Programming assignment 1: Language Modeling

Gitika Meher
Department of Electrical and Computer Engg.
University of California San Diego
9500 Gilman Dr, La Jolla, CA 92093
gkarumur@eng.ucsd.edu

1 Language Model Implementation

1.1 Model Description

I implemented the n-gram model with linear interpolation. Specifically, I implemented the Trigram model.i.e)a model which depends on the previous two word for the estimating probability of the current word. To model the unseen words in the test or the validation data set, I changed all the words appearing only once in the train set vocabulary to 'UNK'. Since this number is almost 10% of the vocabulary size, I modelled all the words appearing only once and starting with 'Z' as 'UNK' so that my language model still has a representation of words unknown to the vocabulary and the number of 'UNK's is not too many. This way, When I try to generate new sentences using my language model, I don't see many 'UNK's. I appended '*', '*' in front of every sentence to mark the beginning of the sentence and 'EOS' at the end of a sentence to mark the end. The model was implemented using a dictionary where the keys of the dictionary are tuples storing the previous two words following the current word. The values of this dictionary are again implemented using a dictionary to store the number of times each word was observed given the tuple key.

1.1.1 Trigram Model

A trigram language model consists of a finite set V, and a parameter q(w|u, v) for each trigram u, v, w such that w belongs to V combined with the set {STOP}, and u, v belong to V combined with the set {*}. The value for q(w|u, v) can be interpreted as the probability of seeing the word w immediately after the bigram (u, v). For any sentence x1 . . . xn where xi belongs to V for i = 1 . . .(n 1), and xn = STOP, the probability of the sentence under the trigram language model is

$$p(x_1 \dots x_n) = \prod_{i=1}^{n} q(x_i | x_{i-2}, x_{i-1})$$

1.2 Smoothing method Description

I used the Linear interpolation smoothing model to account for overcoming the short-comings of the trigram model. We define the trigram, bigram, and unigram maximum-likelihood estimates as shown by the equations where we have extended our notation: c(w) is the number of times word w is seen in the training corpus, and c() is the total number of words seen in the training corpus.The trigram, bigram, and unigram estimates have different strengths and weaknesses. The unigram estimate will never have the problem of its numerator or denominator being equal to 0: thus the estimate will always be well-defined, and will always be greater than 0 (providing that each word is seen at least once in the training corpus, which is a reasonable assumption). However, the unigram estimate completely ignores the context (previous two words), and hence discards very valuable information. In contrast, the trigram estimate does make use of context, but has the problem of many of its counts

$$q_{ML}(w|u,v) = \frac{c(u,v,w)}{c(u,v)}$$

$$q_{ML}(w|v) = \frac{c(v,w)}{c(v)}$$

$$q_{ML}(w) = \frac{c(w)}{c()}$$

being 0. The bigram estimate falls between these two extremes. The idea in linear interpolation is to use all three estimates, by defining the trigram estimate as shown by the equation below. Here

$$q(w|u,v) = \lambda_1 \times q_{ML}(w|u,v) + \lambda_2 \times q_{ML}(w|v) + \lambda_3 \times q_{ML}(w)$$

$\lambda 1$, $\lambda 2$, $\lambda 3$ are three additional parameters of the model, each of them lying between 0 and 1 and they add up to 1. This is similar to taking a weighted average of all three estimates.

1.3 Hyper-Parameter tuning

For ease of adjusting the hyper-parameters, a heuristic dependency is introduced such that there is only one hyper-parameter γ to adjust. This assignment respects the conditions that $\lambda 1$, $\lambda 2$, $\lambda 3$ have to

$$\begin{aligned}
\lambda_1 &= \frac{c(u,v)}{c(u,v)+\gamma} \\
\lambda_2 &= (1-\lambda_1) \times \frac{c(v)}{c(v)+\gamma} \\
\lambda_3 &= 1-\lambda_1-\lambda_2
\end{aligned}$$

follow. Under this definition, it can be seen that $\lambda 1$ increases as c(u, v) increases, and similarly that $\lambda 2$ increases as c(v) increases. In addition we have $\lambda 1 = 0$ if c(u, v) = 0, and $\lambda 2$ = 0 if c(v) = 0. The value for γ can be chosen by maximizing log-likelihood of a set of development data. The value of γ is chosen such that the lowest perplexity is observed on the validation/ development dataset.

2 Analysis of In-Domain Text

2.1 Perplexity

The table below shows the perplexity values after hyperparameter tuning for different corpora.

Model	Train	Dev	Test
Reuters	29.73	225.3	231.53
Brown	110.47	812.35	822.66
Guetenberg	87.89	373.77	378.25

2.2 Comparison with the unigram model

The table below presents the perplexity of the unigram model for different data corpora.

