Waht kinda typoz do poeple mak?

Anonymous

1 Introduction

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 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.

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. **(Add more here.....)**

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

3.3.1 Parameters

3.3.2 Implementation

3.4 N-Gram

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

Table 4: Evaluation of Damerau-Levenshtein Distance

4 Discussion

Method	F-Score
LD	0.4024
DLD	0.4888
WLD	0
N-Gram	0.5656

Table 5: 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.