Minimum Edit Distance Performance for Auto-spell Correction*

Miaomiao Zhang
University of Windsor
zhang3s2@uwindsor.ca

Kitty Duong
University of Windsor
duongy@uwindsor.ca

1 INTRODUCTION

Many tasks in Natural Language Processing are concerned with measuring how similar two strings are. The Minimum Edit Distance (MED) algorithm, also known as the Levenshtein distance or the edit distance, is a measure of the minimum number of editing operations needed to transform one string to another one. The Levenshtein Distance and the underlying ideas are widely used in areas like computer science, computer linguistics, and even bio-informatics, molecular biology, and DNA analysis. As for the dictionary in this experiment, we would use PyDictionary, which is a Dictionary Module for Python, using WordNet to get meanings, translations, synonyms, and Antonyms of words. Birkbeck would be the spelling error corpus, which includes the most common misspelled tokens and the correct spell in pairs.

2 MOTIVATION

The Minimum Edit Distance (MED) algorithm is a versatile tool that is widely applied in various fields where measuring the similarity or dissimilarity between sequences is crucial. This algorithm is designed to quantify the cost of transforming one sequence into another making it valuable in solving a wide range of computational problems. In Natural Language Processing and spell-checking applications, the MED algorithm is used to suggest corrections for misspelled words. It helps identify the minimum number of edits needed to convert a misspelled word into a correctly spelled one.

3 PROBLEM DEFINITION

Given WordNet as the dictionary $\mathcal{D}$ and Birkbeck as the corpus of misspelled tokens C, and a random token $t \in C$, the goal is to calculate the average success at k (s@k), k=1, 5, 10, for the minimum edit distance (MED) algorithm for all misspelled tokens in C.

3.1 Example

For instance, given "desing" from Birkbeck corpus, the top-5 most similar (least distant) tokens to 'desing' based on MED from WordNet are ['desi', 'design', 'designer', 'designate', 'despair']. Then, s@1 is 0 since the correct spell from Birkbeck is 'design' which is not happening at the first item. However, s@k is 1 when k is greater or equal to 2.

4 EXPERIMENT

4.1 Datasets

In this experiment, we used two datasets, the first dataset is the "Birkbeck" spelling error corpus. This contains a list of misspelled words along with their correct spellings. The second one is "Wordnet", which consists of a list of all English words against which we compare our misspelled words in this experiment.

4.2 Results

Based on the Minimum Edit Distance (MED) Algorithm, we divided this problem into two sub-problems, thus there are two parts of the final solution algorithm. In the first part, we intended to take the incorrect spellings and return the correct similar tokens from the WordNet dictionary.

```
Correct Words list ['visited', 'magnificent', 'opposite', 'gallery', 'splendid', 'purple']
Incorrect Words list ['visit', 'magnefision', 'aposit', 'galleroy', 'spenlid', 'purpal']
```

Figure 1: correct tokens output

The second part aims to return a list of only k(k = 1, 5, 10) numbers of tokens. For instance, for the token "gallery", the output is:

(1) s@k for k = 1: 0.15789473684210525
(2) s@k for k = 5: 0.0
(3) s@k for k = 10: 0.0

For the final output for the tokens from the Birkbeck Spelling corpus, we take k values from (1,5,10). The average probability of s@10 is the highest among the three types of values.

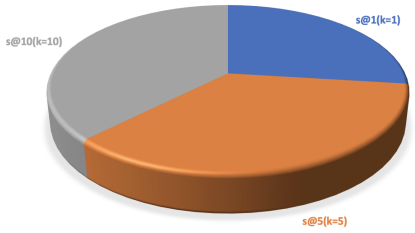

Figure 2: average s@k output

5 CONCLUSION AND FUTURE DIRECTION

The Minimum Edit Distance (MED) algorithm is a dynamic programming algorithm commonly used in natural language processing for spell-checking and related fields. It could also be used in machine learning techniques, especially deep learning with further advancements in algorithm efficiency in the future.

REFERENCES

https://en.wikipedia.org/wiki/WordNet
https://ota.bodleian.ox.ac.uk/repository/xmlui/handle/20.500.12024/0643

*https://github.com/duongy18418/NLP-Assignments

Conference'17, July 2017, Washington, DC, USA
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn