N-Gram Models Report

Yuqing Qiao
Khoury College of Computer Sciences
Northeastern University
qiao.yuqi@northeastern.edu

1 Introduction

NGram models are foundational in statistical natural language processing and involve predicting the next item in a sequence based on the N-1 previous items. This technique is fundamental for tasks such as text prediction, auto-completion, and language modeling. The perplexity of NGram models is highly dependent on the chosen value of N, and the training corpus, which balances memory usage and context sensitivity.

2 Method

The method involves constructing and comparing NGram models with varying N values. Each model is trained on the same dataset to ensure comparability. The models are evaluated based on their perplexity scores, which measure how well a probability model predicts a sample.

3 Dataset and Preprocessing

The dataset consists of a collection of English text from the Reuters and Gutenberg NLTK library, a standard dataset in natural language processing. Text preprocessing steps include tokenization, converting text to lowercase, lemmatization, and removing punctuation and stopwords to reduce model complexity and focus on meaningful words.

4 Model Architecture

The NGram models were implemented using a simple count-based approach(frequency). For each model, N-grams were extracted and their frequencies in the corpus were recorded. These frequencies were then used to compute the conditional probabilities of each word given the preceding N-1 words, applying Laplace smoothing to handle the zero-probability issue for unseen N-grams.

Bigram: $$P(w_n \mid w_{n-1}) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$$

N-gram: $$P(w_n \mid w_{n-N+1}^{n-1}) = \frac{C(w_{n-N+1}^{n-1}w_n)}{C(w_{n-N+1}^{n-1})}$$

5 Results

N-Gram Models Report

Yuqing Qiao
Khoury College of Computer Sciences
Northeastern University
qiao.yuqi@northeastern.edu

The performance of NGram models varies with the value of N and the dataset, illustrating the trade-off between bias and variance. Unigram models generally underfit on complex datasets, while 10-gram models introduce more data sparsity. Predicting the effectiveness of intermediate NGrams (e.g., bigrams to n-grams) without testing is challenging. Thus, evaluating each model on a substantial, independent text sample is crucial, as it provides direct insights into their predictive accuracy, reflected by their perplexity scores.