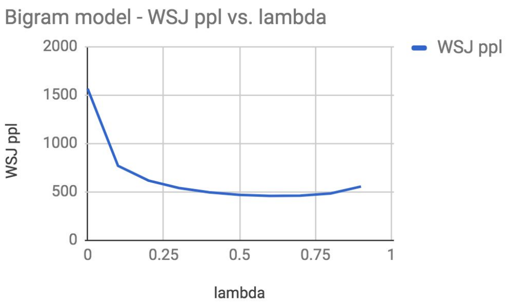
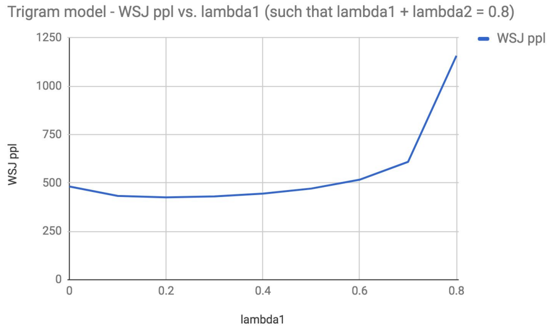
Homework 1 Solutions

Ankur Parikh

This is an example solution to homework 1. You may have gotten a bit different results if you chose to do slightly different experiments. However, for questions 1 and 3 there were clear explanations you should have given since the trend was clear.

1.1: Effect of interpolation weights lambda, lambda1, lambda2

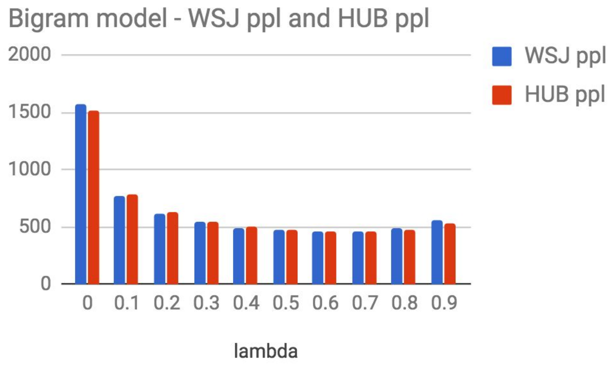




Consider the bigram model shown in the figure above. lambda=0 corresponds to the unigram only model which is a impoverished model, but generalizes well because it has less data sparsity issues. Lambda=1 corresponds to the bigram model which is a much more expressive model, but overfits due to data sparsity.

Interpolating the models balances these two effects. As shown in the graph above, at first increasing lambda decreases the test perplexity, but after a certain point a too large lambda overfits and the test perplexity increases. I found that lambda=0.6 gave the best WSJ perplexity. A similar effect is observed for the trigram model, where in the figure above, I fixed lambda1 + lambda2 = 0.8 and varied them. lambda=0.2, lambda2=0.6 gave me the optimal WSJ perplexity.

1.2: Relationship between WSJ and HUB performance



Interestingly, the perplexity for the unigram model is lower for HUB than for WSJ. This could be due to the lack of OOV words in HUB, or maybe that HUB is just a relatively easier dataset than WSJ.

Once bigrams/trigrams are added (see bigram graph to the left), the WSJ perplexity tends to be lower, which is expected since the training data is from WSJ so the train/test distributions are identical (as opposed to HUB which is a different domain). There are a few cases however, where the HUB perplexity tends to dip lower which is somewhat surprising. I observed a similar effect for the trigram case.

1.3: Relationship between HUB perplexity and HUB error rate

Bigram model		
lambda	**HUB ppl**	**HUB WER**
0	1519.12	0.091
0.1	783.35	0.083
0.2	628.03	0.0827
0.3	548	0.0781
0.4	500.4	0.0761
0.5	471.8	0.0767
0.6	457.3	0.073
0.7	**456.5**	0.0761
0.8	474.55	0.0707
0.9	536.2	**0.066**

While we generally hope that perplexity and error rate correlate they are not always guaranteed to. So it is not surprising to see that the HUB error rate does not always decrease with the perplexity. What is more interesting is that for the bigram model, the HUB error rate is lowest at lambda=0.9 well after the bigram model has started overfitting in terms of both WSJ perplexity and HUB perplexity.

One good plausible explanation for this is as follows:

Perplexity is an intrinsic evaluation measure that measures the likelihood of a sequence and is the metric that the model is trained to optimize. In particular, perplexity is particularly sensitive to rare sequences (i.e. perplexity will be infinity if a probability of a particular sequence is zero) and so tends to prefer models that are better at rare sequences. An intuitive way of seeing this is to visualize the log(x) function which rapidly goes to negative infinity as x->0. Thus perplexity favors more smoothing (i.e. lower lambda).

HUB error rate on the other hand is less sensitive to rare sequences since if a model cannot recognize a particular bigram it will most likely only add a couple extra insertions/deletions/substitutions to the computation of the word error rate.