L555 Final report
Jinhua Qian

Converting Sino-Korean words into Chinese characters

In this program, I was trying to detect Sino-Korean words written in Korean characters (Hangul) and convert them into Chinese characters (Hanza). Additionally, I planned to calculate the frequency of the appearance of Sino-Korean words in certain sentences. In order to realize this program, I applied four steps: segment >> tokenization >> conversion >> counting. (1) Segment. Each sentence will be separated onto its own line. (2) Tokenization. Split the text into tokens. Korean is a typical agglutinating language. There are a lot of affixes that follow or precede words. Thus, the tokenizer is required to separate the affixes from the words. I have tried to separate the particles from words by replacing particle with '\n particle' and split them according to the space. However, there are still a lot of ambiguities needed to be resolved. (3) Conversion. I found a Sino-Korean words dictionary[1] which includes both Chinese words and Korean words. I put one to one correspondent Chinese words and Korean words into a dictionary (my_dict), which could help us detect the Sino-Korean words and convert them into Hanza. (4) Counting. I counted the number of Sino-Korean appeared in a sentence and it was divided by the total tokens from that sentence. Additionally, I calculated the frequency of the Sino-Korean words in the entire text. The program can be run as follows:

```
$ cat final_text.txt | python3 final_segmenter.py | python3 final_tokenizer.py | python3 final_converter.py
```

I have tested 105 sentences and the frequency of the Sino-Korean words in the entire text was 36%. By increasing the accuracy of the tokenizer, I found that the rate of Hanza will increase. This program realized the goal as I excepted. However, further works on tokenization is required.

[1] https://github.com/dahlia/seonbi/blob/main/data/ko-kr-stdict.tsv