CMPT 413/713: Natural Language Processing **(Due: 09/16/21)**

Programming Homework #0

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

1 Objective

In goal of this programming assignment is to improve the accuracy of a simple English word segmentation program. English word segmentation entails partitioning a string of characters into the sequence of English words which is most probable based on some language model. The assignments provides an implementation of a segmentation algorithm, a simple language model that scores the probability of individual words, and a training set for the language model. We aim to tune the segmentation and language model to achieve the optimal performance possible with the given dataset.

2 Method

The default solution produces several inaccuracies when segmenting the testing data; several strings were not segmented at all. Upon reviewing the 5 highest probability candidate segmentations for these strings, we observed that the correct segmentation was present but had a lower score than the result of not segmenting the string at all. In the default solution, the p-score for words not present in the corpus is assigned statically, regardless of the length of the word. When using the default model, if even if a single word in a segmentation is not present in the corpus, the p-score for the correct segmentation is guaranteed to be less than the p-score for the segmentation is one single word consisting of the entire string. Initially, we attempted to address this problem by setting the p-score for unknown words to 0. However, this was unsatisfactory as it meant that any segmentation that contained unknown words would have p-sore of 0. Instead, we adjusted the p-score for unknown words to depend on the length of the word. If w is an unknown word of length l, then $P(w) = \frac{1}{N^l}$. In addition, we made a small change to the segmentation algorithm to avoid separating digits in to separate words.

3 Results

	Dev Score
Default	0.82
Optimized	1.00

The changes we implemented in our improved segmentation algorithm were successful at tackling two of the major problems present in the default solution. First, the improved version can successfully segment strings containing words absent in the corpus. Second, the improved version does not segment numbers into separate words consisting of individual digits. However, the performance of the segmentation system is still not ideal. For example, the improved algorithm produces the incorrect segmentations "current rate sought to go down" and "...lies a small un regarded yellow sun". This is because that the language model is not designed to generalize words found in the corpus by adding common suffixes/prefixes.

4 Contributions

Hewei Cao: Implemented improved segmentation algorithm; **Keenan Byun:** Composed Jupyter Notebook detailing results; **Joshua Malmberg:** Wrote/typeset PDF report detailing results.