COMP90049 Project 1 Report

Analysis of Spelling Correction Methods on a Dataset from UrbanDictionary

1. Introduction

The aim of this report is to implement, compare and analyze the performance of spelling correction methods i.e. Levenshtein distance(LD), N-Gram Distance (ND) and Soundex (SD), on a peculiar dataset taken from UrbanDictionary, that have been automatically identified as being misspelled (Saphra and Lopez, 2016).

2. Dataset

The dataset contains a list of 716 headwords, one per line, that have been automatically identified as misspelled by Saphra and Lopez (2016). Hence, these misspelled words are to be corrected by using spelling correction methods and utilizing another dataset i.e. a dictionary dataset, consisting of 393954 words (tokens) from the English language, which is compiled from various sources. The resultant correct words generated after processing are compared with another dataset which comprises of a corresponding correct word for each of the 716 misspelled words respectively from which we analyze our methods for precision and recall. Hence, we use three datasets here i.e. the misspelled words dataset, the dictionary of English words dataset and the identified words with correct spellings dataset.

2.1 Dataset Analysis

There are anomalies identified in the dataset which effect the computation of the correct spelling word with respect to a misspelled word.

- 122 words which are in the correct spelling words dataset are not present in the Dictionary dataset, some of the words are asterix, brillant, uber etc.
- 175 words which are in the misspelled words dataset are also present in the Dictionary dataset, some of the words are aeroplane, centre, craftwork etc.
- 9 words which are in the misspelled dataset is actually the correct spellings. The words are “bratwurst, brillant, butthead, colour, corney, jeopardy, metre, separate and waterboarding”.

The anomalies or noise in the dataset should be cleaned for higher accuracy, precision and recall.

3. Evaluation Metrics

Throughout this paper, the following terms will be used to evaluate each spelling correction method:

3.1 Precision:

Precision deals with “How many of the predicted words from the dictionary are relevant (correct)?”. It is the ratio of the total number of correct prediction words made and the total number of predicted words.

3.2 Recall:

Recall deals with “How many relevant correct words are selected?”. Recall is the ratio of total number of correct predictions and the total number of absolute correct answers. Here, In the current scenario since every misspelled word has a single correct spelling, the total number of correct answers is just the total number of misspelled words or vice versa.

Note that Accuracy, cannot be the primary evaluation metric, because accuracy is used for systems that predict only one word, wherein precision and recall are used for spelling correction methods that predicts more than one resultant spelling as correct spellings for a misspelled word.

4. Methodology and Evaluation

4.1 Levenshtein distance (LD)

The Levenshtein distance as stated by Levenshtein, Vladimir I. (1966) is a measure of the similarity between the misspelled word and word in the dictionary. Levenshtein distance between each possible misspelled word and a word in the dictionary is calculated and the word for which the lowest distance is returned is predicted as the most likely to be the correct spelling. In the case of ties, i.e. when two or more number of words from the dictionary dataset have the same distance, then all the words are predicted to be the correct spelling.

- **Results:**
  While using Levenshtein Distance as the spelling correction method for the provided datasets as mentioned in section2, the following table can be formulated.

Precision	0.0458
Recall	0.3534
Total Correct Predictions	253
Avg. Predictions per word	7.72
Total Predictions	5528
Average Time for 716 words	520 secs

Table 1: Results for spelling correction using Levenshtein Distance

4.2 N-Gram Distance (ND)

The N-Gram distance as stated by Kondrak, Grzegorz (2005) is used to predict the correct word for each possible misspelled word, m(Gn(m)) and a word in the dictionary dataset, d (Gn(d)) by calculating the distance using the below formulae,

$|Gn(m)| + |Gn(d)| - 2\times|Gn(m)\cap Gn(d)|$.

After calculating, The N-Gram distance between each possible misspelled word and a word in the dictionary dataset, the word in the dictionary dataset for which the lowest distance (value) is returned is considered and predicted as the most likely to be the correct spelling.

- **Results:**

  While using N-Gram Distance as the spelling correction method for the provided datasets as mentioned in section2, the following table can be formulated.

Precision	0.0955
Recall	0.1956
Total Correct Predictions	140
Avg. Predictions per word	2.05
Total Predictions	1466
Average Time for 716 words	481 secs

Table 2: Results for spelling correction using N-Gram Distance

4.3 Soundex (SD)

Soundex is a phonetic algorithm for indexing words by sound as stated by Zobel, Justin and Dart, Philip (1996). The goal is for both misspelled and dictionary words to be encoded to the same representation so that they can be matched and if there is a match that dictionary word is predicted as the correct word and when more number of words have the same Soundex Code, then all those words are predicted.

