Waht kinda typoz do poeple mak?
COMP90049 Project 1 Report

1 Introduction

What kind of typos do people make is a wide question. Some existing solutions, like spelling checker or auto spelling corrector, help people make less misspell. On another hand, these solutions need to know how people make misspells, to improve their performance. For spelling correctors, a suitable algorithm could increase the correctness of their recognition and prediction on the words. This report implements a misspelling predictor based on the basic global edit distance algorithm to correct misspelling, and discuss the performance of those particular algorithms on those particular corpuses.

2 Data

There two set of corpuses used for this project: the wiki misspell and the birkbeck misspell. The wiki misspell list is from Wikipedia contributors (nd), there are 4453 tokens that have been identified as common errors made by Wikipedia editors. Corresponding with misspell list, wiki correct list is a list of the truly intended spellings. Another set of corpuses is from Roger Mitton (1980), which contains 34683 misspellings words, comprising the "Birkbeck spelling error corpus". It has a big difference with the wiki misspell that this list is a machine-readable transcription of hand written by schoolchildren, university students, and adult literacy student. In additional, the dictionary is from (dwyl, nd) which contains approximately 370K English entries. The implementation used it as a referenced correct words during comparison.

3 Methodology

Python is chosen as the language we implemented on, due to the wide range of packages, and there are several available algorithms able to complete the task.

3.1 Globla Edit Distance

The global edit distance(GED) is the main algorithm in the implementations. Package editdistance in Python performs well and fast which is based on the rule of Levenshtein distance. It calculates the "distance" between the misspelling word and the word from the dictionary by the numbers of character edit. In this implementation, these parameters are set to $(0, 1, 1, 1)$ for $match, insert, delete, replace$. The one with the lowest distance will be provide as the correct word. The naive version of GED is that only provides one prediction to the user. The predicted word has lowest distance with the word in dictionary. However it is not considered the situation when several words in dictionary have the same lowest distance with the misspelling word. The naive GED simply outputs the one appear earlier in the dictionary. The naive version of GED get improved that rather than storing the possible word as a single candidate, to store them in a list. The improved GED is able to provide multiple possible options in a row. However, multiple response is hard to get familiar by people. The second algorithm is used to refine the outputs.

3.2 N-gram

By testing the usabilities of the packages on Python, it is found that N-gram has the lowest efficiency. So it is used as a filter to refine the predictions from the GED. Since the less amount of responses from GED led the less amount of input for Ngram, then it could executes in a reasonable time. N-gram splits the tokens to substring with size of N. The number of substrings from two different tokens indicates the similarity of the two tokens. In this implementation, the parameter N is set to 2 and the distance between two tokens is calculated as

$$\frac{Number\ of\ same\ substrings}{Total\ number\ of\ unrepeated\ substrings}$$

4 Evaluation Metrics

Since the output number from different implementations are slightly different. There are different evaluation metrics for them. Throughout the report, accuracy, precision, and recall will be considered separately to evaluate each implementation system.

- *Accuracy*: For only single output system, the fraction of correct output that the system response to.

$$Accuracy = \frac{Num\ of\ correct\ predictions}{Total\ num\ of\ words}$$

- *Precision*: For multiple outputs system, the fraction of correct response among attempted responses.

$$Precision = \frac{Num\ of\ correct\ predictions}{Total\ num\ of\ predictions}$$

- *Recall*: For multiple outputs system, the fraction of word with a correct response.

$$Recall = \frac{Num\ of\ words\ with\ correct\ resps}{Total\ num\ of\ words}$$

5 Result

Overall, the accuracy from single output GED is as low as expected. The recall from the multi output GED perform good enough, but it has about 0.263 of precision. In GED plus Ngram refiner, the recall is only dropped about 10 percent, while its precision is raised up to 0.6 which means for over half of misspelling words it only makes single prediction. The result are summarised in Table 1.

Method	Accu / Recl	Prec
Single Output GED	0.549	0.549
Multi Outputs GED	0.797	0.263
GED + Ngram	0.717	0.600

Table 1: Result from Wiki misspell

In Birbeck misspell corcup, all the evaluation are lower than expect. The method with best performance, multiple output GED even not get half of accuracy and an extremely low precision. The result are summarised in Table 2.

Method	Accu / Recl	Prec
Multi Outputs GED	0.429	0.079
GED + Ngram	0.324	0.240

Table 2: Result from Birkbeck misspell

6 Analysis

Looking at the result from Birkbeck misspell in detail, the misspelling words are terribly wrong. It is possible that the misspells may also contains the machine-reader problem, rather than pure misspelling problem. Wrong recognition on the hand writing cause the low quality misspelling. So the usability of the corcup in this preoject have to be reconsidered. The further work will focus on the wiki misspell corcup only.

Misspell	Correct
afefew	relief
befertl	beautiful
aacock	helicopter

Table 3: Some phonetic examples

Back to the topic: what kind of typos do people make? Some interesting discovers is listing:

- People often only make a small mistyping when they typing which may not noticed by the writer. During fast typing, disordered letters, doubled characters, a missing letter and typing an adjacent key on the keyboard can happened frequently. Table 4 is an example.

Potential reasons	Misspell	Correct
Disorder	wiht	with
Double	libell	libel
Missing	indpendent	independent
Adjacent Key	tje	the

Table 4: Some potential reasons

The data also support the guess. In the wiki corpus, there are totally 4453 misspelling words, and in which there are 3345 words with a GED distance of 1. It means over three quarter of misspelling words have only one character less, more, or two adjacent characters get disorders.

- Phonetics is one of the misspelling reason. The similar sound of words leads people misspell. Table 5 is some misspelling due to phonetic problems.

The consequence due to phonetic problem behaves similar to the consequence due to

Misspell	Correct
careing	caring
diferrent	different
fourty	forty
littel	little

Table 5: Some phonetic examples

mistyping problem. But we still believe that phonetics is one of the reason. The list of words has same or similar pronunciation. It has high probability that people confuse with them. In this project, soundex is not implemented which could be improved in future.

7 Conclusions

This report implements a misspelling predictor, using both global edit distance and the N-gram algorithms. The final system can get about 70 percent of the accuracy and 60 percent of the precision. Also we discussed some potential reasons why people make misspelling. It is important to notice that since most of the results and further works are based on the wiki misspelling corcup, a restrict nature that the corcup is from typing editors limits the scope of discussions. These typos may not works on handwriting people.

References

dwyl. n.d. English words. https://github.com/dwyl/english-words.

Oxford Text Archive Roger Mitton. 1980. Birkbeck spelling error corpus. https://www.dcs.bbk.ac.uk/ ROGER/corpora.html.

Wikipedia contributors. n.d. Wikipedia:Lists of common misspellings. In *Wikipedia, The Free Encyclopedia.* https://en.wikipedia.org/w/index.php?title=Wikipedia:Lists_of_common_misspellings&oldid=813410985.