COMP90049 Project 1 Report

Utilizing Approximate String Matching for Typographical Error Correction

1 Introduction

Spelling corrector automatically checks and corrects wrong words, which occupies a significant role in information retrieval area, since it offers convenience and humanization (Elmi & Evens, 1998). The report based on the hypothesis that the major typographical error of dataset applied in experiments is dittography. The core purpose of the report is to verify the hypothesis by utilizing different approximate matching methods, including 3-grams and Global Edit Distance (GED) in word correction experiments. And based on the scores of evaluation metrics of these algorithms to analyze strengths and weaknesses of them, and then confirm or reject the hypothesis.

2 Dataset

The dataset applied in this report includes a misspelled word list, a corresponding correct word list, and a dictionary for running approximate string search methods, which is collected by Wikipedia contributors (https://en.wikipedia.org). The list of common misspellings is in alphabetical order, and contains major kinds of typographical errors such as homophones, repetitions, and grammar and miscellaneous. The dataset comprises of 4453 misspelled words, 4453 of correct words.

3 Methodology

Throughout the report, the following algorithms and terms will be utilized to conduct and evaluate the experiments.

3.1 3-Gram

An n-gram method is a contiguous sequence of n item from a given dataset. And 3-gram is a subset of n-gram method which is a contiguous sequence of n(n=3) letters and space from a given text (Islam & Inkpen, 2009). The goal of the method is to determine the "best" match of the target word from the dictionary. Moreover, 3-gram distance between string $s\big(G_n(s)\big)$ and string $t\big(G_n(t)\big)$ is

$$|G_n(s)| + |G_n(t)| - 2 * |G_n(s) \cap G_n(t)|$$

3.2 Global Edit Distance

global edit distance is one of the approximate string matching methods which can find a string that matches the one that has the largest similarity. Ukkonen (1985) states that GED has the same goal as the former algorithm, and the edit distance of this method is the minimum editing steps (match, replacement, insertion, and deletion) cost that transforms from the substitution of the target word.

4. Evaluation Metrics

For each word in the misspelling file, after running the algorithms mentioned above, the predictions for the target word have more than one candidate; therefore, the precision and recall evaluation metrics are more realistic and effective than the accuracy.

4. 1 Recall

In information retrieval and spelling checking area, for the system which has more than one matching candidates, the recall is the fraction of words with a correct spelling solution.

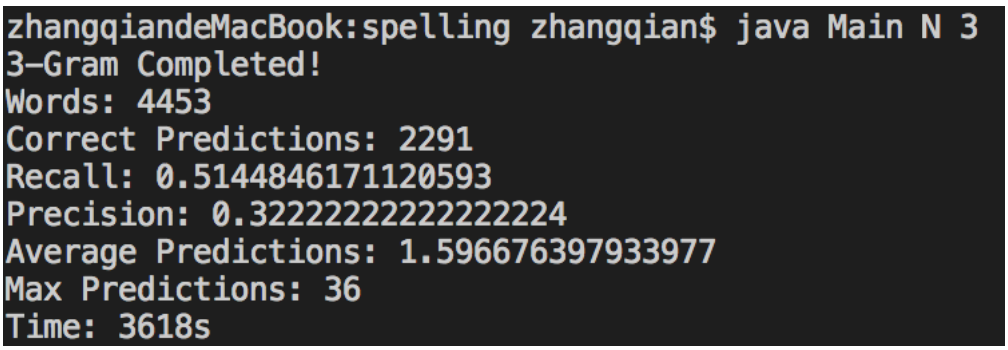
4.2 Precision

For the system which has more than one matching selection, the precision is the proportion of correct responses among all attempts. Both of them are therefore based on the understanding and measure of relevance.

5. 3-Gram Distance

5.1 Result

This method is based on the theory that if string A is the match of string B with several typographical errors, they will most likely contain at least one same substring of length 3 (Kondrak, 2005). The algorithm is written in java, with "#" filling the blanks of substrings, and returns all attempts that has minimum distance. The result is shown in Graph 1:



Graph 5.1: Result of 3-Gram

5.2 Analysis

From the result above, we can see that the recall of 3-gram method is 51.45%, and the corresponding precision is 32.22%.

The score of the 3-gram distance consists of three sections, the length of candidate string c, the length of the target string t, and 2 times of the count of same substrings. Since the length of the target string t is constant in each matching process, other two factors influence the results.

- On the one hand, when candidate strings take the advantage of the length, and substring numbers are the same, shorter strings attain lower distance scores. This means that when the typographical error is haplography, trigram is ineffective in dealing with spelling correction. Table 5.1 illustrates part of the examples:

Misspelling	Correct	Candidates	T/F
adequit	adequate	requit quit	✗
alchol	alcohol	chol	✗
irelevent	irrelevant	relevent	✗

Table 5.1 3-Gram in Checking Haplography

- On the other hand, when the typographical error is dittography, 3-gram algorithm has high-efficiency in matching the correct word with the wrong one. Table 5.2 shows part of the results:

Misspelling	Correct	Candidates	T/F
janurary	january	*january*	✓
marrtyred	martyred	*martyred*	✓
occour	occur	our *occur* scour succour	✓

