Waht kinda typoz do poeple mak?

Anonymous

1 Introduction

Automatic spelling correction is invented and first implemented by a computer scientist called Warren Teitelman in 20th century, and it is very common function that most editors have. The recent spelling correction systems are mainly based on Winnow algorithm (Machine Learning), probability scoring (Bayesian), Neural Networks, Levenshtein Edit Distance, etc. Some systems can even correct the misspelled tokens based on the context. [1]

The goal of this report is to determine what kind of typographical errors people make. In this report, one baseline algorithm and three advanced approximate string matching algorithms will be discussed and used to find different types of typographical errors.

1.1 Dataset

The dataset used in this report involves 4453 common misspelling errors made by the editors of Wikipedia[2], and their corresponding truly intended spellings.

Evaluation Metric	20%	100%
Precision	0.2700	0.2604
Recall	0.7899	0.7905

Table 1: Compare 20% and 100% of dataset

According to the evaluation metrics of the baseline algorithm shown in Table 1 (results are rounded to 4 decimal places), there is no much difference between 20% and 100% of the dataset. Therefore, only 20% random selected tokens of the dataset (890 tokens) will be used for the rest of the algorithms in the interest of speed.

All the tokens in the dataset and words in the dictionary are lower-cased, therefore, no preprocessing is required for the spelling corrections.

1.2 Evaluation Metrics

In this report, the algorithms applied will give multiple predictions for each misspelled word, therefore, precision and recall will be used as the evaluation metrics.

In order to compare between the baseline and advanced algorithms, precision and recall can be combined into a single evaluation metric called F-Score, which is the harmonic mean of precision and recall.[3]

$$F_1 = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$$

2 Hypothesis

There might be various kinds of typographical errors in the Wikipedia dataset. However, this paper will only be focusing on three possible types of typographical errors listed below, since finding all of the error types in the dataset will become a clustering task which is much more complex.

1. **Transposition** of two adjacent characters
2. **Substitution** of a truly intended character to a wrong character
3. **Duplication** of characters

These typographical error types could be tested with Damerau–Levenshtein Distance, Weighted–Levenshtein Distance, and N-Gram Distance respectively.

3 Method

3.1 Levenshtein Distance (LD)

The Levenshtein Distance (LD) gives the Global Edit Distance (GED) between the misspelled words and the dictionary entries with parameters *(m,i,d,r) = (0,1,1,1)*.

This method is used as a baseline method in this paper. The comparisons of the results between this baseline algorithm and the other algorithms could indicate the presence or absence

of the corresponding types of typographical errors in the dataset. If the advanced algorithms have better results than this baseline algorithm, it is probable that the advanced algorithms have corrected the corresponding errors stated in the hypotheses.

Evaluation Metric	LD
Precision	0.2700
Recall	0.7899
F-Score	0.4024

Table 2: Evaluation of Levenshtein Distance

3.2 Damerau–Levenshtein Distance (DLD)

Damerau–Levenshtein Distance is very similar to LD, but it also takes the transposition of two adjacent characters into account, and treat transposition as an operation with cost 1.

This additional character operation allows the DLD algorithm to give the misspelled tokens with transposition error a lower distance to truly intended word in the dictionary.

Evaluation Metric	DLD
Precision	0.3422
Recall	0.8551
F-Score	0.4888

Table 3: Evaluation of Damerau-Levenshtein Distance

3.3 Weighted–Levenshtein Distance (WLD)

Weighted-Levenshtein is another type of Levenshtein Edit Distance (or Global Edit Distance) which allows us to modify the cost of replacing a specific character with another character. This algorithm should be able to correct more "Substitution" errors in the dataset if the hypothesis is true.

3.3.1 Parameters

By analysing the Wikipedia dataset, the most common substitutions are shown at Table 4.

3.3.2 Parameters

For the implementation, the replace cost of the top 5 frequent: *(a,e), (e,a), (i,e), (e,i), (a,i)*, are setted to 0.5, while the cost of others are remained at 1.

Wrong Char	True Char	Frequency
a	e	231
e	a	185
i	e	129
e	i	127
a	i	125
e	o	83
i	a	73

Table 4: Substitution frequency in Wikipedia dataset (show frequency greter than 70 only)

Evaluation Metric	WLD
Precision	0.3168
Recall	0.7618
F-Score	0.4475

Table 5: Evaluation of Weighted-Levenshtein Distance

3.4 N-Gram

The N-Gram distance is also known as Q-Gram distance. The distance between n-grams of string s and t can be calculated by the following equation:

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

where $G_n(s)$ and $G_n(t)$ are the n-grams of string s and t respectively.

3.4.1 Principles

The bigram (N=2) method could be useful for the spelling correction. For the hypothesis: "Duplication of characters", there could be two typical cases:

1. The misspelled token has a duplicated character that the truly intended word does not contain, for example:

   misspelled: caat
   bigrams: {#c, ca, aa, at, t#}
   correct: cat
   bigrams: {#c, ca, at, t#}
   distance = 1

2. The misspelled token ignored the duplicated character that the correct word contains, for example:

misspelled: mis
bigrams: {#m, mi, is, s#}
correct: miss
bigrams: {#m, mi, is, ss, s#}
distance = 1

For both cases above, the bigram distance between the misspelled token and the truly intended word is 1, which is the minimum distance between two different strings.

Therefore this algorithm can correct the two types of misspelled tokens demonstrated above.

Evaluation Metric	N-Gram
Precision	0.4815
Recall	0.6854
F-Score	0.5656

Table 6: Evaluation of N-Gram Distance

4 Discussion

4.1 With evaluations

By comparing the F-Score's of the advanced algorithms (*DLD, WLD, N-Gram*) with the baseline algorithm (*LD*). It seems that there is some evidence to suggest that the three hypotheses are true, since the F-Score's of the corresponding algorithms are all higher than baseline's.

Method	F-Score
LD	0.4024
DLD	0.4888
WLD	0.4475
N-Gram	0.5656

Table 7: Comparing LD with other methods

4.2 With examples

Besides, the hypotheses could also be supported by the examples in the Wikipedia dataset. A few examples for each type of error are shown below in Table 8.

5 Conclusion

In conclusion, there is strong evidence showing that the dataset might include three types of typographical error: Transposition, Substitution, and Duplication. There might also be many other kinds of typographical errors such as: insertion, deletion, etc., which could be proven by changing the parameters of the GED and evaluating the results.

Misspelled	Correct	Type
theri wiht tlaking	their with talking	Transposition
chasr protocal consept	chase protocol concept	Substitution
possesing committment mispelled	possessing commitment misspelled	Duplication

Table 8: Transposition examples

References

[1] E. Mays, F. J. Damerau, and R. L. Mercer, "Context based spelling correction," *Inf. Process. Manage.*, vol. 27, no. 5, pp. 517–522, Sep. 1991, ISSN: 0306-4573. DOI: 10.1016/0306-4573(91)90066-U. [Online]. Available: http://dx.doi.org/10.1016/0306-4573(91)90066-U.

[2] Wikipedia contributors, "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, n.d.

[3] C. J. V. Rijsbergen, *Information Retrieval*. London: Butterworths, 1979.