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, 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 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 Previous Work

2 Hypothesis

Text.

3 Method

3.1 Levenshtein Distance (Baseline)

Levenshtein distance between the misspelled words and the dictionary entries are used as the baseline method for comparison.

Evaluation Metric	Levenshtein
Precision	0.2604
Recall	0.7905
F-Score	0.3918

Table 2: Compare 30% and 100% of dataset

3.2 Global Edit Distance (GED)

3.2.1 Parameters

3.2.2 Implementation

3.3 N-Gram Distance

4 Evaluation

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. [citation]

5 Discussion

6 Conclusion

Concluding text.