Methods

The overall method includes several steps. First, I found the list of heteronyms contained in the sentences of the corpus. Then, I ignored those that does not contain any heteronyms. Then, I annotated the pronunciation of heteronyms. Then, I scored each sentence based on the criteria for ranking. Finally, I ranked the sentences in terms of the score.

In the process of finding heteronyms in a sentence, I used crawling to check the online dictionary *Dictionary.com.* I searched each word of the sentence on the website and collected IPA pronunciation and corresponding example sentences. Then, I could determine if the word can be used as a heteronym based on the number of IPA pronunciations and the number of meanings. The inflictions of each word are not considered, thus cases such as past tenses and plurals are ignored unless itself was another word. For example, the word 'moped'(a kind of a motorcycle) was not counted as a heteronym because the meaning as "the past tense of mope" was not a valid use case. I had to process IPA pronunciations because some of them contained other information than the pronunciation, such as stress, literary cases, and position relative to vowels and consonants. They were cleaned up into a neat map of POS and IPA. The number of such IPA pronunciations was used to determine if a word is a heteronym.

In the pronunciation annotation step, the heteronyms in the sentence were annotated with the help of its POS and example sentences. For each heteronym, I identified which definition entry most likely corresponds to the word. I used the trigram of POS of the heteronym and surveyed the POS patterns in the example sentences. Such similarity was scored for each definition entry to find the correct one. By identifying the definition, the IPA could be identified. Some of the definition entries did not have any example sentences for me to utilize. In such cases, trigram could not be used, but the POS of the word was used instead.

In the scoring step, the occurrence and distribution of heteronyms in the sentence were used to form a value for the score. The number of occurrences of heteronyms is the strongest criteria, so it was positioned in the ten-thousands place. The number of different heteronyms was the second criteria, but the larger number meant lower rank. So it was positioned in the hundreds place, but subtracted from 100. The number of different POS was the last criteria, but the larger number meant lower rank. So it was positioned in the units place. The interval was arbitrarily decided as a 100 because I assumed it is not likely that a sentence contains a 100 heteronyms. Using the value formed through this process ranks the sentences by applying the three criteria in the correct order.

Discussion

The results have decent performance. It identifies heteronyms correctly, even the less widely known ones such as 'said' and 'house'. In the preliminary step of the process, words such as 'the' and 'a' are found as heteronyms. However, they are ignored because they differ in pronunciation only with respect to the position in the sentence or the presence of a stress. If they were included, the ranking would have been incorrect as well because almost all sentences contain at least one of them.

It was beneficial to ignore inflictions such as past tenses and plurals because dictionary does not contain them. They are redirected to the original word, which most likely has a different pronunciation. Thus, it was impossible to obtain the pronunciation of inflictions. Since there is no need to crawl pronunciation of the original word instead of the real word, it was more efficient to skip such cases.

The more challenging step was annotating the correct pronunciation. Some annotations have multiple versions inside the brackets because they are all correct ones, but they are still heteronyms since they can be pronounced differently in other cases. An example would be 'pass'. When it is used as a verb or a noun(the act of passing), it can be pronounced as both [pæs] and [pɑs]. However, when it is used as a noun to indicate the name of an American guitarist, it is pronounced as [pæs] only. I counted such case as a heteronym and believe it is reasonable.

In some cases, the POS of a word was enough to identify the correct pronunciation if it is pronounced differently for different POS. However, if it had different pronunciations for a same POS, the problem got tricky. To address this issue, I used the trigram of POS's around the heteronym and compared with those used in the examples. With a fair number of examples, it improved the accuracy slightly. Without the trigram POS method, the algorithm could not annotate 'wind' correctly in "The Georgia Legislature will wind up its 1961 session …", although not included in the csv file (33rd in the terminal output).

There were some cases I thought was an error in identifying heteronyms, but instead was a proof of robustness of the algorithm. Words such as 'periodic', 'won' and 'intern' were more of less widely-known heteronyms. I first thought they were not supposed to be identified as a heteronym, but in fact they were heteronyms: periodic ([ˌpɪər iˈɒd ɪk] / [ˌpɜr aɪˈɒd ɪk]), won ([wʌn] / [wɒn]) and intern ([ɪnˈtɜrn] / [ˈɪn tɜrn]).

The ranking of sentences is greatly affected by the performance of the heteronym identification. If the heteronyms are not correctly found, the ranking will be incorrect. Observing the output file, the ranking is accurate. The sentences follow the ranking criteria: those contain more occurrences are at the top, and those with various heteronyms tend to move to the bottom, etc.

Possible Improvements

The performance of pronunciation annotation would have been better if there were more example sentences in the online dictionary. There were a lot of cases in which a word is not commonly used that it did not have any example sentences. But to perform well in such edge-cases, corresponding example sentences would have been helpful.

There were some cases of verbs that had different meaning when used with or without the object. Devising a method to determine the presence of an object could help in heteronym identification.

Although I took care of common IPA forms, there were unique cases in which the pronunciation is too diverse and complex to extract IPAs, such as for words that come from other languages. If the pronunciation description was more general, the pronunciation annotation would have been cleaner.