Lab2 Report

fmonteroperez1 - aca15fm

For this lab our task was to implement three different language models to identify the missing word in a given sentence, one using unigrams, one using bigrams and one using bigrams and laplace smoothing.

1 Unigram Language Model

The unigram language model predicted correctly 5/10 sentences in the file (figure 2), returning only 0 values for one of the sentences (figure 1). This can be considered as an expected result because of how the language model calculates the probability of sentences by multiplying the probability of each word, thus the model will choose the most popular word.

This language model can be improved further by taking into context the previous words in the sentence as well as the beginning and end of a sentence.

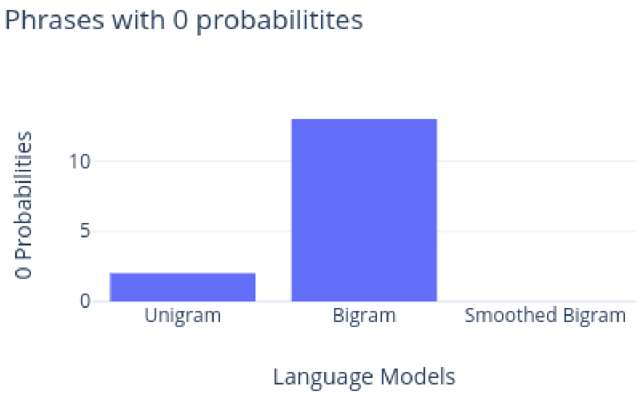

Figure 1

2 Bigram Languange Model

This model based on the Markov assumption assumes that the probability of a word is only based on the previous one, thus it is a bigram. In order for this to happen we remove the punctuation signs from the training data and from the test sentences, then the start and end tokens where added to the sentences.

Using this model we make the counts for words more scarce and can see an improvement over the unigram model as this one correctly predicts 6/10 words (figure 2). Although an improvement over the previous model, this model's flaw is that it returns 0 as the probability for phrases more frequently than the previous model (figure 1). This is because we might not have seen a given combination of words thus its probability is 0.

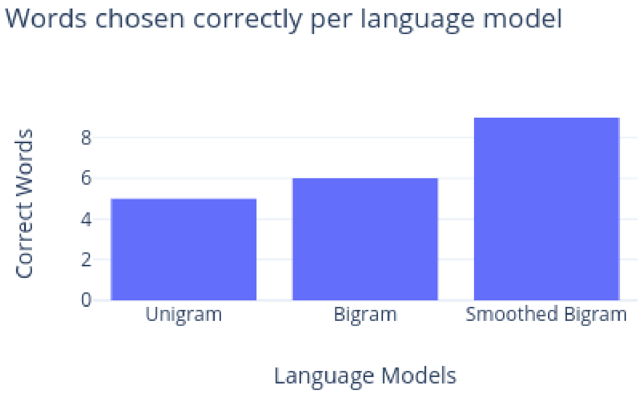

Figure 2

3 Smoothed Bigram Languange Model

To reduce the impact of an unseen word on the prediction of the model we apply smoothing to the classifier. Smoothing pretends like we've seen every word at least once and avoids returning 0 probabilities for unseen words (figure 1). This alone should present an increase in accuracy over the previous language.

A value of 0.5 was used for the smoothing to not alter the distribution as much as 1-up smoothing.

This model predicts correctly 9/10 words (figure 2). This is the best out of all the language models, basing the probability of a word on the previous one increases the chance of selecting the correct words by comparing using its chance of occurring in combination with the previous and the next words.