N Gram Language Models

Kunal Lad : KL28697

Overview

A language model is a model of legal sentences in a language. Statistical language models have generally been found to be more useful than formal grammars in various NLP applications like Speech Processing, Machine Translation, OCR, context based spell correction etc. N-Gram language models are statistical models based on (N - 1) order Markov model (i.e. any word in a sentence is dependent only on previous N - 1 words). In this project we implement a Backward Bigram Model and a Bidirectional Bigram Model using the library of given Bigram Model and compare their performances on 3 standard data sets.

Methodology

We generalized the provided BigramModel class by adding a mode (Forward/Backward). In the backward mode model, sentences (including start and end tokens) are reversed. We then used this generalized class to implement a BidirectionalModel class which contains a BigramModel of both types and combines their result by linearly interpolation providing equal weight to results from both models.

Results and Discussion

In this section we present the obtained results for all 3 models and compare the performance based on various parameters like data set, length of sentences in test set etc. For all the experiments we used 90% data for training and 10% for testing. We have measured the performance in terms Perplexity and Word Perplexity metrics which measure weighted branching factor in predicting next word.

DataSet	Type	Bigram Model Perplexity | Word Perplexity	Backward Bigram Model Perplexity | Word Perplexity	Bidirectional Bigram Word Perplexity
Atis	Train	9.04 | 10.59	9.01 | 11.63	7.23
Wsj	Train	74.26 | 88.89	74.26 | 86.66	46.51
Brown	Train	93.51 | 113.35	93.50 | 110.78	61.46
Atis	Test	19.34 | 24.05	19.36 | 27.16	12.70
Wsj	Test	219.71 | 275.11	219.51 | 266.35	126.11
Brown	Test	231.30 | 310.66	231.20 | 299.68	167.48

Table 1 : Performance on training set and all sentences in test set.

From the table 1 we can conclude that perplexity of backward model is surprisingly almost same as forward model but word perplexity of backward bigram model is worse than forward model for small data set and is better for bigger datasets. Results suggest that backward model performs better and better as compared to forward model as the size of dataset increases. Possible explanation is that backward dependencies are stronger than forward ones and require more data than forward ones to be learnt reasonably.

Also we can clearly observe that bidirectional model outperforms individual models and the margin of improvement is huge for large data sets. This is intuitively clear because bidirectional model leverages information from both the models and thus has more context.

DataSet	Set Type	Bigram Model Perplexity | Word Perplexity	Backward Bigram Model Perplexity | Word Perplexity	Bidirectional Bigram Word Perplexity
Atis	Train	9.44 | 13.06	9.43 | 15.34	9.33
Wsj	Train	29.43 | 73.50	29.44 | 79.29	40.56
Brown	Train	34.84 | 83.48	43.84 | 65.16	36.98
Atis	Test	14.28 | 20.99	14.21 | 24.13	12.15
Wsj	Test	63.37 | 226.56	63.37 | 247.82	138.29
Brown	Test	70.08 | 185.70	70.09 | 150.37	80.98

Table 2 : Performance on Short (<= 5 words) sentences in training and test set.

DataSet	Set Type	Bigram Model Perplexity | Word Perplexity	Backward Bigram Model Perplexity | Word Perplexity	Bidirectional Bigram Word Perplexity
Atis	Train	8.96 | 10.20	8.92 | 11.07	6.91
Wsj	Train	74.91 | 89.00	74.91 | 86.71	46.55
Brown	Train	94.41 | 113.63	94.40 | 111.25	61.71
Atis	Test	21.34 | 24.99	21.41 | 28.08	12.85
Wsj	Test	222.14 | 275.45	221.94 | 266.47	126.04
Brown	Test	237.72 | 313.79	237.62 | 303.73	169.87

Table 3 : Performance on long sentences (> 5 words) in training and test set.

Tables 2 and 3 show the results of evaluating the models on short and long sentences in training and test sets. Results show that for all models, both perplexity and word perplexity are much smaller for short sentences as compared to long sentences. This is expected because long sentences are more likely to have long term dependencies which are not captured by our models. We can try using separate models for long and short sentences to see if it works better.

One thing which we can observe for all datasets and all types of sentences is that perplexity is always less than word perplexity. This is expected because predicting start and end of English sentences is generally much easier than predicting words in middle of the sentences. But what is surprising is that it reduces the measure by a huge amount for large datasets. This shows that models learnt from huge datasets are extremely good at predicting start and end of sentences.