Son Pham

CSCI-5832 NLP

Assignment 1 Part 2

In this exercise, the goal was to determine how many words does the standard BERT vocabulary really knows. The dictionary is formatted so that each line consists of a single word that may or may not be real. Although the vocabulary contains 30522 types, many are not actually unique words. Non-real words contained symbols, numbers, and/or morphemes that were combined with prefixes and suffixes to create another "word", although cannot be considered as a unique word. The goal of looking for morphemes is to find the smallest meaning-bearing unit after removing the additional prefix and suffix word segments that were used to modify grammatical function of the word.

The first step was to use Unix commands to remove characters that created false words. First, any line with brackets at the beginning of the line were removed, with the remaining list sorted in alphabetical order. This left the total token size to be 29522. Second, any single character words were removed, assuming that no single character is considered a feasible word. This resulted in a total token size of 28526. Third, since the list contains several lines staring with "##" that makes them false words, they were removed, which updates the tokens to be at 22698. Since there were several lines that included numbers, any line that included them were removed. This decreased the total size to 21731.

After using the preliminary analysis using Unix, the next step was to run spell-check to remove any word that is not present in an English dictionary. Using a text file of 466k English words created by *dwyl*, the remaining corpus was looped through and checked against the dictionary to determine if they are accurate words. To ensure that the search does not look for words internal to a whole word, an underscore was added to both the dictionary and corpus to delineate between words and their boundaries. The corpus is then run through the dictionary to parse the results, which is 19677 tokens.

The penultimate step was to perform lemmatization to process through words with the same stems but different wordforms. This step was performed using the Natural Language ToolKit package in Python. The stemmerporter function was ran on the corpus to extract the stems for the words, where applicable. This step did not remove any additional tokens, as the stem words remained inside the corpus. The final step is thus to again use Unix to sort the corpus and return only the unique tokens in the corpus. In conclusion, the author's best estimate at determining how many words doe the standard BERT vocabulary really knows now stands at 12321 tokens, or about 40% of the original corpus.