CS 388 Natural Language Processing Spring 2016

Homework 1: N-gram Language Models

Abhishek Sinha

1 Methods

Both the backward bigram model and the bidirectional bigram model were implemented using the provided bigram model as a blackbox. For the backward bigram model, each of the sentences was reversed and then passed as input to the train and test functions of an instance of the bigram (forward) model. I did not implement the backward bigram model independently using a reverse iterator since the additional time cost of reversing the sentences was only around 2-3% (found by profiling) and the implementation was much more cleaner. For the bidirectional bigram model, I maintained an instance of the backward bigram model and the bigram (forward) bigram model. For every sentence, the log probabilities were computed for all the tokens (excluding start and end token) using both the forward and the backward model. The predictions from the models were averaged using equal weights of 0.5. Other weight combinations were also tried but did not lead to significant improvements in the word perplexities. Since the performance of the 2 individual models is similar, it makes sense to have equal weights.

2 Results

1. **Comparision of Word Perplexities**

   (a) **Training**

Dataset	Bigram	Backward Bigram	Bidirectional
atis	10.591	11.636	7.235
wsj	88.890	86.660	46.514
brown	113.359	110.782	61.469

   (b) **Test**

Dataset	Bigram	Backward Bigram	Bidirectional
atis	24.053	27.161	12.700
wsj	275.118	266.351	126.113
brown	310.667	299.685	167.487

2. **Comparison of Perplexities**

(a) **Training**

Dataset	Bigram	Backward Bigram
atis	9.043	9.013
wsj	74.268	74.268
brown	93.519	93.509

(b) **Test**

Dataset	Bigram	Backward Bigram
atis	19.341	19.364
wsj	219.715	219.520
brown	231.302	231.206

3 Discussion of Results

The perplexity results are found to be almost identical for the bigram model (forward) and the backward bigram model. In terms of word perplexities, bigram is found to be slightly better on the smallest (atis) dataset while backward bigram is found to be better on the wsj and the brown datasets though the difference is less than 10 %. Thus the data does not show that any of 2 models is clearly better than the other. The main reason the 2 models give similar perplexity values is that they both compute reasonable approximations of the same actual probability. The actual probability is the one we would get by computing the probability of the sentence precisely using the chain rule of probability without making any Markovian simplifications.

For the bidirectional model, the word perplexity becomes almost half of the value obtained by taking only the individual models. This is because of the ensemble effect. Both the forward and the backward models are similar strength models but provide complementary information. For example, sometimes the words on the right of a given word may provide more information about it whereas sometimes the word to the left may be more informative. The backward model would work well in the first case whereas the forward model would work in the latter. But the bidirectional model, by averaging the 2 models, would work well in both the scenarios.

Another thing we observe from the results is that in all the 3 models, the word perplexity on the test data is around 2-3 times the perplexity on the training data. This means that despite the fact the we smoothen the bigram model to some extent by interpolating it with the unigram model, there still does seem to be considerable overfitting.