ktproject1

by Tianyu Chen

Submission date: 04-Sep-2018 06:33PM (UTC+1000)
Submission ID: 996629174
File name: KTPROJECT1.pdf (136.75K)
Word count: 1664
Character count: 9023

COMP90049 Project 1: What Kind Typos Do People Make?

1. Introduction

Typist or authors usually make spelling mistakes when they type the words and the reasons underlying those typographic errors are various. The aim of this report is to find the most accurate spelling correction methods to predict the intended (correct) word of each misspelling word made by Wikipedia editors. The main methodology implemented is implementing Global Edit Distance (GED) with different parameters. The underlying criteria of customising parameters is based on the hypotheses made on four main reasons for typographic errors: insertion, deletion, substitution and transposition. Finally, the analysis and evaluation will be given based on two metrics: precision and recall rate.

2. Dataset

There are three data sets used in this report: dictionary, wiki misspell and wiki correct [1]. The dictionary will be utilised as the dictionary for approximate string match, with 370099 words. Wiki misspell data contains 4453 misspell words made by Wikipedia editors. Wiki correct data contains 4453 correct words which are used as the benchmark to measure whether the spelling correction methods predict the correct words that Wikipedia editors intend to type.

3. Related Researches

The words spelling errors can be classified into several categories. One of the main errors is the typographic (typo) error. Typo error arises when the typist or author know the correct spelling of a word but misspell it because of figure slipping or pressing invalid key [2]. Kukich [3] concluded this kind of error as the non-words problem.

There are four main reasons underlying this problem, including deletion, insertion, substitution and transposition. People might miss certain letter (insertion), double type certain letter (deletion), substitute with another letter (substitution) or transpose adjacent letters (transposition) [4]. For example, when editors intend to type "accounting", they may mistakenly type it as "acounting" (missing words) "acconuting" (transpose), "accountting" (double type), or "acsounting" (substitute).

He also mentioned that the main work of solving this problem is done by exploring efficient string matching and comparison techniques for determining whether the input string is in the predefined acceptable work list or dictionary or not [2]. Some researchers find that through applying the weighted edit distance to spelling correction model, a lower score of one operation will help the system more precisely to retrieve the correct (intended) words from the dictionary if people's real typing mistakes coming from that operation [4].

4. Hypothesis

The hypothesis is that the four causes of typo error are evenly distributed. This hypothesis is based on the idea presented in the past literature as mentioned before.

Therefore:

H_0: *four types of mistakes are evenly distrubited*

H_1: *missing character is the most common cause of typo* error

Through implementing experiments, if the results can not indicate enough evidence to support, and then must be accepted.

5. Methodology

The global edit distance (GED) is used in approximate string matching to rank the word with the lowest GED as the "best match" of the intended word. The ranking criteria are based on the total costs given to operations (insertion, deletion, substitution, transposition) transferring the entire misspell words to the words in the dictionary [5].

There are three types of methods to test:

- Levenshtein Distance (LD)
- Damerau-Levenshtein Distance (DLD)
- Damerau-Levenshtein Distance with weighted parameters

The LD method scores three operations (insertion, deletion, substitution) equally as one point, while the DLD take the transposition operation into consideration. Therefore, the DLD could detect people's mistake of transposing adjacent letters.

DLD with weighted parameter method scores four operations as a half of point separately to seek which operation is the most common source of misspelling mistake. If those four scenarios all generate the similar amount of matching to correct (intended) words, the null hypothesis could not be rejected. However, if one result is significantly higher than others, the reason of typo error tends to be that operation with a lower score. In addition, after finding the most common typo error, the last three operation's score will be changed to 0.75 to see which one is the second common cause.

6. Evaluation Matrix

In order to assess different spelling correction systems' capability of correcting editors' spelling mistakes, precision rate and recall rate are used to serve as criteria.

Three spelling correction methods (mentioned before) will return different multiple predicted words and comparing them to correct (intended) words stored in wiki correct data. The precision rate is a fraction of correct prediction among all attempted prediction, which measures the efficiency of methods returning correct words. The recall rate is a proportion of words with correct prediction, which quantify the capability of methods returning correct words [5].

