Homework 1: N-gram Language Models

Guangyu Lin, EID: gl8429

***Abstract*—The short report is intended to present homework 1, N-gram Language Models. We modified a simple, typical N-gram language model into backward bigram model and bidirectional bigram model. Then we trained and experimentally tested them on English datasets, Airline Booking Conversations(ATIS), Wall Street Journal(WSJ), and Brown. Finally, we present tables and figures of comparative results, then we concludes backward bigram model is a little better than forward bigram model in diverse dataset and forward bigram model is better in spoken discourse. And bidirectional bigram model is the best one of three, because it is a combinational prediction of another two models.**

I. Introduction

This report explores a modification of a simple, typical N-gram language model [2] and experimentally test it on English dataset, which includes Airline Booking Conversations(ATIS), Wall Street Journal(WSJ), and Brown. The first modification is to produce a "backward" bigram model that models the generation of a sentence from right to left. The second modification is to combine the forward and backward model. Additionally, we weight the prediction of both models equally when interpolating a probability for each word/token in the sentence. For each bigram language model, we produce a trace analogous to the sample BigramModel trace file. Then, we draw several tables and several figures of comparative results and states a discussion about the difference and similarities which we found. [1]

The report is organized as following, the algorithms of backward bigram model and bidirectional bigram model are described briefly in Section II. The results table and figure is provided in Section III. We discuss our results in Section IV. And Section V is the conclusion of the report.

II. Algorithms

A. Backward Bigram Model

The one significant difference between a left-to-right and a right-to-left(backward) model is that the backward model's final prediction is to predict sentence-start($< S >$) rather than sentence-end($< /S >$). Therefore, I modify the original code of Bigram Model before I train data and calculate the sentence probabilities by reversing the sentences. For smoothing part, we use the same weights as in forward bigram model, $\frac{1}{10}$ on the unigram and $\frac{9}{10}$ on the bigram and we also replace the leftmost token of each type with "$< UNK >$" token. Then we derived from the following formula:

$$\frac{1}{10}p_{unigram} + \frac{9}{10}p_{backward-bigram}$$

B. Bidirectional Bigram Model

Bidirectional Bigram Model combines the forward and backward model. When determining the probability of a word in the sentence, it should use both the estimate from its forward and backward contexts by linearly interpolating the probability predicted for that token by BigramMode and BackwardBigramModel. By default, just weight the prediction of both modes equally when interpolating a probability for each word/token in the sentence. [1]

To calculate the combination probability of these to models, I derived from the following formula:

$$\frac{1}{10}p_{unigram}+\frac{9}{10}\left(\frac{1}{2}p_{forward-bigram} + \frac{1}{2}p_{backward-bigram}\right)$$

We can also play with different ratio of weight of the prediction.

III. Experiment Results

We train and test three datasets (ATIS, WSJ, BROWN) with a ratio of 9:1. Table I is the results of training dataset: ATIS and Table II is the results of testing dataset: ATIS. The results include preplexity and word preplexity. Similarly, Table III - Table VI are training and testing results of the other two datasets.

TABLE I
Training Data: ATIS

Model	Preplexity	Word Preplexity
Forward Bigram	9.043	11.636
Backward Bigram	9.013	10.592
Bidirectional Bigram	6.079	7.235

TABLE II
Testing Data: ATIS

Model	Preplexity	Word Preplexity
Forward Bigram	19.341	24.054
Backward Bigram	19.364	27.161
Bidirectional Bigram	10.227	12.700

TABLE III
Training Data: WSJ

Model	Preplexity	Word Preplexity
Forward Bigram	74.268	88.890
Backward Bigram	74.268	86.660
Bidirectional Bigram	40.204	46.514

IV. Discussion

A. Forward V.S. Backward Bigram Models

As far as we know, forward bigram models is to predict where the sentences end, whereas, backward bigram models is to predict where the sentences begin.

TABLE IV
TESTING DATA: WSJ

Model	Preplexity	Word Preplexity
Forward Bigram	219.715	275.118
Backward Bigram	219.520	266.352
Bidirectional Bigram	104.796	126.113

TABLE V
TRAINING DATA: BROWN

Model	Preplexity	Word Preplexity
Forward Bigram	93.519	113.360
Backward Bigram	93.509	110.783
Bidirectional Bigram	52.375	61.469

TABLE VI
TESTING DATA: BROWN

Model	Preplexity	Word Preplexity
Forward Bigram	231.302	310.667
Backward Bigram	231.206	299.686
Bidirectional Bigram	130.157	167.487

From the previous six tables, we find that the preplexity or word preplexity of the training data of each dataset is better than of the testing data of each dataset, which is obvious. We also find that the preplexity of ATIS is better than the other two datasets because of its small. For ATIS(airline booking conversations), the forward bigram model is better than the backward bigram model, which may because of spoken discourse. At the beginning of a sentence, we would like to insert greetings or other interjections.

