COMP90049 Assignment 1 Report: Cause of Typographical Errors

1. Introduction

This report will focus on identifying the causes of typographical errors. Several spelling matching tools, N Gram, Levenshtein Distance, sondex, as well as a method modified from Levenshtein Distance, will be implemented into spelling correction, based on evaluation metrics. Then, the results of the experiment will be demonstrated and analyzed to reveal the cause that may lead to the typographical errors in the dataset.

2. Dataset

The dataset in this report consists of two parts, Wiki_mispell and Birkbeck_misspell, which are quite different.

2.1. Wiki_misspell

Wiki_misspell is a list containing common mistakes made by Wikipedia users, which appearance at least once a year [1]. All these errors belong to typographical errors, which are created during the typing process [2]. On consequence, this list will be main dataset of this report.

2.2. Birkbeck_misspell

The entries in Birkbeck_misspell, which was published by Roger Mitton in 1980, differ from those in Wiki_misspell. It comprises misspellings created by schoolchildren, university students and adult literacy students among both native and foreign speakers [3]. Such kind of errors do not belong to typographical types.

3. Hypothesis

3.1 Literature Review

From Wikipedia, typographical errors are generally considered to be made in typing process. A typical instance can be “fat finger”, where users make over-input, less-input and wrong-input due to fat fingers or small key zones [2]. Moreover, in Baird’s article, he suggested that mis-input on touch screen can be solved by returning functions near the touching area [4]. Besides, Hitachi, a famous electronic company, actually uses such concept in their products [5].

4.1. N-Gram

Therefore, we can assume such mistakes to be caused by: users miss letters while typing (less-input), users write more letters than goal input (more-input) and users type wrong letter (wrong-input).

4. Methodology

This section will illustrate string matching tools implemented in the report and the way where they used in misspelling correction.

4.1. N-Gram

Specifically speaking, 2-Gram distance was implemented in this report. Meanwhile, the terminal “#” was used. For each entry in misspelling list, the distances to each word in the dictionary were calculated and the dictionary words with smallest distance were returned as candidates of possible input.

4.2. Phonetic

Soundex distance was implemented as a representation of Phonetic methods. Letters share similar “sound” were assigned with the same number and entries were transformed into new expresses based on that. Finally, the dictionary words with similar form to the misspelling entries were returned.

4.3. Levenshtein Distance

According to Michael Gilleland andMerriam Park [6], Levenshtein Distance is a special case of GED with parameter [0,1,1,1], which stands for matching, insertion, deletion and replacement. The larger Levenshtein Distance is, the less similar two words would be. Hence, the program of this report returned the words in dictionary with least Levenshtein Distance to the entries.

4.4. Modified Levenshtein Distance

In Leveshtein Distance, if one of the parameters was reduced, the results may more “prefer” to the action related to such parameter since Lvenshtein Distance selects actions with less

cost. In modified Levenshtein Distance, parameters were primarily set as [0,1,3,3] and the limit on the distance was set to be smaller than 3. That means the output only could be obtained by making 1 or 2 insertions. Then the parameters were set as [0,3,1,3], and so on. Later, the limit was changed into 4 and parameters became [0,1,4,4] and so on to test the erros within 3 steps.

5. Evaluation

This section will explain the evaluation metrics applied in the paper and analyze the results obtained in the previous section.

5.1. Evaluation Metrics

A single word in misspelling list may be related to several reponses. Hence two evaluation metrics were implemented in this report, recall and precision. Recall is defined as the fraction of entries with a correct attempt. Precision is the ratio of correct responses among all attempts.

5.2. Analysis

Firstly, corrections based on Soundex, Levenshtein and 2-Gram were conducted among Wiki_misspell and Birbeck_misspell (see table 1, table 2).

