Bangla Spell Checking and Correction Using Edit Distance

Muhammad Ifte Khairul Islam
Daffodil International University
khairul3182@diu.edu.bd

Rahnuma Islam Meem
Daffodil International University
raahnuma1711@gmail.com

Faisal Bin Abul Kasem
Daffodil International University
mfaisalpasha@gmail.com

Aniruddha Rakshit
Daffodil International University
aniruddha.cse@diu.edu.bd

Md. Tarek Habib
Daffodil International University
md.tarekhabib@yahoo.com

***Abstract* - Automatic spelling correction for Bangla language is a very important thing as the modern world is almost completely dependent on digital devices where electronic keyboards are used. Bangla is a largely used language (3.05% of world population), worldwide securing the 7th most spoken language. But concerning the natural language processing it is still in its initial stage. Automatic spelling correction of Bangla language is just crossing its primitive stage. A few works have been done on automated Bangla spelling correction but those numbers are not still adequate. This research work is based on correct word prediction of typed Bangla sentences by using edit distance method and fractional accuracy. The results are promising. So, there is a high chance that the proposed model will help and have a great impact on automated Bangla spelling correction.**

Keywords - Bangla Spelling, Spelling Correction, Non-word error, Fractional Accuracy.

I. INTRODUCTION

Written form of a language is a very important factor in communication; whether it is handwritten or typed. When speaking it is not important to know proper spelling of a word and even without proper maintenance of grammar communications can be conducted, but when writing it is very important to have correct spelling. Spelling mistake is a common issue especially if it is a complex language such as Bangla. In aspect of most spoken language, Bangla stands 6th in the world [1]. It has complex word construction, so it is common to have spelling error while writing. It is not necessary that the reasons for spelling mistakes are only because of its complexity, it might also occur due to fast typing or lack of knowledge. As it is a common phenomenon to make spelling mistakes, automatic spelling correction processes are a must. There are lots of work on this issue but when considering Bangla there aren't many notable works. The main focus of this research work is checking the spelling of Bangla word following by the correction of it using fractional accuracy. Direct dictionary lookup and edit distance method will be used to check correctness and gain possible corrections of the word then the accuracy will be calculated. There are some works that used the edit distance method but, in this research, a much larger dictionary and new method to calculate accuracy are used which will give better results.

The remainder of the paper is structured as follows: a literature review is conducted in section II where previously published works that are similar to our work has been discussed. Section III is about the methodology of our work where the process of our work is described. Next, in the implementation section, results of our work is shown. Section V shows the comparison results with previous works. Finally section VI concludes the paper describing how this work is promising.

II. LITERATURE REVIEW

There are quite a number of works done on spelling mistake detection and correction. Spanish, English, Mandarin, Bangla etc. are some of the most spoken languages on earth, so it is normal to work on these languages but automated spell checker and correction is of such an importance that a number of works were also done on less spoken compared to the languages mentioned before or local languages such as Myanmar, Persian and Hindi [2], [3] and [4]. All over the world people are working on this topic not necessarily focusing on any specific language but also on language independent spell checker and corrections [5].

Languages such as English, Chinese and Arabic are largely spoken languages thus having more works focusing on them. Some notable works on English are done in [6], [7], [8], [9] and [10] using different methods such as edit distance, tree traversal etc. [11] and [12] are two works done on Chinese and Arabic.

As the main focus of this work is on Bangla, so if works done on automated spell checking in Bangla is considered then number of papers are small [13], [14], [15], [16], [17], [18], [19], [20] and [21] that focus on this common concern with this paper. All these papers have different techniques deployed but there are two common facts about all these papers and those are small corpus size and lack of balance. As a solution to the Bangla spell checking problem, clustering and edit distance is used by [13] Prianka Mandal, B M Mainul Hossain claiming to show an accuracy of 99.8%, which was applied on 2450 words [13]. Naushad UzZaman and Mumit Khan claimed accuracy of 91.67% with the data set of 1607 words using mapping based on double metaphor and edit distance as a solution [16]. To identify phonetic error, string matching is used by Bidyut Baran Chaudhuri [17]. He used 2 dictionaries to check and correct errors but his process mainly considered phonetic errors and as it uses dictionary and the reverse dictionary for correction of misspelled words it needs double memory. In another one of M. Z. Islam, M. N. Uddin and M. Khan's paper they used stemming algorithm in which if the stem is

