CMPT 413/713: Natural Language Processing **(Due: 09/29/21)**

Programming Homework #1

Group Members: Joshua Malmberg, Hewei Cao, Keenan Byun
Group Name: Mochi

1 Objective

The aim of this assignment is to segment Chinese sentences into the words composing them. Unlike English sentences in which words are separated by spaces, there is no explicit separation between words in Chinese sentences and the segmentation must be inferred from the meanings of the characters. We aimed to implement an efficient Chinese language segmentation algorithm and explore various language models to obtain a high performance.

2 Method

Initially, we implemented a baseline segmentation algorithm based on the pseudocode algorithm provided in the assignment description, using a unigram language model with Laplacian smoothing. We then adapted the baseline segmentation algorithm to use a bigram language model with Laplacian smoothing, providing minimal improvement on segmentation performance. In these early iterations of the segmentation algorithm, we observed that the segmentation algorithm was not robust against unknown words. Specifically, when unknown words are encountered, they were not added to the heap which resulted in the algorithm not reaching the end of the sentence before terminating. This issue was resolved by adding a second loop to ensure the end of the sentence is reached before the program terminates.

To improve performance, we implemented a third iteration of the segmentation algorithm using Linear Interpolation with both the unigram and bigram probabilities, tuning the hyperparameters to maximize the dev-score. Finally, we implemented some simple rules to better handle unknown words. Observing that when several unknown characters were found in sequence, they were often parts of a single unknown word, we combined all adjacent unknown words together.

3 Results

	Dev Score
Laplacian Unigram Segmentation	0.81
Laplacian Bigram Segmentation	0.89
Linear Interpolation Segmentation w/ Rules	0.90

The use of linear interpolation smoothing provided minimal increase in F-score over bigram segmentation with Laplacian smoothing. We implemented a version of the segmentation algorithm using backoff smoothing which did not provide performance improvements. Despite trying several different smoothing approaches and achieving significant improvement, the final iteration of our algorithm still is limited in its ability to handle words which do not occur in the vocabulary.

4 Contributions

Hewei Cao: Implemented baseline unigram segmentation algorithm and wrote function to tabulate bigram counts; **Keenan Byun:** Implemented JM smoothing and Laplace Smoothing for Unigram Segmenter, composed Jupyter Notebook detailing results; **Joshua Malmberg:** Implemented Bigram Segmentation w/ Laplacian smoothing, debugged issues in unigram segmentation algorithm, wrote/typeset PDF report detailing results.