Sentence Completion using Language Modelling

Vijeta Agrawal

$28^{th} February$ 2019

1 Aim

The main focus of this report is to show the results of three languge models(Unigram Model, Bigram Model and Bigram Model with smoothing) which perform sentence completion, i.e. given a sentence with a missing word to choose the correct one from a list of candidate words. The way we use a language model for this problem is that we consider a possible candidate word for the sentence at a time and then ask the language model which version of the sentence is the most probable one.

2 Training & Test Sets

The corpus that trained my language models is "news-corpus-500k.txt", which is a small subset of the 1 Billion Word Benchmark. The text was already tokenized and split into sentences.

The sentences to be completed together with the candidate words are in "questions.txt". The word to be completed is denoted by '____', while the pair of candidate words is at the end of the line (e.g. weather/whether). The character ':' between the sentence and the candidates is not part of the sentence.

3 Accuracy discussion for language models

These language models uses the dictionary structure to store the probabilities for unigrams and bigrams. Before building these language models, the unnecessary punctuations from both training and test sets were removed for better tokenization. In bigram models, begin ('<s>') and end ('<\s>') symbols were appended before the first and after the last words of the sentences respectively.

Model	Smoothing	Accuracy Percentage
Unigram Model	No	50
Bigram Model	No	70
Bigram Model	Yes	90

Table 1: Accuracy Results

3.1 Unigram Model

This model calculates the probability of each word in the corpus for training the model and based on those probabilities predicts the sentence with the candidate word which has the highest probability. The accuracy computed for this model is **50%** because the probabilities of the unigrams that we are using gives us its importance in our training corpus but not in the sentence we are predicting it for. It means that we are not taking the context into account while making our predictions. Moreover, for the words in the test set which were not present in the training set, it assigns a zero probability which results in the zero probability prediction of the sentence which further reduces the accuracy of the model.

3.2 Bigram Model without Smoothing

This model calculates the probability of each bigram in the corpus for training the model and based on the conditional probability of each bigram given the probability of the first word in it, predicts the sentence with the candidate word which has the highest probability. The accuracy computed for this model is **70%** because the conditional probabilities of the bigrams that we are using, takes the context of the sentence into account while making the predictions. But, for the bigrams in the test set which were not present in the training set, it assigns a zero probability which results in the zero probability prediction of the sentence which reduces the accuracy of the model.

3.3 Bigram Model with Smoothing

This model also calculates the probability of each bigram in the corpus for training the model and based on the conditional probability of each bigram given the probability of the first word in it, predicts the sentence with the candidate word which has the highest probability. The accuracy computed for this model is **90%** which is highest of the three models because we are using a concept called smoothing while calculating the sentence probability in our predictions. Smoothing refers to assigning some of the total probability mass to

unseen words or n-grams, in this case bigrams. We are using a simple (Add-one) smoothing i.e. assigning a count of 1 to unseen bigrams.

The accuracy difference of these models is evident in the figure below.

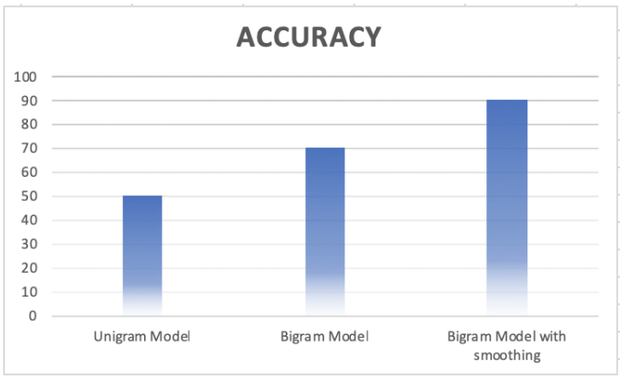

Figure 1: Comparison of accuracy percentage for different models

The preview of the result is shown in the figure below. The total time taken for running these models is around 24 secs.

```
(base) pc-207-52:Assignment 2 vijetaagrawal$ python3 lab2.py news-corpus-500k.txt questions.txt

RUNNING UNIGRAM MODEL--------

The predicted sentences are:

Correct Sentence:  i don't know whether to go out or not .
Correct Sentence:  we went through the door to get inside .
Incorrect Sentence:  they all had a peace of the cake .
Correct Sentence:  she had to go to court to prove she was innocent .
Correct Sentence:  we were only allowed to visit at certain times .
Correct Sentence:  she went back to check she had locked the door .
Incorrect Sentence:  can you here me .
Incorrect Sentence:  do you usually eat cereal for breakfast .
Incorrect Sentence:  she normally choose with her mouth closed .
Both sentences return zero probability.

Accuracy of the Unigram Model is:  50.0 %


RUNNING BIGRAM MODEL--------

The predicted sentences are:

Correct Sentence:  i don't know whether to go out or not .
Correct Sentence:  we went through the door to get inside .
Correct Sentence:  they all had a piece of the cake .
Correct Sentence:  she had to go to court to prove she was innocent .
Correct Sentence:  we were only allowed to visit at certain times .
Correct Sentence:  she went back to check she had locked the door .
Correct Sentence:  can you hear me .
Both sentences return zero probability.
Both sentences return zero probability.
Both sentences return zero probability.

Accuracy of the Bigram Model is:  70.0 %


RUNNING BIGRAM MODEL WITH SMOOTHING--------

The predicted sentences are:

Correct Sentence:  i don't know whether to go out or not .
Correct Sentence:  we went through the door to get inside .
Correct Sentence:  they all had a piece of the cake .
Correct Sentence:  she had to go to court to prove she was innocent .
Correct Sentence:  we were only allowed to visit at certain times .
Correct Sentence:  she went back to check she had locked the door .
Correct Sentence:  can you hear me .
Correct Sentence:  do you usually eat cereal for breakfast .
Incorrect Sentence:  she normally choose with her mouth closed .
Correct Sentence:  i'm going to sell it on the internet .

Accuracy of the Bigram Model with smoothing is:  90.0

Total time taken for all models to run:  23.96330499649048
```

Figure 2: Prediction results

4 Conclusion

The language modelling for sentence completion is fairly improved when we take the context of the prediction sentence into account and its further improved when we use the concept of smoothing to avoid zero probability predictions.