Table 5.2 3-Gram in Checking Dittography

- In terms of metathesis and error letters, the error letter(s) in the word affect the score of the distance to a large degree, which cause ineffeciency in spelling correction. Table 5.3 explains part of the results:

Misspelling	Correct	Candidates	T/F
omre	more	mo re	✗
percieved	perceived	sieved peeved percid	✗
reciepents	recipients	repents	✗
amalgomated	amalgamated	amated	✗

Table 5.3 3-Gram in Checking Metathesis

6. Global Edit Distance

6.1 Result

The Levenshtein parameters utilized in the system is (m, i, d, r) = (1, -1, -1, -1), and the algorithm is written in java. The metric results of GED are summarized in Graph 6.1:

```
zhangqiandeMacBook:spelling zhangqian$ javac *.java
zhangqiandeMacBook:spelling zhangqian$ java Main G
Global Edit Distance Completed!
Words: 4453
Correct Predictions: 3355
Recall: 0.7534246575342466
Precision: 0.44272895223013986
Average Predictions: 1.7017740848865932
Max Predictions: 27
Time: 1737s
```

Graph 6.1: Result of 3-Gram

6.2 Analysis

From the result above, we can see that the recall for GED is 75.34%, and the precision is 44.27%. The score of GED is the minimum number to transform the candidate string to the target string by match, replacement, insertion, and deletion. The system assigned same weight for (r, i, d), which are (1, 1, 1).

- When the typographical error is haplography, according to the algorithm, the

major transformation step is deletion (distance+1), and GED is effective in correcting this kind of error. Table 6.1 illustrates part of the examples:

Misspelling	Correct	Candidates	T/F
archaology	archaeology	*archaeology*	✓
becouse	because	bedouse *because*	✓
caluclated	calculated	calyculated *calculated*	✓

Table 5.1 GED in Checking Haplography

- Similarly, when the typographical error is dittography, the major transformation step is insertion (distance + 1), and GED is effective in correcting this kind of error. Table 6.2 shows part of the examples:

Misspelling	Correct	Candidates	T/F
caluculated	calculated	calyculated *calculated*	✓
charactersistic	characteristic	*characteristic*	✓
constinually	continually	*continually*	✓

Table 6.2 GED in Checking Dittography

- In terms of metathesis, if there are two letters in the word are switched, the operations are (r,r), or (d,i), or (i,d), which doubles the distance score and decreases the efficiency. Table 6.3 explains part of the results:

Misspelling	Correct	Candidates	T/F
constatn	constant	constat constate constantan	✗
percieved	perceived	sieved peeved percid	✗
countires	countries	counties	✗

Table 6.3 GED in Checking Metathesis

- The last category is the same length but contains a wrong letter in the word, which can be corrected by replacement (distance + 1). Table 6.4 shows part of the examples:

Misspelling	Correct	Candidates	T/F
critisize	criticise	criticize	✗

cxan	cyan	oxan *cyan* can cran coan clan chan	✓
extention	extension	extentions	✗

Table 6.4 GED in Checking the Wrong Letter(s)

7 Verification

According to the analyze of 3-gram and GED, the Table 7.1 summarizes and compares the evaluation metrics:

	3-Gram	GED
Recall	51.45%	75.34%
Precision	32.22%	44.27%
Haplography	ineffective	high-effective
Dittography	high-effective	high-effective
Metathesis	ineffective	ineffective
Error letter	ineffective	effective

Table 7.1 3-Gram and GED

From the table, the obvious difference of the evaluation metrics for two algorithms is observed. In fact, the recall and precision of GED method is higher than 3-gram's. Moreover, based on the four types of typographical errors, the most conspicuous distinction in correction efficiency is haplography, rather than the dittography which is the hypothesis that made at the beginning of the report. Therefore, the hypothesis is rejected.

8. Conclusion

In summary, the report applies 3-gram and GED methods to verify that the hypothesis that the major typographical error of the dataset is dittography is false. GED achieves 37.40% higher in precision, and 46.38% higher in recall than 3-gram algorithm, and the major effective correction difference between these two methods is in haplography; therefore, the hypothesis cannot be accepted. Furthermore, the system

still has room for improvement in increasing effective, for instance, the dictionary should be up to date to include more correct words, and other methodology like Soundex (Holmes & McCabe, 2002), can be utilized to solve the issue, and verify the hypothesis.

Reference

Elmi, M. A., & Evens, M. (1998, August). Spelling correction using context. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1* (pp. 360-364). Association for Computational Linguistics.

Holmes, D., & McCabe, M. C. (2002, April). Improving precision and recall for soundex retrieval. In *Information Technology: Coding and Computing, 2002. Proceedings. International Conference on* (pp. 22-26). IEEE.

Islam, A., & Inkpen, D. (2009, August). Real-word spelling correction using Google Web IT 3-grams. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3* (pp. 1241-1249). Association for Computational Linguistics.

Kondrak, G. (2005, November). N-gram similarity and distance. In *International symposium on string processing and information retrieval* (pp. 115-126). Springer, Berlin, Heidelberg.

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

Ukkonen, E. (1985). Algorithms for approximate string matching. *Information and control*, 64(1-3), 100-118.