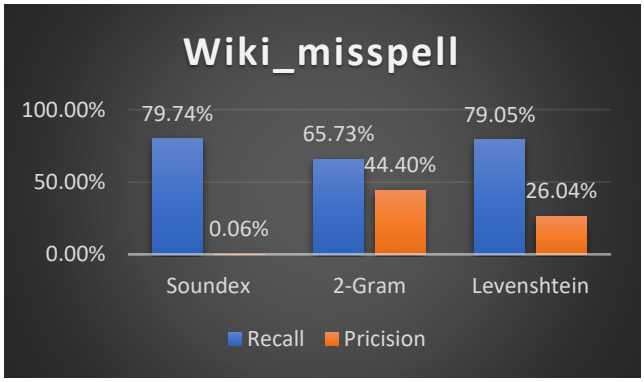

Table 1: Evaluation metrics in Wiki_misspell

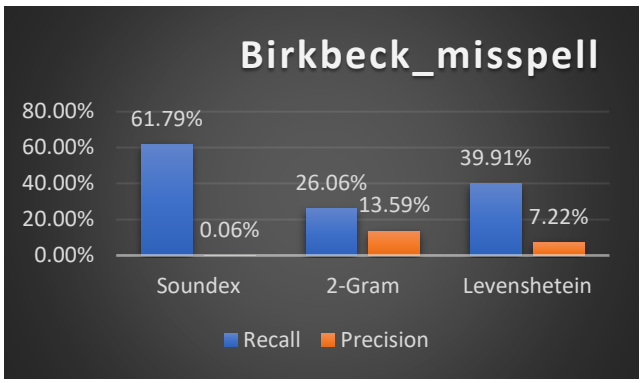

Table 2: Evaluation metrics in Birkbeck_misspell

The recall of Soundex performs well in both lists but the precisions were quite low, since it returns too many attempts. Therefore, it might not be a good measurement to the topic. The recall of 2-Gram in Wiki_misspell was higher than that in Bikrbeck_misspell because typolographical errors might only change a small part of the words, majority of alphabets were preserved. Those changes can be specified into insertion, deletion and replacement. The change of Levensstein Distance was the same as 2-Gram's.

Furthermore, as for Levenshtein Distance, parameters were changed to get the fraction of each possible cause in Wiki_misspell. The results can be viewed in table 3.

Parameter	Recall	Precision
[0,1,3,3]/2	35.86%	21.65%
[0,3,1,3]/2	19.04%	10.86%
[0,3,3,1]/2	35.84%	2.14%
[0,1,4,4]/3	36.04%	8.31%
[0,4,1,4]/3	19.32%	4.21
[0,4,4,1]/3	37.21%	0.23%

Table 3: Evaluation metrics of modified Levenshtein Distance in Wiki_misspell

Four parameters represent matching, insertion, deletion and replacement respectively. From the table, approximately 35% of errors can be corrected by insertion, which means 35% of the

errors could be created by less-input once or twice. Similarly, 35% of the errors were made by replacement once or twice. Besides,19.04% of the mistakes might be caused by more-input. Only about 10% of errors were created by the combination of these operations. When constraint on distance increased to 4, the proportion went to 36.04%, 37.21% and 19.32% respectively. It demonstrates that the majority of errors were caused within 2 steps.

6. Conclusion

Soundex, 2-Gram, Levenshtein and modified Levenshtein method were applied in this report. Words in dictionary with smallest distances to misspellings were returned as responses. Since Soundex returned too many attempts and 2-Gram did not identify specific typing action, modified Levenshtein distance seemed to be the most appropriate tool. Based on the results explained previously, the main causes of typographical errors can be identified as mis-input and less-input. More-input is a less-weighted reason but also should be focused on. Only around 10% of mistakes were created due to multiple kinds of operations. That could be observed in real life. Users are more likely to make single kind of mistakes in typing process, and the error number would not be large.

Furthermore, based on the statistics, different string-matching tools should be used in different datasets. Levenshtein, or edit distance, performs better in typographical errors. Because its parameters represent each mistake that users may make while they are typing. In addition, 2-Gram could be an alternative since its recall is acceptable and precision is even higher. Moreover, even though Soundex distance generates too many responses in both typographical errors and misspellings, which may be a waste of storage, it still provides a helpful concept in misspelling correction among the dataset such as Birjbeck_misspell. Because most written mistakes are related to “pronunciation”, which matches the concept of phonetic method. In conclusion, the implementation of a proper method may solve the problem more efficiently.

References

[1] 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

[2] Wikipedia contributors (2018, August 27) Typographical error. In Wikipedia: The Free Encyclopedia, https://en.wikipedia.org/wiki/Typographical_error

[3] Mitton, Roger (1980) Birkbeck spelling error corpus. In University of Oxford Text Archive, http://ota.ox.ac.uk/headers/0643.xml

[4] Baird, Randall B (2009), "Fast Typographical Error Correction for Touchscreen Keyboards.", U.S. Patent Application No. 12/506,564

[5] Kumai, Hiroyuki, et al (1991), "Method and apparatus for determining character and character mode for multi-lingual keyboard based on input characters.", U.S. Patent No. 5,634, pp.134

[6] Gilleland, Michael&&Park, Merriam (n.d.) Levenshtein Distance, in Three Flavors. In University of Pittsburgh, https://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm