COMP90049 Project 1

# 1 Introduction

Approximate string search algorithms are broadly applied as spelling correction methods to resolve typographical errors in languages which are omnipresent. This report discusses several implementations of these methods along with the evaluation towards them according to related experiments.

The main datasets involved in this study include:

***wiki\_misspell.txt*** and ***wiki\_correct.txt***

Retrieved originally from Wikipedia: Lists of common misspellings [1], both consisting of 4453 entries occupying a single line respectively, in identical alphabetical sequences. They don’t contain phonetic errors. (Hereinafter referred to as Wiki Data)

***birkbeck\_misspell.txt*** and ***birkbeck\_correct.txt***

Retrieved originally from Birkbeck spelling error corpus in University of Oxford Text Archive [], containing 34683 entries. Unlike above, the Birkbeck data contain phonetic errors. (Hereinafter referred to as Birkbeck Data)

***dict.txt***

A large dictionary of the study comprising 370099 English words as single lines in alphabetical order, with each word existing only once. This file serves as the reference that is useful in the string search algorithms. (Hereinafter referred to as Dictionary)

# 2 Overview of Methods

Overtime, people have developed numerous methodologies regarding to approximate string search. Edit distance (EDIT, also known as Damerau-Levenshtein distance) counts the minimum number of edits required to transform one string into the other. It can be further classified as *global* (GED) and *Local* (LED) [2] The calculation of N-Gram Distance also aims to choose among the similarities between strings for the optimal result of correction. [3] There are also Phonetic String Matching algorithms which concentrate much on the sounds of the words, e.g. Soundex. [4]

According to the overlook of the typos in Wiki Data, the typographical errors can be classified into basically four types, which are:

* Insertion, with unnecessary character(s) inserted or duplicated. E.g. abandonned (abandoned)
* Omission, with necessary character(s) omitted. E.g. abilties (abilities)
* Transposition, with two characters swapped unexpectedly. E.g. abritrary (arbitrary)
* Substitution, with some character(s) wrongly replaced by other character(s). E.g. acadamy (academy)

In addition, phonetic errors are discovered in Birkbeck Data apart from the four types above:

* Phonetic errors, usually with words whose pronunciations are quite similar. E.g.

The experiments are designed based on the hypothetic errors above.

As the testing data are solely single words without a linguistic context, the entire prediction is based on the similarities between two words with no additional information. Therefore, the result is not as satisfactory as those in terms of paragraphs or articles.

Occasionally there are situations where multiple predictions with same distances. As there are no extra information provided by Dictionary (frequency, context, etc.), here we just randomly choose one out of the candidates.

# 3 Experiments and Results

In this study, we implement global edit distance and n-gram distance to perform spelling correction, using the following evaluation metrics to evaluate:

* Accuracy
* Precision
* Recall

Global edit distance:

N-gram distance:

In this project, bigram distance is applied, which means substrings are of length 2.

For each misspelt character in ***Wiki Data*,** we calculate its bigram distances with every word in the ***Dictionary***, and finally pick the one with minimum bigram distance.

However, there are situations where two or more words in the dictionary have same bigram distance to a misspelt word. In that way

Soundex:

Especially implemented to test Birkbeck Data.

• Recall: out of all of the correct Latin names, what proportion are found by the system? (This

is particularly meaningful if your system cannot predict a Latin name for some of the Persian

names.)

# 5 Critical Analysis

A discussion of how the results provide evidence supporting the presence/absence of theoretical

types of typographical errors;

# 6 Conclusion

The identification of typographical errors with the help of approximate matching methods

# 7 References

[1] 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

[] Mitton, Roger (1980) Birkbeck spelling error corpus. In University of Oxford Text Archive,

http://ota.ox.ac.uk/headers/0643.xml

[2] Levenshtein, Vladimir I. (1966). “Binary codes capable of correcting deletions, insertions, and reversals”. Soviet Physics Doklady 10 (8): 707710.

[3] Kondrak, Grzegorz (2005). “N-Gram Similarity and Distance”. In Proceedings of the 12th international conference on String Processing and Information Retrieval (SPIRE’05), pp. 115-126, Buenos Aires, Argentina.

[4] Zobel, Justin and Dart, Philip (1996). “Phonetic String Matching: Lessons from Information Retrieval”. In Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR’96), pp. 166-172, New York, USA.

B. John Oommen,

String alignment with substitution, insertion, deletion, squashing, and expansion operations,

Information Sciences,

Volume 83, Issues 1–2,

1995,

Pages 89-107,

ISSN 0020-0255,

https://doi.org/10.1016/0020-0255(94)00110-W.

(http://www.sciencedirect.com/science/article/pii/002002559400110W)