7. The Experiments

Three experiments will be conducted to test the hypothesis. The first one is a comparison between LD and DLD to test if the inclusion of transposition will improve prediction of typo error cause. The second one is separately customising four operations to 0.5 to test if four results are approximately equivalent or not. The last one is designed to find out if one operation is more common than others, which one is the second common operation.

8. Results

8.1 Comparison between LD and DLD

	LD	DLD
Precision	0.2766	0.3524
Recall	0.8394	0.9025
Avg time pw/s	1.5741	0.5425

Table 1: experiment 1 result

After including the transposition operation, the precision rate increase from 0.2766 to 0.3624 and the recall rate increases from 0.8394 to 0.9025. For example, in the case of "ahppen", LD method predicts 23 potential words and only one is correct, while DLD only make one correct because DLD considers transposing letter "a" and "h". Therefore, the overall prediction and recall rate is higher using DLD and surely transposition operation is one of the typo error reasons.

Misspell	LD	DLD	Correct
ahppen	'alpeen', 'alpen', 'appel', 'append', 'appet', 'arpen', 'aspen',	'happen'	'happen'

	'atopen', 'cuppen', 'happen', 'heppen', 'hippen', 'keppen', 'kippen', 'koppen', 'lippen', 'mappen', 'pippen', 'rappen', 'shapen', 'shippen', 'tappen', 'wippen'		

Table 2: experiment 1 example

8.2 Customise One Operation Parameter

DLD	Insert=0.5 Others=1	Delete=0.5 Others=1
Precision	0.5051	0.3642
Recall	0.8482	0.6816
Avg time pw/s	1.1490	1.2215

Table 3: experiment 2 result

DLD	Substitute =0.5 Others=1	Transpose =0.5 Others=1
Precision	0.4415	0.4415
Recall	0.8718	0.8718
Avg time pw/s	1.2013	1.2002

Table 4: experiment 2 result

From this results, it could be concluded that lower insertion's parameter has significantly improved the precision rate from 0.3524 to 0.5051, which indicates that insertion is the most common cause behind typo errors compared to other three operations. However, all of the recall rates have experienced a slight decrease because the higher precision rate illustrated decreased the number of the prediction made on each word, therefore, the total corrected response also declines. There is a tradeoff between higher precision rate and lower recall rate. Although under the cases of substitution (0.5) and transposition (0.5) have a higher recall rate than insertion scenario, the higher precision rate could compensate insertion's drawback on recall rate. Therefore, the precision rate is still the most important measurement to measure a spelling correction method's capability in this case.

8.3 Customise Two Operations' Parameters

DLD	No.1	No.2	No.3
	Insert = 0.5 Delete = 0.75 Others=1	Insertion =0.5 Substitute = 0.75 Others=1	Insert = 0.5 Transpose =0.75 Others=1
Precision	0.6031	0.5429	0.5963
Recall	0.7669	0.7539	0.8329
Avg time Pw/s	1.5373	2.0065	3.8355

Table 5: experiment 3 results

The result of this experiment indicates that the second common reason of typo error might be transposition or deletion because the precision rate has been improved from 0.5051 (only change insertion parameter to 0.5) to 0.6031 (deletion = 0.75) or 0.5963 (transpose =0.75). However, the recall rate of the no.1 experiment is much lower than the no.3, which could not be compensated by a small larger precision rate. Therefore, to solve Wikipedia editors' specific spelling problem, ranking transposition as the second common cause is more reasonable.

9. Discussion

The results from three experiment provide sufficient evidence to reject the null hypothesis that each operation equally contributes to typo errors.

Furthermore, the result gained from the second experiment proves that the insertion is the most possible cause of typo errors, hence people tend to omit letters when they are editing rather than other possible mistakes. Therefore, the alternative hypothesis that missing letter is the most common cause of typo error could be accepted.

Moreover, the result of the third experiment investigates the second common causes of typo error, which is the transposition.