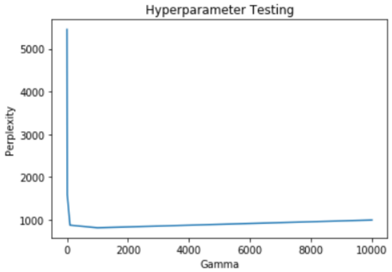
Model	Train	Dev	Test
Reuters	1471.21	1479.1	1500.695
Brown	1513.80	1589.387	1604.1
Gutenberg	982.572	991.5	1005.79

2.3 Comparison with hyper-parameters

γ is the hyper-parameter chosen for this model. Values of gamma are plotted against different development set perplexities. The gamma which outputs the lowest perplexity on the dev set is chosen.

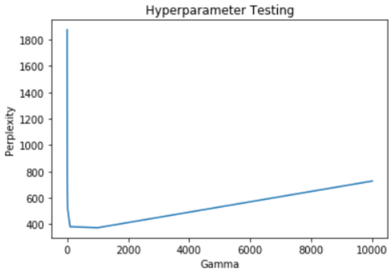
2.3.1 Brown

Gamma	1	10	100	1000	10000
Perplexity of dev-set	5453.65	1588.58	873.6	812.35	993.4



2.3.2 Gutenberg

Gamma	1	5	10	20	100	1000	10000
Perplexity of dev-set	1875.4	831.7259	638.48	515.545	381.77	373.77	728.1



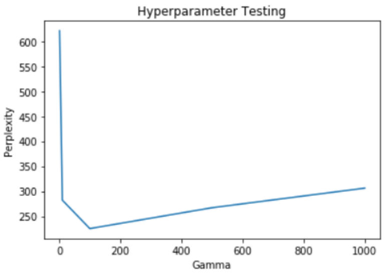
2.4 Reuters

Gamma	1	10	100	500	1000
Perplexity of dev-set	621.03	282.20	225.30	267	306.2

2.5 Examples of Sampled sentences

2.5.1 Brown

1. This impressed prefer and paso noted law because he went the street and tide
2. This is for words and detachment ran literature event in architecture circumstances to Red



2.5.2 Gutenberg

1. And as much comfort and to be forward to be renewing what always made an agitation mizzen had done.
2. She says the hall that hath the priests had believed other the flesh and set my house

2.5.3 Reuters

1.lt MUNICIPAL FINANCIAL CORP lt GOR COMPLETES SALE OF UNIT Bell and there has also been decline in the year ending December 1986 withdrawal of the area said Exxon Corp lt
2.Poor stock purchases it would suspend reduced lt WMNG gas reeserves in the current account balance of 250 mln stg in outstanding least 92 cts Net 40 pct

3 Analysis of Out-of-Domain Text

3.1 Perplexity

Model Trained on	Model Tested on	Perplexity of the test data
Reuters	Brown	2874.23
Reuters	Gutenberg	4260.18
Brown	Gutenberg	1070.14
Brown	Reuters	2917.21
Gutenberg	Brown	1356.34
Gutenberg	Reuters	3683.68

3.2 Comparison with the unigram model

Model Trained on	Model Tested on	Perplexity of the test data
Reuters	Brown	3865.15
Reuters	Gutenberg	4887.46
Brown	Gutenberg	982.57
Brown	Reuters	6736.60
Gutenberg	Brown	2626.05
Gutenberg	Reuters	12392.53

3.3 Comparison of performance of different corpora

From the table in the section 3.1, we can see that the test perplexity values of Data trained on Brown corpus and tested on Gutenberg and vice-versa are the lowest when compared to all. This is probably because these two corpora are similar in nature when compared to the Reuters corpus.
When compared to the Unigram model trained on Brown and tested on Gutenberg, similar linear interpolated model was observed to score worse perplexity. This is probably because, since similar sentences are present in both Brown and the Gutenberg corpus, the speculation of the unigram model is better.

4 Adaptation

4.1 Approach

Since from the out of domain analysis above, we notice that Brown corpus and Gutenberg corpus are similar, I trained my model on Brown corpus with different proportions of the Gutenberg Corpus to capture the perplexity.

Ratio added	0.25	0.5	0.75	1
Perplexity of test-set	557.52	458.619	430.30	410.37

4.2 Relevant Comparisons

According to the table above, we notice that by adding 0.5% of Gutenberg trainset to the Brown corpus while training, we see a perplexity relatable to 378.25 which is the test set perplexity of Gutenberg corpus using the full training set. We see that the perplexity observed on the test set becomes better as we keep adding more of the Gutenberg Trainset.