Language Model Evaluation Report

Introduction

In this report, we will analyze the performance of four language models (LMs) on two different corpora, "Pride and Prejudice" and "Ulysses." We have used two different smoothing techniques, Good-Turing Smoothing and Linear Interpolation, along with tokenization to build these LMs. The primary evaluation metric for these LMs is perplexity.

Perplexity Scores

LM1: Tokenization + 3-gram LM + Good-Turing Smoothing

- Average Test Perplexity: 323.45
- Average Train Perplexity: 7.28

LM2: Tokenization + 3-gram LM + Linear Interpolation

- Average Test Perplexity: 418.57
- Average Train Perplexity: 24.13

LM3: Tokenization + 3-gram LM + Good-Turing Smoothing

- Average Test Perplexity: 562.76
- Average Train Perplexity: 22.69

LM4: Tokenization + 3-gram LM + Linear Interpolation

- Average Test Perplexity: 687.93
- Average Train Perplexity: 76.33

Analysis

General Trends

- It's evident that LM1, which utilizes Good-Turing Smoothing, outperforms LM2 in terms of test perplexity, indicating better language modeling capabilities for LM1.
- LM3 with Good-Turing Smoothing also outperforms LM4, which uses Linear Interpolation, in terms of test perplexity on the "Ulysses" corpus, suggesting that Good-Turing Smoothing may be a more suitable smoothing technique for this specific corpus.

Overfitting

- The average train perplexity for all models is significantly lower than the average test perplexity. This discrepancy suggests that all models may suffer from overfitting, as they

perform much better on the training data than on unseen test data. However, LM1 has the smallest gap between train and test perplexity, indicating it might generalize better.

Corpus Differences

- The "Ulysses" corpus generally results in higher perplexity scores compared to "Pride and Prejudice." This discrepancy may be due to the complexity and unique linguistic characteristics of "Ulysses," making it more challenging to model accurately.

Smoothing Techniques

- LM1 and LM3 both employ Good-Turing Smoothing, and they tend to have lower perplexity scores compared to LM2 and LM4, which use Linear Interpolation. This suggests that Good-Turing Smoothing may be more effective in capturing the underlying language structure in both corpora.

Conclusions

LM1 with Good-Turing Smoothing consistently performs better in terms of test perplexity compared to LM2 with Linear Interpolation. Therefore, for these specific corpora, Good-Turing Smoothing appears to be a more suitable smoothing technique.

The overfitting observed in all models indicates a need for more robust techniques to prevent overfitting and improve generalization.

The choice of corpus greatly impacts language model performance, with more complex and diverse corpora resulting in higher perplexity scores.

Further experimentation with different smoothing techniques, n-gram sizes, and tokenization methods may lead to improved language models tailored to specific text domains.

In summary, LM1 with Good-Turing Smoothing appears to be the best-performing model among the ones evaluated here, but further research and experimentation are necessary to build more robust and domain-specific language models.