CS 388 NATURAL LANGUAGE PROCESSING

HomeWork 1 N-Gram Language Models

VIvek Natarajan

Introduction:

The goal of the assignment is to understand the working of N-Gram Language Models. This involved reading and understanding a simple implementation of the Forward Bigram Model in Java. The Forward Bigram Model was then modified to implement a Backward Bigram Model that models the generation of a sentence from right to left Experiments were performed on three different corpus to evaluate the comparative performance of the Forward and Backward Bigram Model based on Word Perplexity measures. Then the models were combined to implement a Bidirectional Bigram Model which used a weighted sum of the estimates from the Forward and the Backward Bigram Model. Experiments were again performed on the same corpus to obtain a comparative performance analysis of the three models.

Experiments:

The experiments were performed using atis, wsj and brown test corpus. The lambda parameters (weights for combining unigram and bigram probabilities) and the weights of the forward and backward model were also modified. Besides, experiments were also performed on varying the extent of the percentage of test data. The results have been summarized in the tables and are further analyzed in the analysis and discussion sections.

Results:

EXPERIMENTAL RESULTS ON TRAINING DATA:

Word Perplexity Values (Lambda1 = 0.1, Lamda2 = 0.9, forward = 0.5 and backward = 0.5)

Testing Corpus	Forward Bigram	Backward Bigram	Bidirectional Bigram
Atis	10.5919539986291	11.636203016013205	7.235173934082009
WSJ	88.89008380713945	86.6602647021424	46.51444509063141
Brown	113.35954408376686	110.78289581654053	61.46886647115352

EXPERIMENTAL RESULTS ON TESTING DATA:

Word Perplexity Values (Lambda2 = 0.1, Lambda2 = 0.9, forward = 0.5 and backward = 0.5)

Testing Corpus	Forward Bigram	Backward Bigram	Bidirectional Bigram
Atis	24.053999598153656	27.16138806179997	12.700210156018738
WSJ	275.1178149885602	266.3515745993808	126.1131573880389
Brown	310.66735613437913	299.68570342320015	167.48711091425574

EXPERIMENTAL RESULTS ON TESTING DATA:

Word Perplexity Values (Lambda2 = 0.1, Lambda2 = 0.9, forward = 0.4 and backward = 0.6)

Testing Corpus	Forward Bigram	Backward Bigram	Bidirectional Bigram
Atis	24.053999598153656	27.16138806179997	12.949317973874642
WSJ	275.1178149885602	266.3515745993808	127.07234618406198

Brown	310.66735613437913	299.68570342320015	168.3118703042146

EXPERIMENTAL RESULTS ON TESTING DATA:

Word Perplexity Values (Lambda2 = 0.1, Lambda2 = 0.9, forward = 0.6 and backward = 0.4)

Testing Corpus	Forward Bigram	Backward Bigram	Bidirectional Bigram
Atis	24.053999598153656	27.16138806179997	12.696005051120713
WSJ	275.1178149885602	266.3515745993808	127.80855365063704
Brown	310.66735613437913	299.68570342320015	169.86008221231313

Observations:

The key observations from the experiments are summarized here:

1> The backward model outperformed the forward model on corpus with larger sizes(brown and wsj).
2> The Bidirectional model gave approximately 50% improvement on the Forward and Backward models in all three test corpus. The Word Perplexity scores reduced by half on all three test corpus
3> Varying the lambda parameters (the weights to combine the unigram and bigram estimates) resulted in increased Word Perplexity scores on all three corpus.
4> Equal weights to combine the estimates from the Forward and Backward model in the Bidirectional Model gave the best results.
5> Decreasing the size of the training sentences resulted in increased word perplexity scores.

Discussion:

The word perplexity scores for the Backward model showed an increase with respect to the atis corpus. However the atis corpus is too small to make any inference on performance characteristics. On the other hand, with the brown and wsj corpus, the backward model gave reduced word perplexity scores. Given that the datasets are comparatively larger in size, we can say that the backward model performs better than the forward model. My hypothesis is that the backward bigram model provides complementary information to the forward model. However, in some cases the backward model has more information. Consider, the phrase "He is playing". On seeing the word "is", there are variety of verbs which could potentially follow it. Hence the bigram "is playing" has reduced probability. However, when we see the word "playing", the likely combination of words with it are reduced to a subset of tense indicators like "is" and "was". Because the brown corpus has its source from literature works, its likely to contain action phrases with the verb coming at the end. This enables better prediction for the backward model. In case of wsj corpus, the tendency is for the corpus to contain noun or verb phrases which end with the noun like "bear markets" or "accrued interest". Here again the terms at the end contain more information and better model the text corpus generation.

The bidirectional model definitely gives much better results than either of forward or backward models. This is because the two models generally contain and complementary information and the combination of the two models gives much improved results. Consider the phrase, "He is playing cricket". In my opinion, the forward model better the phrase "playing cricket" while the backward model better models the phrase "is playing" as previously explained. Thus because of their complementary nature, the combination of the two models give 50% improvement in results. Other observations show that varying the weights of the two models in the bidirectional model does not make a significance difference even though the backward model is slightly better than the forward model. Increasing the unigram weights in the interpolation probabilities also gives reduced performance. More experiments need to be performed to evaluate the best smoothing measures for the two models.