978-1-7281-3445-1/19/$31.00 ©2019 IEEE

not detected, the suggestion generation process will produce a list of possible suggestions [18]. They used edit distance algorithm to discover the best match. Another work focusing on phonetic error is done by Naushad UzZaman and Mumit Khan in which they modified phonetic encoding based on soundex algorithm to match Bangla phonetic [15]. Using finite state automaton Md. Munshi Abdullah et. al. suggested a strategy to suggest probable list of correct words against the error word [19]. NH Khan, GC Saha, B Sarker, MH Rahman used character-based N-gram for checking error but it didn't suggest any correction to the misspelled word [21]. To detect error words binary search method and a direct dictionary look up; to suggest correct word against the incorrect word recursive simulation method is used by A.B.A. Abdullah, A. Rahman in their paper [20].

There exist many spell checking and correction works for English which have accuracy of 86.95%, 96.1% and 96.4% for both real word and non-word errors [8], [7]. Even though authors of [6], [7] has studied English spell checking and correction in their papers, corpus size or test data size were not mentioned. In [11] a work on Chinese language spell checking and correction, two algorithms (edit distance and Soundex) combined with pinyin is used. In [12], Arabic language spelling checker system uses context words and N-gram language models. In the experiment they used a corpus of 41,170,678 words and 28 confusion sets. Average accuracy was 95.9% handling real-word errors as well as non-word errors both.

A. Types of Error

In classification done by Kukich [22], all spelling errors are of two types; real-word error and non-word error. Contextually inappropriate though the word is valid is called real-word error. Let's consider, "I eat milk.", "eat" is not an invalid word but contextually it's inappropriate. An example in Bangla would be, "আমার আব্বু বাটি যাবে", বাটি is not an invalid word but contextually it is inappropriate. Lexically invalid words are called Non-word error. For example, in the sentence "we aer going to play", 'aer' is an invalid word. Similarly in Bangla, 'নাটি' is an invalid word in "আমার আব্বু নাটি যাবে". Some more classifications are done by Kukich [22] for non-word spelling errors. Mainly cognitive error and typographical error. Cognitive error is, when the spelling is not known by the writer. When the writer makes typing mistake that is called Typographical error. For example, different errors of the Bangla word "চলমান" caused by insertion, deletion, substitution and transposition while typing are, "চলম্মান", "চমান", "চমমান" and "চলমনা".

III. METHODOLOGY

In this work we are handling non-word, especially typographical errors. We used direct dictionary lookup method to detect an error word. We have a dictionary from which we are searching if the typed word is available. If the typed word is not available then we can consider it as an error and then use edit distance algorithm to correct the erroneous word. To correct the error, we estimate similarity of structure between the word which is misspelled and possible correct words using minimum edit distance. It measures the least number of overall operations required to convert one string into another. There can be various operations like insertion, deletion and/or substitution. For calculation of edit distance, Levenshtein algorithm [23] is used as written in Algorithm 1. The working principle of Algorithm 1 is that, if we consider two strings X and Y (X= misspelled word, Y= possible correct word) with the lengths i, j and create a matrix which has order of i×j edit distance between different sub-strings of X and Y. The corresponding values will express the minimum distance between X and Y. Let's assume that cost for each insertion and deletion is 1 and 2 is the cost for substitution.

Algorithm 1. Algorithm to calculate minimum edit distance

```
minimum_dist(misspelled, possible_word)
   i ← len(misspelled)
   j ← len(possible_word)
   create distance matrix disMat [i+1, j+1]
   for each column k ← 0 to j
   for each row l ← 0 to i
         disMat[k,l]←minimum(disMat[k-1,l]+
         insert-cost(possible_word1),
         disMat[k-1,l-1]+substitute-cost(misspelled1,
         possible_wordk),
         disMat[k,l-1]+insert-cost(misspelled1))
   return min-distance
```

System chooses the word with minimum edit distance for the misspelled word. We calculate the accuracy using fractional accuracy method [24]. The accuracy is calculated for fractional accuracy by taking into account the matching of the suggested word with the intended word and matching order. Accuracy rate is therefore used to deal with the mentioned issue. Suppose w_m matches with v_{mi}, where w_m is intended word and v_{mi} is suggested word (i.e. w_m equals v_{mi}, where $1 \le i \le n + 1$), the accuracy would be,

$$Acc = \frac{n+1-i}{n} \times 100\% \quad (1)$$

Here n is length of suggestion list and i is the position of matching word. If i is equal to $n + 1$, failure occurs. The meaning of $(n + 1)$-*th* match is, there is no match and the accuracy equals 0. The whole methodology is shown in Figure 1.