10. Conclusion

To conclude from the previous analysis, there are four common causes of typographic error. Among those four causes, the most common reason that Wikipedia editors misspell words is the omission of letters and the second one is the transposition. It is also suggested that the spelling correction system should pay more attention to investigating those two kinds of mistakes to efficiently correct misspell words.

11. Reference

[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] Bhatti, S., Ismaili, I., Shaikh, A. Javid, W. (2012). Spelling Error trend and Patterns in Sindhi. *Journal of Emerging Trends in Computing and Information Science*, 3(10), 1435-1439

[3] Kukich, K. (1992). Techniques for Automatically Correcting Words in Text. *Acm Computing Surveys (CSUR)*, *24*(4), 377-439.

[4] Samuelsson, A. (2017). Spelling Correction in a Music Entity Search Engine by Learning from Historical Search Queries. Retrieve from: http://www.nada.kth.se/~ann/exjobb/axel_samuelsson.pdf

[5] Zobel, Justin and Philip Dart. (1996). Phonetic String Matching: Lessons from Information Retrieval. In *Proceedings of the Eighteenth International ACM SIGIR Conference on Research and Development in Information Retrieval*. Zu¨rich, Switzerland. pp. 166–173.

ktproject1

GRADEMARK REPORT

FINAL GRADE

96/100

GENERAL COMMENTS

Instructor

PAGE 1

PAGE 2

PAGE 3

Comment 1

Probably single parameter variation is enough to rank the considered typo types frequencies.

Comment 2

No examples

PAGE 4

RUBRIC: 2018S2 P1 RUBRIC		9.60 / 10
METHOD (30%)		10 / 10
10 (10)	• System design is admirably clear and unquestionably structured to provide testable hypotheses which will provide knowledge for the given problem	
8 OR 9 (9)	• Utilises relevant methodological strategies which are connected to logical hypotheses • System design is clear and reproducible, but some minor ideas are overlooked • Evaluation is systematic and logical	
8 OR 9 (8)	• Utilises relevant methodological strategies which are connected to logical hypotheses • System design is clear and reproducible, but some minor ideas are overlooked • Evaluation is systematic and logical	
7 (7)	• Utilises relevant methodological strategies which are connected to plausible hypotheses • Description of system design is missing some important idea, making the design questionable or dubious • Evaluation is logical but not systematic	
5 OR 6 (6)	• Utilises methodological strategies, but disconnected from corresponding hypotheses, or fundamentally limit the prospect of gaining knowledge • Description of system design lacks several crucial methodological components • Evaluation is attempted but illogical	
5 OR 6 (5)	• Utilises methodological strategies, but disconnected from corresponding hypotheses, or fundamentally limit the prospect of gaining knowledge • Description of system design lacks several crucial methodological components • Evaluation is attempted but illogical	
0 TO 4 (4)	• Methodological strategies are incomplete or absent	
0 TO 4 (3)	• Methodological strategies are incomplete or absent	
0 TO 4 (2)	• Methodological strategies are incomplete or absent	
0 TO 4 (1)	• Methodological strategies are incomplete or absent	
0 TO 4 (0)	• Methodological strategies are incomplete or absent	
CRIT ANALYSIS (40%)		9 / 10
10 (10)	• Clearly identifies the knowledge gained about the task • Argumentation is logical and incontrovertibly supported by evidence • Theoretical properties of methods are well-understood and linked to practical behaviour • Demonstrates a very high level of abstract thought • Admirably situated with respect to the academic community • Publishable with perhaps minor changes	

