Programming assignment 1: Comparing Language Models

CSE 256: Statistical NLP

University of California, San Diego

Wei-Cheng Huang (A53267614)

1. Introduction

N-gram model is used to predict the occurrence of a word based on the occurrence of the previous N-1 words. In this project, we focus on comparing different n-gram Language models with smoothing method and hyper-parameter tunings.

2. Language Model Implementation

In this project I implemented a trigram and a bigram model with add-one / Laplace smoothing. Firstly when we read every sentence from the corpus, we need to add two token [‘*’,’*’] and a [‘EOS’] (end of sentence) at the beginning and the end of the sentence respectively. In order to compute the conditional probability of the first word and to know the end of the sentence. We first compute all counts of all words, then if the word count is less than a specific threshold, we tend to define it as ‘UNK’, which means that our model doesn’t know this word any more. Here, we choose the threshold as 4, which means if the total count of specific word is less than 4, it would be defined as ‘UNK’ by the model. Such method could help our model be more general on frequently appearing words.

There are some hyper parameters I tried to tune. The rare word threshold and the δ in Laplace smoothing (See Laplace smoothing formula below).n

$$\text{Laplace Smoothing : } q(\omega_i \,|\, \omega_{i-2}, \omega_{i-1}) = \frac{count(\omega_{i-2}, \omega_{i-1}, \omega_i) + 1}{count(\omega_{i-2}, \omega_{i-1}) + \delta \,|\, \mathscr{V} \,|}$$

A. Delta δ tuning

The following table shows that the lower δ is, the lower perplexity we can get. I used model trained by brown training dataset to compute the perplexity of brown’s train, dev, and test dataset given delta from 0.1 to 0.9. Therefore, I chose 0.001 as my final δ to run the comparison test.

Delta Tuning									
	0.1	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
Train	542.9	1047.8	1548.0	2046.2	2543.4	3040.0	3536.1	4032.0	4527.6
Dev	1041.7	2059.7	3074.4	4087.7	5100.2	6112.2	7123.9	8135.4	9146.7
Test	1044.0	2064.4	3081.6	4097.3	5112.3	6126.8	7140.9	8154.9	9168.6

B. UNK threshold tuning

I used training set of Brown to train the model by adjusting UNK threshold value, and found that the higher the threshold is , the low perplexity I can obtain, which might seem like the higher threshold the better. However, if excessively high threshold would result in too many 'UNK', which would ruin the whole model , and make no sense to sample sentence. Therefore, I keep threshold at 4 to run my comparison test, which construct relative sample sentences.

UNK threshold tuning								
	2	3	4	6	8	10	20	30
Train	973	694	542	380	296	242	125	78.3
Dev	1747	1291	1041	767	621	526	305	41.9
Test	1752	1294	1044	769	622	527	306	41.8

3. Analysis of In-Domain and Out of Domain Text

- Empirical Evaluation: By observing from the table below, we can see that Trigram model performs much better no matter on Brown, Reuters, or Gutenberg dataset, according to the perplexity. The reason might be that Trigram model considers the sequence of the sentence, rather than only the frequency of each word.

- Qualitative Analysis: I started sample sentences with "With", and "Next". We can see that the sample sentences of Unigram model are much shorter and make no sense, since it only consider the frequency of each word and the 'EOS' appears more frequent, which leads to shorter sentences. On the other hand, sample sentences of Trigram model tend to be longer and make more sense compared with unigram. They are more grammatically correct than unigram, although the paragraphs are still composed of unrelated words.

4. Analysis of Out-of-domain Text

- Empirical Evaluation: By observing from the table below, we can see that unigram model trained on Brown has similar performance on Gutenberg dataset, which means that they might have similar word set. However, there is a different story in trigram model, that trigram model trained on brown has similar perplexity on Reuters and Gutenburg. It is possible that Reuters has more similar word sequence as brown, or although Gutenberg has similar word set, the word combinations are much different than the training set Brown.
- Qualitative Analysis: Unigram model performs worse than Trigram model in both in-domain or out-of-domain text, since it considers only the word frequency but not the sequence of the sentence.

