**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Statistical Language Models Analysis**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Weston Hawes**

California State University, Fresno

March 12, 2023

**Abstract**

The report presents an analysis of a statistical language model trained with a large set of text to predict the probability of certain words or phrases occurring in a given context. The text is preprocessed by removing special characters and numbers and converting it to lowercase. The report explains the concepts of statistical N-gram models and confusion matrices and their application in natural language processing and machine learning. The report outlines the approach, including the dataset used for training, the code breakdown, and the results. The report also includes various plots to visualize the distribution of unique words and bigram occurrences. The project identified 69599 words in the text and generated a plot of the occurrence of each unique word and a plot of the frequency of each unique word. There were 7740 unique words and 40180 unique word pairs identified in this book. The most common word was 'the' with 3260 occurrences and a frequency of .047 and the most common word pair is ‘of the’ with 288 occurrences and a frequency of .0041. The report provides insights into the structure and meaning of natural language and demonstrates the effectiveness of statistical language models in natural language processing.

**1 Introduction**

The goal of the project is to predict the likelihood of certain words or phrases occurring in a given context. The text has been preprocessed by converting it to lowercase, removing special characters and numbers, and removing extra whitespaces. The project involves various text analysis tasks, such as counting the number of words in the preprocessed text, the number of unique words, and the occurrence of each word (uni-gram) in the preprocessed text. Additionally, the project analyzes the occurrence of each pair of words (bi-gram) in the preprocessed text and generates text based on the probability of the words. The project also generates various plots to visualize the distribution of unique words and bigram occurrences.

**2 Background Material**

Key concepts utilized in this project will be defined and linked to this project in this section.

**2.1 Statistical N-Gram Models**

Statistical N-gram Models are widely used in natural language processing and computational linguistics. The term "N-gram" refers to a sequence of N words or characters that occur together in a piece of text. In statistical N-gram modeling, the frequency of these N-gram sequences is analyzed in order to make predictions about the likelihood of certain words or phrases occurring in a given context. This approach is particularly useful for tasks such as language modeling, text classification, and machine translation. By analyzing the frequency and patterns of N-grams in large datasets, statistical N-gram models can provide valuable insights into the structure and meaning of natural language and can be used to develop highly accurate and effective natural language processing applications. Hence unigram and bigram models are just N-Gram models with 1 and 2 grams respectively.

**2.2 Confusion Matrices**

Confusion matrices are a popular tool employed in the domain of machine learning to assess the efficacy of a model by comparing its forecasts with actual outcomes. It is essentially a table that summarizes true positives, true negatives, false positives, and false negatives generated by the model's predictions. These values help compute various performance metrics like accuracy, precision, recall, and F1 score, which provide an insight into the model's efficiency. Confusion matrices are employed in numerous applications such as natural language processing, image recognition, and fraud detection.

**3 Approach**

**3.1 Dataset Used for Training**

The dataset that the statistical language model will be trained on is the book "A Room with a View" written by British author E.M. Forster and first published in 1908. The novel has a total word count of approximately 80,000 words. The version of the book that was used in this project is on Project Gutenberg at the following URL: <https://www.gutenberg.org/ebooks/2641>.

**3.1 Code Breakdown**

The MATLAB code used to produce this project performs various text analysis tasks on the text file "room\_with\_a\_view.txt", the name given to the text file storing the contents of the book. Here is a breakdown of the approach:

1. The text file is read and stored as a string using the function readTextToStr().
2. The text is preprocessed using preprocessText() function. The text is converted to lowercase, special characters and numbers are removed, and extra whitespaces are removed.
3. The number of words in the preprocessed text is counted using countNumWords() function.
4. The number of unique words in the preprocessed text is counted using countNumUniqueWords() function.
5. The number of unique words with a minimum length is counted using countNumUniqueWordsMin() function.
6. The occurrence of each word (uni-gram) in the preprocessed text is counted using countUnigram() function.
7. The occurrence of each pair of words (bi-gram) in the preprocessed text is counted using countBigram() function.
8. The generateText() function generates text based on the probability of the words. The function takes the word pairs and counts as input, and starts with a given word. It generates a sequence of words based on the probability of the next word given the previous word.
9. The generateTextImproved() function also generates text based on the probability of the words. In addition, it takes an input for the maximum word length and a list of end words. The function generates a sequence of words such that the generated text ends with one of the specified end words and the maximum length of any word in the generated text is less than or equal to the specified maximum length.
10. Various plots are generated to visualize the distribution of unique words and the bigram occurrences.

**4 Results**

After cleaning the string of punctuation and other non-letter characters, this project found that there are a total of 69599 words in the book.

4.1 Unigram

Figure 1 is a Plot of the occurrence of each unique word. Figure 2 shows the frequency of each unique word; if all frequencies are added up, we get the sum of 1. The most common word is ‘the’ with 3260 occurrences and a frequency of .047. This is followed by ‘and’ at 1872 occurrences and a frequency of .027. These words are thirdly followed by ‘to’ at 1740 occurrences and a frequency of .025. There were 7740 unique words identified in the book.

**4.2 Bigram**

Figure 3 is a Plot of the occurrence of each unique word pair. Figure 2 shows the frequency of each unique word pair. The most common word pair is ‘of the’ with 288 occurrences and a frequency of .0041. This is followed by ‘in the’ at 257 occurrences and a frequency of .0037. These words are thirdly followed by ‘miss bartlett’ at 186 occurrences and a frequency of .0026. There were 40180 unique word pairs identified in this book.

A picture containing text, screenshot, plot, diagram

Description automatically generated

Figure 1

A picture containing text, screenshot, display, plot

Description automatically generated

Figure 2

A picture containing text, screenshot, plot, parallel

Description automatically generated

Figure 3

A picture containing text, screenshot, diagram, plot

Description automatically generated

Figure 4

**5 Discussion**

To further improve this project, additional information can be used to train the n-gram model to help it perform more effectively. One of the limitations of the model is that it only considers the probability of the next word given the previous word or word pair, and does not take into account the overall context or meaning of the text. This means that the generated text may not always be coherent or make complete sense. If the eight parts of speech such as nouns, verbs, adverbs, etc. are taught to the algorithm, it could be used to know how to structure probable word pairs. A dictionary of synonyms could also be used to better predict the next word because this could extend the to the words(synonyms) that other text may choose to use, that the training text did not. A dictionary could be used to reduce the probability of words such as pronouns from a training text because pronouns are usually used in specific contexts. Another limitation of the model is that it only considers up to bigram models. While this is sufficient for many applications, higher order n-gram models may be needed to capture more complex patterns and relationships in the text. The improvement to address this limitation is to expand the model to consider higher order n-gram models. This would require more computational resources and larger datasets but may provide more accurate predictions and insights into the structure and meaning of natural language.

**6 Conclusion**

In conclusion, this project successfully implemented an n-gram model in MATLAB to predict the likelihood of certain words or phrases occurring in each context. The model was trained on the book "A Room with a View" by E.M. Forster, and various text analysis tasks were performed on the preprocessed text, such as counting the number of words and unique words, analyzing the occurrence of each word (uni-gram) and pair of words (bi-gram), and generating text based on the probability of words. Additionally, various plots were generated to visualize the distribution of unique words and bigram occurrences. The project highlights the usefulness of statistical N-gram models in natural language processing and computational linguistics and how they can be used to develop highly accurate and effective natural language processing applications.