COMP90049 Project 1

# 1 Introduction

A short description of the problem and data set

Typographical errors are common in almost all languages

Spelling correlation is one of the main applications for Approximate String Search

In this study, we mainly make use of the typo list retrieved originally from Wikipedia: Lists of common misspellings. []

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

# 2 Existed studies

Overtime, numerous algorithms have been developed to solve the spelling correction and typographical errors. Among them, edit distance (EDIT, also known as Damerau-Levenshtein distance) counts the minimum number of edits required to transform one string into the other.

F.J. Damerau. 1964. A technique for computer detection and correction of spelling errors. Communications of the ACM 7, pages 171–176.

(Originally in Russian, published in English as:) Levenshtein, Vladimir I. (1966). “Binary codes capable of correcting deletions, insertions, and reversals”. Soviet Physics Doklady 10 (8): 707710.

* Neighbourhood Search
* longest common subsequence (LCS)

Vacl´av Chv´atal and David Sankoff. 1975. Longest common subsequences of two random sequences. *Journal of Applied Probability*, 12:306–315.

* N-Gram Distance

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.

There are also Phonetic String Matching algorithms which concentrate much on the sounds of the words, e.g. Soundex, Editex, Ipadist, etc.

However, we are taking advantage of orthographical methods due to the features of our data set.

Hypothesis:

* Non-word spelling errors

Detection: any word not in a dictionary is an error

Correction:

1. Generate candidates
2. Choose the best of the candidates:
   1. Shortest weighted edit distance
   2. Highest noisy channel probability

* Real-word Errors
  + Typographical errors
    - Three -> there
  + Cognitive Errors (homophones)
    - piece -> peace
    - too -> two
* For each word *w*, generate candidate set:
  + Find candidate words with similar ***pronunciations***
  + Find candidate words with similar ***spelling***
  + Include *w* in candidate set
* Choose best candidate
  + Noisy Channel
  + Classifier

3 Overview of methods

An overview of your approximate matching method(s)

focus on how they are applied to this task

4. Results

The results, in terms of the evaluation metric(s) and illustrative examples

Evaluation Metrics for Spelling Correlation:

* Accuracy
* Precision
* Recall

Examples

5 Discussion

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

Zobel, Justin and Philip Dart. (1996). Phonetic String Matching: Lessons from Information

Retrieval. In Proceedings of the Eighteenth International ACM SIGIR Conference on

Research and Development in Information Retrieval. Z¨urich, Switzerland. pp. 166–173.

Birkbeck

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

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