * (-1.58) = -0.16$$

Perplexity(S_3) = $2^{-l_3} = 2^{0.16} = 1.12$

Add-one Smooth Bigram:

$$l_1 = \frac{1}{M_1} logP_{add\_one\_bi}(S_1) = \frac{1}{9} * (-26.98) = -3.00$$

Perplexity(S_1) = $2^{-l_1} = 2^{3.00} = 8.00$

$$l_2 = \frac{1}{M_2} logP_{add\_one\_bi}(S_2) = \frac{1}{9} * (-26.98) = -3.00$$

Perplexity(S_2) = $2^{-l_2} = 2^{3.00} = 8.00$

$$l_3 = \frac{1}{M_3} logP_{add\_one\_bi}(S_3) = \frac{1}{10} * (-30.38) = -3.04$$

Perplexity(S_3) = $2^{-l_3} = 2^{3.04} = 8.22$

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.

brown-test.txt
Unigram model: Perplexity: 365.14681747225984

Bigram model: Perplexity: Undefined

Bigram model with Add one smooth: Perplexity: 266.20705584529674

learner -test.txt
Unigram model: Perplexity: 409.64007748758974

Bigram model: Perplexity: Undefined

Bigram model with Add one smooth: Perplexity: 336.29905882362823

The unigram LM calculates the probability of each independent word in training data, so its perplexity is the greatest which means not good to fit the language module.

The bigram LM is undefined because there are unseen bigrams in test file. These unseen bigrams probability is undefined, so we cannot calculate the perplexity value.

The Add–one smooth bigram LM can handle the unseen data in test file by add one for each unseen bigrams. And the Add –one smooth bigram LM works better than bigram LM and unigram LM. This LM is better to fit the language module.

Learner-test file has a less perplexity in unigram LM than Brown-test file because it is written by a non-native English speaker who cannot write many unseen words in the corpus.
In Add–one smooth bigram LM, the perplexity is greater than Brown-test file because the non-native writer cannot write correct grammar or bigram words so that there are many unseen bigram data leading the perplexity to be great.