Son Pham

CSCI-5832 NLP

Assignment 1 Part 1

To “know” a word is to understand the meaning of a word and to have used it at least once in a document or to have spoken it at least once. A word may also be considered known if being used before but cannot precisely remember the meaning. To be more technical in the meaning of words “word”, “how many”, and “know”, a more detailed explanation is discussed. Type, which is defined as an entry in a vocabulary. Token is defined as an instance of a type in the text. A corpus assessment would focus on finding the base, or headword, of a given word, i.e., eat is the headword for ate. This assessment also considered the stems and affixes of the words by using inflectional and derivational morphology, which were used to find the smallest meaning-bearing unit and the additional prefix and suffix word segments that are added to the stems to change the meaning or grammatical function of the word.

To find out how many words the author think he may know, two tests were performed. The first test was to utilize a set of previously written reports and documents and parse the corpus for unique words used. In this study, the corpus gathered are written reports from the authors senior year as an aerospace engineering undergraduate student. This set of documents are technical in nature, so it may not represent commonly used words, known words from reading books, and may have more words that are repeated very few times. Since the corpus was from a set of technical papers, the words may not reflect which words the author would use if were to write other types of documents, such as a fictional novel where the topic is non-technical in nature.

The second test was to run an online vocabulary test to determine how many words that may be known based on a list of questions asking whether a participant knew a word or not. This test, at best, gives a rough indication of how many words are known by the author. Another reason the estimated results may be inaccurate is due to the unknown methods of how the online vocabulary test is set up. Without knowing too much detail on how this online test determines what the user may “know”, it is impossible to determine the precision and accuracy of the results.

For the first test, Unix commands were used to parse and find vocabulary words within the corpus. This is a rough assessment process that was used sort, clean, and assess the list of words. The first step was to remove possible duplicate words due to upper- and lower-case differences. In the final corpus, single characters were removed of the list of unique vocabularies, thus were not counted. Typically, a word is considered a word if it is can be defined in a comprehensive English dictionary, but in this case, any string that is greater than one character is considered a word.

Duplicate words were also removed from the corpus. Typically, “how many” is determined by the uniqueness of the word, where a word is counted only once and words with the same base, but different suffixes/prefixes are not counted multiple times. For this study, algorithms were not added to determine inflectional and derivational morphology, so a word with the same base but different or additional prefixes or suffixes are all included in the results. After cleaning the corpus of upper- and lower-case differences, the word count totals 5057. The most common words are *the*, *of*, *and*, *to*, and *in*. If Heaps' Law was used on the given corpus, with k=10 and beta=0.7, the results are 3914 words. For a fun assessment, the Unix commands were used on this document, which returned 260 unique words.

The second test was ran using an online website (http://vocabulary.ugent.be/), which allows the user to select whether a word is known or unknown from a bag of words. Fake words are added to ensure the participants are honest when taking the test, and any fake word that the user select as known negatively deduct points from the performance metric. The results were 87% of English words, though as mentioned earlier, it is debatable what is defined as the "total" English word count.