SPELL CORRECTION OF ISOLATED WORDS: METHODOLGY TO IMPROVE THE SPELL CORRECTION

J.R.K.C. Jayakody*
Department of Computing and Information System, Faculty of Applied Sciences, Wayamba University of Sri Lanka
**Corresponding author (email: kithisrij@wyb.ac.lk)*

Introduction

Spell correction is an important functionality in information retrieval based systems and text processing applications. Further, more spell correction is used to correct the documents to be indexed and correcting the user queries to retrieve the right answers. Spell checkers are implemented in application programs [1] to flag words that are not spell correctly. It is an isolated word correction. All modern commercial spelling error detection and correction tools [2] work on word level with an appropriate dictionary. Usually edit distance, Frequency and n-grams are the main methodologies of spell correction based on individual word correction with a dictionary. This study is based on ongoing research to investigate the key methods of individual word correction and then to identify the best combination of methods to improve the spell corrections. Improvements of spell correction by combining different methods were not addressed appropriately in the literature.

Methodology

N-gram, Edit distance and the frequency of the word are the key methods which were combined to improve the word correction accuracy. N-gram is a contiguous of items from a given sequence of text. When a query was given; first enumerate all character sequences within a present. Jaccord coefficient was used as the methods to validate the accuracy in N-gram techniques. In edit distance error correction, given two words A1 and A2, then minimum number of operations to covert a misspelled word to a correct word is counted. That operation either be insert, delete or transform. Weighted edit distance is an extended form of edit distance error correction method in which weight would be assign for an operation based on the character involved[3].

Dataset

The correct word list which was used to build the index was taken from several public domain books from project Gutenberg and most frequent words from Wiktionary and the British national corpus. Testing documents were taken from Roger Mitton's spelling error corpus from the oxford text archive. Those corpuses were used to build the index file. Tree Map <String, List<String>> to store the index file. The research was conducted as illustrated in Figure 1. Natural Language processing preprocessing techniques [4] were used to extract, transform and load the data. Dataset file content was read and stringtokenizers [5] were used to separate the strings into tokens. Word tokenizer, Wordpunctokenizer, tree bank tokenizer and regextokenizer have been used to identify the most appropriate

tokenizer. Since Regexptokenizer has developed to token based on spaces of the sentences, it produced the less number of appropriate tokens to proceed with the subsequent steps of the process. Once the tokenization process was over it was used to generate the inverted index with the bigrams, frequency table with the value and key pairs. In addition to that existing edit distance function was used. Validation function was developed to test the accuracy of individual as well as combined spell correction methodologies with provided Dataset.

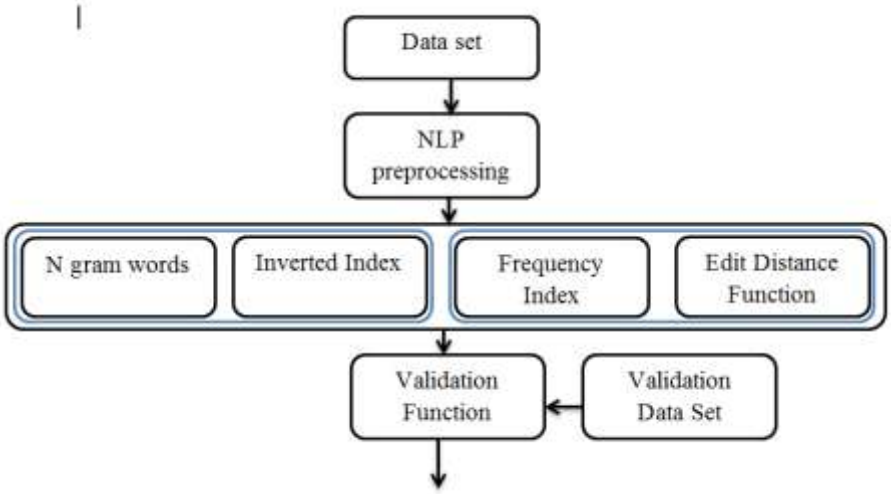

Figure 1: Process of testing spell correction methodologies appropriateness

Results and discussion

Spell correction was tested with individual methods as well as by combining different methods. N-gram, edit distance and frequency correction was tested individually as illustrated in Table 1. Furthermore those methods were combined proportionally to correct the words.

Table 1: Spell correction accuracy for individual spell correction methodologies

Type	Accuracy
N-gram (Jaccord Coefficient)	68.0
Edit distance	67.0
Frequency	3.0

According to the final validation frequency was given the least accuracy and the n-gram with Jaccord coefficient was given the highest accuracy of 68%. In this phase n-gram was limited to bigram methods. Generic edit distance function was given 67% accuracy which is almost equivalent to N-gram methodology. In the next phase, all the methods were combined proportionally to improve the spell corrections.

N-gram, Frequency and edit distance methodologies were used proportionately to improve the spell correction accuracy. Frequency and edit distance were proportionately changed within 10-90% range. N-gram, Frequency [90%] and edit

distance [10%] were given 66% accuracy whereas N-gram with Frequency [10%] and edit distance was given [78%] accuracy.

Table 2: Spell correction accuracy by combining spell correction methodologies

Type		Accuracy
N-gram with Frequency[10%]	Edit Distance[90%]	78.0
N-gram with Frequency[20%]	Edit Distance[80%]	78.0
N-gram with Frequency[40%]	Edit Distance[60%]	76.0
N-gram with Frequency[60%]	Edit Distance[40%]	73.5
N-gram with Frequency[80%]	Edit Distance[20%]	72.0
N-gram with Frequency[90%]	Edit Distance[10%]	66.0

Conclusion and recommendations

N-gram with jaccard coefficient, Frequency and the edit distance methods were tested individually for accuracy. Then each method combined proportionally to test the word accuracy. According to the generated results with the tested corpus, frequency was not a good option for spell correction which is 3%. Bigram with jacard coefficient was given the highest frequency among methods which is 68%. Once the methods were combined, accuracy was improved up to 78% which is 10% improvements over the best individual method. The portion of edit distance and the frequency checking was changed appropriately to get the highest accuracy. N-gram with frequency (20%) and edit distance (80 %) was given the best accuracy among the combinations. Improvements of the edit distance afterwards would not make any impact of the accuracy.

While the research effort presented in this paper have been fruitful in achieving the objectives there are few notable future research extension to be mentioned. The bigram must be extended to trigram and quadram which can be suggested for extension of this work with future research.

References

[1] J.L. Peterson. "Computer programs for detecting and correcting spelling errors." *Communications of the ACM* 23, no. 12, 1980, pp. 676-687.

[2] R.C. Angell, E.F. George, P. Willett. "Automatic spelling correction using a trigram similarity measure." *Information Processing & Management* 19, no. 4, 1983, pp. 255-261.

[3] M. Mohri. "Edit-distance of weighted automata: General definitions and algorithms." *International Journal of Foundations of Computer Science* 14, no. 06, 2003, pp. 957-982.

[4] M. Srivastava, G. Rakhi, P. K. Mishra. "Preprocessing techniques in web usage mining: A survey." *International Journal of Computer Applications* 97, no. 18, 2014.

[5] Y. He, K. Mehmet. "A Comparison of 13 Tokenizers on MEDLINE." *Bethesda, MD: The Lister Hill National Center for Biomedical Communications*, 2006.