8 OR 9 (9)	• Clearly identifies the knowledge gained about the task • Argumentation is logical and thoroughly supported by evidence • Theoretical properties of methods are well-understood and linked to practical behaviour • Demonstrates a moderate level of abstract thought • Attempts to situate with respect to the academic community, but perhaps not clearly
8 OR 9 (8)	• Clearly identifies the knowledge gained about the task • Argumentation is logical and thoroughly supported by evidence • Theoretical properties of methods are well-understood and linked to practical behaviour • Demonstrates a moderate level of abstract thought • Attempts to situate with respect to the academic community, but perhaps not clearly
7 (7)	• Attempts to identify the knowledge gained about the task, but vague or unclear • Argumentation is logical, but evidence is lacking in some areas • Theoretical properties of methods are understood, but not clearly linked to practical behaviour • Demonstrates abstract thought, but extended analysis not always clear or successful • Little connection to the academic community
5 OR 6 (6)	• Knowledge gained about the task is fundamentally flawed or lacking • Argumentation is present but illogical, and evidence is inadequate or contradictory • Theoretical properties of methods are not in evidence • No signs of abstract thought and/or analysis • No connection to the academic community
5 OR 6 (5)	• Knowledge gained about the task is fundamentally flawed or lacking • Argumentation is present but illogical, and evidence is inadequate or contradictory • Theoretical properties of methods are not in evidence • No signs of abstract thought and/or analysis • No connection to the academic community
0 TO 4 (4)	• No indication of knowledge gained about the task • Argumentation is generally absent • Mostly data without corresponding analysis • Theoretical properties of methods are not in evidence • No connection to the academic community
0 TO 4 (3)	• No indication of knowledge gained about the task • Argumentation is generally absent • Mostly data without corresponding analysis • Theoretical properties of methods are not in evidence • No connection to the academic community
0 TO 4 (2)	• No indication of knowledge gained about the task • Argumentation is generally absent • Mostly data without corresponding analysis • Theoretical properties of methods are not in evidence • No connection to the academic community
0 TO 4 (1)	• No indication of knowledge gained about the task • Argumentation is generally absent • Mostly data without corresponding analysis • Theoretical properties of methods are not in evidence • No connection to the academic community
0 TO 4 (0)	• No indication of knowledge gained about the task • Argumentation is generally absent • Mostly data without corresponding analysis • Theoretical properties of methods are not in evidence • No connection to the academic community

REP QUALITY (30%) 10 / 10

10	• Ideas and arguments are cohesive, where the components of the report clearly

(10)	indicate how they relate to the whole • Report structure is logical and formal, in line with typical standards in academic writing • Generally clear and easy-to-follow • References are suitably synthesised and chosen discriminately with respect to the given problem • Adequately concise and meets word limits
8 OR 9 (9)	• Ideas and arguments are coherent, and generally the work fits together as a unit • Report structure is logical and formal, with small divergences from typical academic standards • Generally clear, with small disruptions in flow • References are suitably synthesised, but are too few or chosen indiscriminately • Adequately concise and meets word limits
8 OR 9 (8)	• Ideas and arguments are coherent, and generally the work fits together as a unit • Report structure is logical and formal, with small divergences from typical academic standards • Generally clear, with small disruptions in flow • References are suitably synthesised, but are too few or chosen indiscriminately • Adequately concise and meets word limits
7 (7)	• Ideas and arguments are mostly coherent, but do not come together in a unified way • Report structure is logical, but possibly informal or out-of-line with academic standards • Some unclear sections that do not detract from the overall work • References are present, but terse or disconnected from the problem at hand • Perhaps small divergences from the word limits
5 OR 6 (6)	• Ideas and arguments are notably incoherent • Report structure is flawed • Some unclear sections which detract from the overall work • References are disconnected or absent • Possibly way off the word limits
5 OR 6 (5)	• Ideas and arguments are notably incoherent • Report structure is flawed • Some unclear sections which detract from the overall work • References are disconnected or absent • Possibly way off the word limits
0 TO 4 (4)	• Ideas and arguments are missing or impossible to follow • Report has no structure or references • Not a formal report, even at a stretch
0 TO 4 (3)	• Ideas and arguments are missing or impossible to follow • Report has no structure or references • Not a formal report, even at a stretch
0 TO 4 (2)	• Ideas and arguments are missing or impossible to follow • Report has no structure or references • Not a formal report, even at a stretch
0 TO 4 (1)	• Ideas and arguments are missing or impossible to follow • Report has no structure or references • Not a formal report, even at a stretch
0 TO 4 (0)	• Ideas and arguments are missing or impossible to follow • Report has no structure or references • Not a formal report, even at a stretch