

CS5012 - P1
(Small) Language Models
200007413

1 Implementation

1.1 HMM Model

1.2 Viterbi Algorithm

1.3 Accuracy Calculation

1.4 Perplexity Calculation

1.5 Bigram Model

2 Results

Lang	HMM Accuracy	HMM Perplexity	Bigram Perplexity
en	0.98252	95.00279	13.99254
orv	0.82102	8045.24735	146.68261
tr	0.97091	428.37705	21.10778

Table 1: Accuracy of the HMM model, with perplexity values for the HMM and Bigram models using the English, Old East Slavic, and Turkish test corpuses. Given to 5 decimal places.

3 Reflection

The HMM model had high accuracy for English and Turkish test sets. This is consistent with the low perplexity observed on these test sets. The model had a lower accuracy on the Old East Slavic test set, which is consistent with the higher perplexity observed. The dramatic increase in perplexity on the Old East Slavic test set is likely due to the

The HMM model had high accuracy on the English and Turkish test sets. A significant performance drop was observed on the Old East Slavic test set. This could be due to the small size of the training set for Old East Slavic. This may be due to the relatively small training set for Old East Slavic. The model may not have been able to learn the language model effectively due to the small size of the training set. The model may have overfitted the training data, leading to poor generalization on the test set.

The HMM model additionally had low perplexity on the English and Turkish test sets - performing best with the English test set. The model had a higher perplexity on the Old East Slavic test set, which is consistent with the lower accuracy observed on this test set.