- **Results:**

  While using Soundex as the spelling correction method for the provided datasets as mentioned in section 2, the following table can be formulated.

Precision	0.004
Recall	0.5908
Total Correct Predictions	423
Avg. Predictions per word	150.76
Total Predictions	107945
Average Time for 716 words	81 secs

Table 3: Results for spelling correction using Soundex

5. Improvement (or) Extension

A Potential improvement considered is the reduction of the number of tie breaking predicted dictionary words for the misspelled word with equal Levenshtein distance by processing this list of words for their respective N-Gram distance over and the words with the least distance are the final list of predicted correct words for the misspelled word. This increases precision over other discussed spelling correction methods.

- **Results:**

  While using Levenshtein distance followed by N-Gram Distance for predicted word reduction as the spelling correction method for the provided datasets as mentioned in section2, the following table can be formulated.

Precision	0.1129
Recall	0.25
Total Correct Predictions	179
Avg. Predictions per word	2.22
Total Predictions	1586
Average Time for 716 words	580 secs

Table 4: Results for spelling correction using Levenshtein followed by N-Gram Distance.

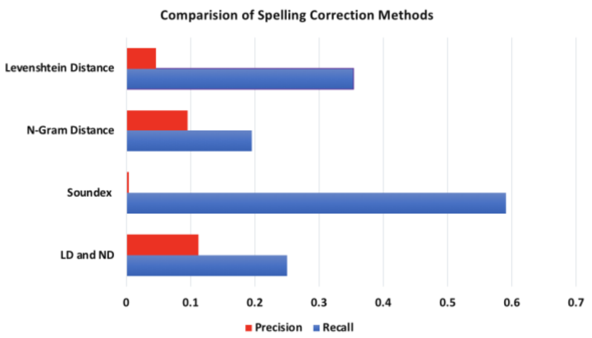
6. Evaluation:

The following table can be formulated, comparing the results and metrics after predicting the correct spellings for the 716 misspelled words.

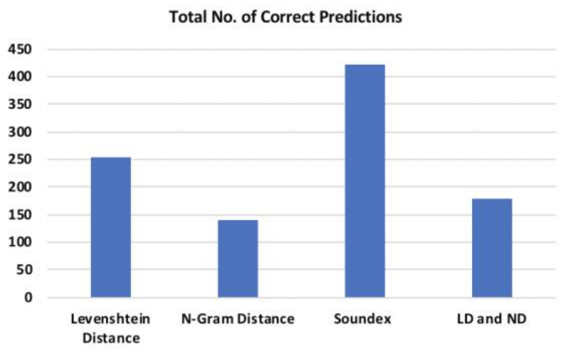
	LD	ND	SD	LD & ND
Precision	0.0458	0.0955	0.004	0.1129
Recall	0.3534	0.1956	0.5908	0.25
Avg. Time (Secs)	520	481	81	580
Total Correct Predictions	253	140	423	179
Total Predictions	5528	1466	107945	1586

Table 5: Comparing the Results of the spelling correcting methods discussed above.

It is clearly shown that the precision of Levenshtein distance followed by N-Gram Distance for the words is greater than that of the rest, whereas the Recall of Soundex is greatest than the rest, but the time taken to compute all the correct words for the misspelled words using Levenshtein distance followed by N-Grams is the greatest i.e. 580 seconds followed by Levenshtein Distance with 520 seconds and the least is Soundex with 81 sec. A bar graph can be plotted mapping the precision and recall results.



Graph1: Comparing the Precision and Recall of the spelling correcting methods discussed above.



Graph2: Comparing the total number of correct predictions for the methods discussed above.

It can also be observed that the total number of correct predictions is the greatest for Soundex followed by Levenshtein distance and a bar graph can be plotted between them as shown in Graph2.

7. Conclusions

This report established and presented spelling correction methods used to predict words with correct spelling for every misspelled word in the misspelled dataset. It is also possible to improve the precision by using various combinations of spelling correction methods together or one after the other as shown in Section 5 where the precision has increased to 0.1129 from 0.0458 and the next step to improve the precision would be by further editing the n-gram metric, or by finding a more effective weight and passing it as parameters to the edit distance.

References:

1) Naomi Saphra and Adam Lopez (2016) Evaluating Informal-Domain Word Representations with UrbanDictionary. In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, Berlin, Germany. pp. 94–98.
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.