Waht kinda typoz do poeple mak?

Anonymous

1 Introduction

Most of the recent spelling correction systems are based on Winnow algorithm (Machine Learning), probability scoring (Bayesian), Neural Networks, Levenshtein Edit Distance, etc.

The goal of this report is to determine what kind of typographical errors people make. In this report, one baseline algorithm and two advanced algorithms will be implemented for comparisons and evaluations.

1.1 Dataset

The dataset used in this report involves 4453 common misspelling errors made by the editors of Wikipedia(Wikipedia contributors, nd), 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 above (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.

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

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.(Rijsbergen, 1979)

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

2 Hypothesis

The possible types of typographical error could be:

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

This paper will only focus on the three types of typographical errors listed above, which 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 parameter (m,i,d,r) = (0,1,1,1).

This method is used as a baseline method. The comparisons of the results between this baseline algorithm and the other algorithms indicate the presence or absence of the corresponding types of typographical errors.

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 (DLD) 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 replace cost for a particular character to another character (default cost = 1).

3.3.1 Parameters

By analysing the dataset, the most common substitutions are shown at the table below.

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)

3.3.2 Implementation

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

3.4 N-Gram

The N-Gram distance is also known as Q-Gram distance, the distance between n-grams

Evaluation Metric	WLD
Precision	0.3168
Recall	0.7618
F-Score	0.4475

Table 5: Evaluation of Weighted-Levenshtein Distance

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.

The bigram (N=2) method could be useful for the spelling correction, according to the hypothesis: "Duplication of characters",

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

Table 6: Evaluation of Damerau-Levenshtein Distance

4 Discussion

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

Table 7: Comparing LD with other methods

5 Conclusion

Concluding text.

References

C. J. Van Rijsbergen. 1979. *Information Retrieval.* Butterworths, London.

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.