(A)Unigram Result

sample 1: With are it the out and

sample 2: Next In but the poem two capsule and microscope Mrs infectious failure seen pre bounds welcome Catherwood one in she hope to anode evident explores patiently hands

```
x train
             brown    reuters    gutenberg
---------  -------  ---------  -----------
brown      1513.8     6780.82     1758.06
reuters    3806.39    1471.21     4882.8
gutenberg  2616.57   12420.1       982.572
-------------------------------
x dev
             brown    reuters    gutenberg
---------  -------  ---------  -----------
brown      1589.39    6675.63      1739.41
reuters    3808.87    1479.09      4833.88
gutenberg  2604.28   12256.3        991.5
-------------------------------
x test
             brown    reuters    gutenberg
---------  -------  ---------  -----------
brown      1604.2    6736.6        1762.01
reuters    3865.16    1500.69      4887.47
gutenberg  2626.05   12392.5       1005.79
```

(B) Trigram Result

sample 1: With the astonishment ambiguous Daniel door uncovered bundles Barnard followed formidable throat logs engendered ripple thrived supportive Tech suspense Pip TSH theft variable veranda Anyway invariably Tilghman gaiety reactors truce gracefully disagreement coincide Dead mentions watch slowed brightly purely newspapers devise tradition ax possessions married flair Wall fellow unemployment dialect intensity brain preparation simply democratic communion sedans material addition sums repeal load Judy ivory Not acquainted broken chimney 71 During Jupiter vest Many intersection specific mee uncommon mimesis aims Slim foresight shrubs North polyether realistically Hengesbach aborigine applies screwed Ministry Unfortunately Dei dismissed consolidation facts spaced law accuse ascribed Later forbids previously

sample 2: Next 1919 Slim attaining Roman lost Walton Ages obscured uncertainty groupings myriad nurse taste construction mystery Institute Hunter stride forms roles emotions frequent Arthur integrity northeast 400 tightened fairly guards packed Wisman realized magnificent deeply Did scented Ramsey Germans Ford blot sidewise Generally spur bands Drug Montgomery minerals offered crew West frantically humbly screen rigid towards concrete stare create inning Dog grow Barnes adventures breast committee stadium Lines autonomy Helen waves awaited choose choked 3rd registrant poetic distorted respondent emancipation knee immature centuries consequences swallowed explode sweaters dragged morality Common Council scope Determine game eternal Rankin warfare cycles Publication spin defending flared

```
x train
              brown    reuters    gutenberg
---------  --------  ---------  -----------
brown       542.903   1182.38      1078.21
reuters    1040.99     360.086     1194.47
gutenberg  1060.09    1404.5        474.985
-------------------------------
x dev
             brown    reuters    gutenberg
---------  -------  ---------  -----------
brown      1041.7    1236.39      1143.08
reuters    1158.92    602.849     1244.74
gutenberg  1214.24   1463.07       787.979
-------------------------------
x test
             brown    reuters    gutenberg
---------  -------  ---------  -----------
brown      1044.05   1235.05      1141.89
reuters    1161.11    606.528     1244.81
gutenberg  1215.89   1459.38       790.225
```

5. Adaptation

In this session, we try to use a small fraction (randomly shuffling) of Brown's training data to see if it can obtain good result on Reuters' dataset. The result shows that the perplexity decreases as we use smaller subset of data, which I think it doesn't make sense. Again, I think that the perplexity doesn't certainly mean that it is better result. We also have to consider the sample sentences.

	1/2	1/4	1/6	1/8	1/10
Brown	60.9	43.8	33.9	27.45	22.49
Reuter	64.8	45.6	34.8	27.96	23.86

6. Conclusion

According to the result above, the Trigram makes more sense than Unigram model on aspect of sample sentences. And the perplexity of unigram is much higher than Trigram model. However, I think there is no positive correlation between if the perplexity is low or the sentence makes sense. It is very hard to determine whether the model performs robust only by perplexity or construction of sample sentences. We might need to develop better method on such kind of problems.