Take Home Exam 2110572 2020/2 NLP SYS
Question 2: Spelling Error Correction Method Specification

Thammakorn Kobkuachaiyapong

Summarize in 4 steps

1. Count incorrect-correct frequency.
2. Remove some error in labeling.
3. Choose a map that exceeds the threshold and maximum respect to it key.
4. Look through words in sentences and change if the word is in the mapping.

Create mapping dictionary

First count occurrence frequency of incorrect-correct pair from the train data by store incorrect as key and tuple of correct and count as value in mapping dictionary for each size of incorrect word (the mapping will store “ค่ะ” separate from “น่ะ”, “ค่ะ”). Then filter the mapping dictionary by selecting only maximum counts that exceed some threshold.

Mapping dictionary observation

After observing the mapping dictionary, it showed that there is not much mapper for incorrect word size that is higher than four (less than 5 pairs), so the mapping will consider only incorrect size three and below.

And if observe the mapping dictionary closely, there are some of the incorrect keys that the train data’s label is wrong such as (“อะ”) which is correct in real life usage but mark as incorrect then correct with (“อ่ะ”) which is incorrect, so the mapping dictionary should get rid of those miss-label.

Lastly, on one word dictionary there are many of incorrect usage but correctly spell while consider as misspelled including “ค่ะ”, “คะ”, “ใช่”, “ใช้”, “ไม่”, “ไหม” which will swap entire document and make the result wrong so we will handle this only in larger sizes which mean these keys would be removed from one word dictionary.

Sentence correction

For correction, apply mapping to the sentence, begin with three size incorrect mapping then two and one in order to correct the label error from larger sizes. By loop through the word in every sentence and if the word is in the incorrect dictionary then correct it with dictionary value.

Then, use train data as an evaluation to check if the mapping map wrong sequence or not and to consider which threshold is the best cutoff some keys that may be wronged by labeling and not occur much in order to not try to correct the corrected sentence.

For threshold selection, it is considered by average count of correction words(10) and how the size of the mapping changes with respect to the threshold starting from zero. This will help scope threshold range down to 0-5, 0-5 and 0-20 for size three, two and one respectively.