Project 1: Waht kinda typoz do poeple mak?

Anonymous

1.Introduction

Typographical errors are some mistakes made in the typing or printing process.[1] Some mistakes may be made by human and some of them may be produced by machine. This report will discuss some reasons for typographical errors and raise some hypotheses about these. The report is based on the results of the experiments applying five spelling correction methods, which are implemented in the C programming language. And the datasets in these experiments comes from 3 text files.

- dict.txt: This is a text file including 370099 ordered English words, which is used as the dictionary for reference.

- wiki_misspell.txt: It has a list of 4453 misspelling words, which comes from the common mistakes in Wikipedia. [2]

- wiki_correct.txt: This is another list of 4453 words that corresponds misspelled tokens in wiki_misspell.txt. All these words have correct spelling.[3]

2.Hypotheses

The reason for missspelling errors may be multiple. Here are some hypotheses about what kind of errors will possibly happen. And each idea is based on my observation for the datasets.

2.1.Losing characters

When people are writing some long words or sentences, they probably miss one or two characters unconsciously. During my observation, some words obviously need one more character to read. For example, the string “confidentally”, on the line 968 in wiki_misspell.txt, should be modified as the word “confidentially”. This word miss the letter “i”. And similar mistakes may happen on other words.

2.2.Adding characters

Adding unwanted characters may be another sort of common mistake made by authors, especially double writing. For instance, it is obvious that the token “addressess” (line 127, wiki_misspell.txt) should delete the last letter “s”, because the “s” are double written incorrectly.

2.3.Transposition

Transposition means that some letters in a word are put on incorrect position. It will happen sometimes. And some letters even exchange the position. It can be seen on the line 481 in wiki_misspell.txt that the string “attaindre” are written incorrectly due to the letter “r” and “e”, while the correct form is “attainder”.

2.4.Phonetic errors

The phonetic errors are also likely to happen because of the authors’ accent and English skills. Many words have similar pronunciation that confuse some people. For example, when we speaks “habsbourg” (line 1905, wiki_misspell.txt) and “habsburg” (line 1905, wiki_correct.txt), they sound like same token. So phonetic reason may be another possible cause of typographical errors.

2.5.Replacement

Since we focus on the mistakes made by human, so it is possible that sometimes authors mistake a letter for another letter. For instance, the word “absence” (line 21, wiki_correct.txt) is written as “absense” (line 21, wiki_misspell.txt). The writer takes the letter “c” for letter “s”. So the possibility of replacement also exists.

3.Methods

The designed system will check each string in wiki_misspell.txt and predict the correct form for each string. Then it will compare the predictions to all words in wiki_correct.txt and calculate number of successful predictions. All the results will be printed in five new text files. This system is based on the following methods.

3.1.Global Edit Distance (GED)

As the edit distance is a classic method that

calculate the dissimilarity of two strings, the global edit distance focuses on the differences globally. It is not actually a "distance" but a figure that reflects how many weighed operations can help people to change a token to a correct word.[4] These operations are insertion, deletion, replacement and matching, and each operations is associated with a score. Best match is the dictionary entry with best aggregate score.

3.2.Local Edit Distance (LED)

Local edit distance is another kind of edit distance. As this method focuses on the best substring, it is more suitable for comparing two strings of very different lengths. Same as the global edit distance, it also deal with four kinds of weighted operations. And the best match is related to the best score of substring.

3.3.Longest Common Subsequence (LCS)

The method of the Longest common subsequence is to find the longest subsequence common to all sequences in a set of sequences. Unlike the substring that is handled by the local edit distance method, subsequences are not required to occupy consecutive positions within the original sequences.[5] The best match will be the two strings that have most common subsequence.

3.4.Hamming distance (Ham)

The Hamming distance measures the minimum number of substitutions required to change one string into the other, or the minimum number of errors that could have transformed one string into the other.[6] However, this sort of distance method can only handle with two strings of equal length. The minimum of the Hamming distance means the best match.

3.5.Soundex

Soundex is a kind of phonetic algorithm. It assign values to words and strings so that they can be compared for similarity of pronunciation. It classifies letters in different groups with different value. For example, all vowels are assigned to value 0, and all labials have the value 1.[7]

4.Evaluation

After implementing all the methods in the system, the correction results come out. All the hypotheses are confirmed or rejected. Here are some figures..

Methods	Total prediction	Correct prediction	Accuracy
GED	4453	2954	0.663
LED	4453	2341	0.526
LCS	4453	1603	0.360
Ham	4453	943	0.212
Soundex	4447	68	0.015

From this table we can see that the GED method has the best performance in correcting words. But other methods are also useful to support or reject my hypotheses.

For the errors of losing characters and adding chatacters, the method GED, LED and LCS can be effective. Because the GED and LED both take the operations, insertion and deletion, into account. And these two operation are suitable to modify these two kinds of errors. When comparing two string by characters, the LCS method always ignore the consecutiveness of each letter so that these two sorts of errors can be skipped, and only the longest subsequence will be treated. So the LCS will be effective. Here are some examples showing that problems of adding characters and losing characters actually exists in missspelling errors and the GED, LED and LCS can solve them effectively.

	aginst	inflamation	dicussed
GED	success	success	success
LED	success	success	success
LCS	success	success	success
Ham	failure	failure	failure
Soundex	failure	failure	failure

	abandonned	carribbean	speciallized
GED	success	success	success
LED	success	success	success
LCS	success	success	success
Ham	failure	failure	failure

Sound ex	failure	failure	failure

As for the errors of transposition and replacement, the Hamming distance often has a good performance. And the GED can also do good job, while the LED and LCS sometimes do not. This is for reason that the Hamming distance and the GED focus on the similarity of two strings globally, whereas the other two methods works locally. Here are example confirming the existence of the transposition and replacement. And the GED and Hamming distance can work well.

	alochol	opprotunity	virutal
GED	success	success	success
LED	failure	success	failure
LCS	failure	failure	failure
Ham	success	success	success
Sound ex	failure	failure	failure

	absail	devestated	profesion
GED	success	success	success
LED	failure	success	failure
LCS	failure	failure	failure
Ham	success	success	success
Sound ex	failure	failure	failure

When it comes to the hypothese of the phonetic errors, we need to admit that it is not the major reason for the typographical errors in our dataset. It is obvious that the accurcay of the Soundex method is very low. So it can be seen as coincidence that some words are corrected successfully by the Soundex algorithm.

5.Conclusions

Overall, I firstly raise some hypotheses about the causes of typographical errors through observation. Then I use five different correcting methods to support or reject my ideas. Each kind of errors have one or more ways to deal with. The Global edit distance method has best performance in correction, while other methods have their limitation but still be useful in some area. We can conclude that the error of losing characters, adding characters, transposition and replacement exists in our datasets. But the phonetic errors may not be the main cause of typographical errors.

References

[1]"Typographical Errors in Text and Figures", JAMA Psychiatry, vol. 74, no. 4, p. 424, 2017.

[2][3]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

[4]S. Konstantinidis, "Computing the edit distance of a regular language", Information and Computation, vol. 205, no. 9, pp. 1307-1316, 2007.

[5]D. Nath, "A Survey on Longest Common Subsequence", International Journal for Research in Applied Science and Engineering Technology, vol. 6, no. 4, pp. 4553-4557, 2018.

[6]K. Wong and M. Kim, "On private Hamming distance computation", The Journal of Supercomputing, vol. 69, no. 3, pp. 1123-1138, 2013.

[7]Census soundex. [Washington, D.C.?]: National Archives and Records Administration, 2003.