For other two datasets, WSJ(Wall Street Journal) and Brown, the backward bigram model is better than the forward bigram model, which may because of their diverse datasets.

B. Bidirectional V.S. Backward V.S. Forward Bigram Models

To deep dive the results tables, we combine some data from the previous tables, then we get Table VII and Table VIII, which are directly in Figure 1 and Figure 2 . For each dataset, forward bigram model is similar to backward bigram model and bidirectional bigram model is almost twice better than the other two models. Why is that? In my opinion, the forward bigram models look like $< left, ? >$ and the backward bigram models look like$<?, right >$. The combinational bidirectional bigram models look like, $< left, ? >$ and $<?, right >$, which combine the predictions of forward and backward bigram models. Therefore, the word preplexity is better than the other two models.

TABLE VII
WORD PREPLEXITY OF TRAINING DATA: ATIS, WSJ, BROWN

Dataset	Forward Bigram	Backward Bigram	Bidirectional Bigram
ATIS	10.591	11.636	7.235
WSJ	88.890	86.660	46.514
BROWN	113.360	110.783	61.469

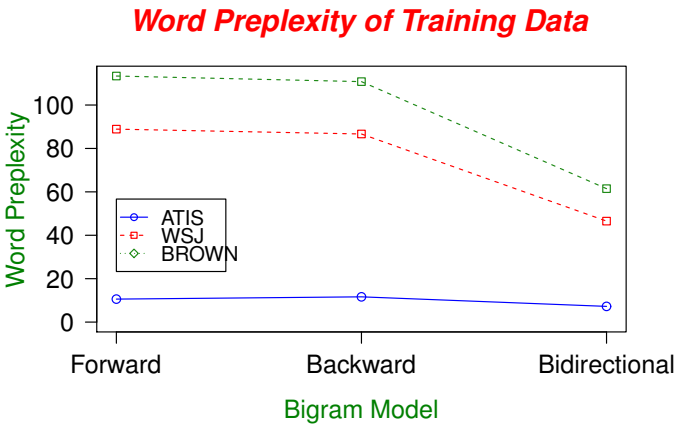

Fig. 1. Word Preplexity of Training Data

TABLE VIII
WORD PREPLEXITY OF TESTING DATA: ATIS, WSJ, BROWN

Dataset	Forward Bigram	Backward Bigram	Bidirectional Bigram
ATIS	24.054	27.161	12.700
WSJ	275.118	266.352	126.113
BROWN	310.667	299.686	167.487

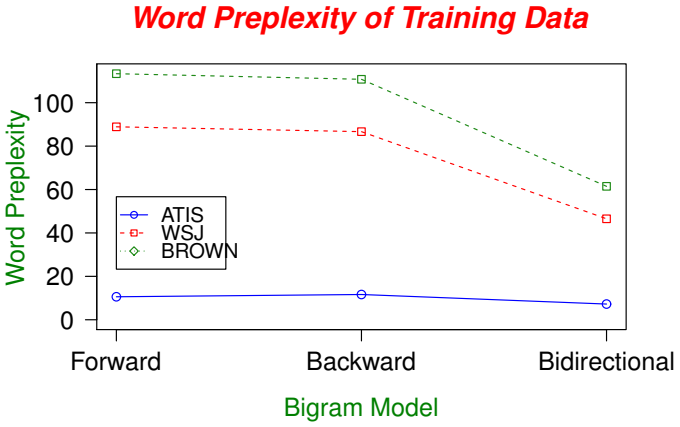

Fig. 2. Word Preplexity of Testing Data

V. CONCLUSION

In this report, we talked about the algorithms of backward bigram models and bidirectional bigram models. From the experimental results, we find backward bigram model is a little better than forward bigram model in diverse dataset and forward bigram model is better in spoken discourse. And bidirectional bigram model is the best one of three, because it is a combinational prediction of another two models.

REFERENCES

[1] https://www.cs.utexas.edu/ mooney/cs388/hw1.html
[2] Martin, James H., and Daniel Jurafsky. "Speech and language processing." International Edition (2000).