Statistical language modeling

Given a sequence of N-1 words, an N-gram model redicts the most probable word that might follow this sequence.

It’s a probabilistic model that’s trained on a corpus of text. Such a model is useful in many Natural Language Processing applications including speech recognition, machine translation and predictive text input. An N-gram model is built by counting how often word sequences occur in corpus text and then estimating the probabilities. Since a simple N-gram model has limitations, improvements are often made through various techniques

The goal of this project is to generate text artificially using the probability of appearance of the different words.

MATLab implementation (turn all steps into functions, CURRENTLY NOT FUNCTIONALIZED)

1. Create a function to read the textfile containing the book (ok)

2. Create a function to get all the document in lower case and extract the words

Get document in lowercase (ok)

Extract words(ok)

3. Create a function to count the number of words (ok)

4.Create a function to get the number of unique words (ok)

5. Create a function to get the number of unique words where you specify as an option the minimum number of characters in the word (ok)

6. Create a function to count the occurrence of each word ( unigram) (ok)

7. Create a function to get a matrix to get the number of words, based on the previous word. (Bi-gram)

8. Create a function that generates text based on the probability of the words, using the previous word as prior information.

9. Create additional functions to improve the automatic text generation process, by adding constratint or other rules for the generation

10. Create a function to plot the distribution of the unique words

Report

Title + Abstract + description of methods (1 point)

Provide the information about the selected book and the information you obtained from it based on the implemented code (1 point)

Plot a graph showing the distribution of the words and analyze the results (2 points)

Explain how additional information from the text such as the meaning of the words, dictionary of synonyms, grammar (subject, verb, complement...)­­­­­ can improve the text generation function (1 point)