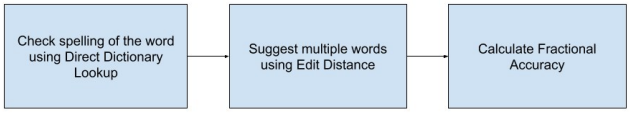

Fig. 1. Block Diagram of Proposed Method

IV. IMPLEMENTATION

As described earlier we are using a dictionary to check if the input word is correct or not. We have a large dictionary consisting of 4,039,603 unique entries. If any word of the entry text is not found in the dictionary then the word is marked as erroneous word. When an error word is detected,

edit distance method is used to generate list of the correct words. Then we find which word from the generated list is similar to the expected word and we calculate fractional accuracy based on the method described above. For example, we have a list of 4 generated words, we found our intended word at position two then our fractional accuracy will be,

$$\frac{4+1-2}{4}\times 100\% = 75\%$$

The text used in our research are collected from social sites as well as auto generated. One third (50000 words) of the data is collected from social media and two third (100000 words) is auto generated. Percentage of accuracy based on edit distance range is shown in Table 1.

TABLE I. ACCURACY BASED ON EDIT DISTANCE RANGE

Edit distance range	Total number of error word	Total number of success word	Accuracy
1	135,000	132,000	97.78%
2-3	15000	13245	88.3%

We can see that if edit distance range is between 2 and 3 then we get higher accuracy. Another table (Table 2) shows the accuracy based on source of input text.

TABLE II. ACCURACY BASED ON SOURCE

Source	Total number of error word	Total number of success word	Accuracy
Social Media	50,000	47,145	94.29%
Auto generated	100,000	98,100	98.1%

V. COMPARATIVE RESULTS

After discussing different types of results of our work it is better to conclude with total accuracy result which shows that this research work is promising enough. As we see from Table 3 and Figure 2 that the higher the suggestion word list length is the higher accuracy rate we get. If the suggestion list length is only 1 then the accuracy is 65.17% whereas the accuracy is 96.83% if the length is 9.

TABLE III. ACCURACY TABLE BASED ON SUGGESTION LIST LENGTH

Suggestion list Length	Accuracy
1	65.17%
2	74.14%
3	82.62%
4	86.69%
5	88.76%
6	92.10%
7	93.79%
8	94.48%
9	96.83%

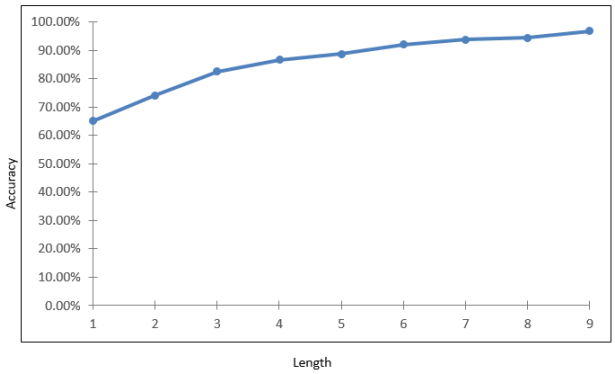

Fig. 2. Accuracy against suggestion list length

Also if we make a comparative table of all works done in Bangla spell checking, we see that our work is promising than others. Table 4 shows the comparative study.

TABLE IV. THE COMPARATIVE ANALYSIS OF REPORTED WORKS

Work	Type of Errors Handled	Algorithm	Test Data Size	Accuracy
This work	Non-word error	Edit distance	150,000 words	96.83%
[13]	Phonetic, typographical	Clustering algorithm	2450 words	99.8%
[14]	OCR generated, typographical errors, Phonetic errors	phonetic encoding, edit distance	1607 words	91.67%
[15]	Typographical, Phonetic	Phonetic encoding algorithm	*Not mentioned*	Above 80%
[16]	Orthographic rules in Bangla	Mapping rule, Double metaphone	1607 words	91.67% accuracy found when used edit distance 0. No. of correct words is 1473 and error words is 134
[17]	Phonetic error	String matching	25,000 words	Showed high accuracy. 5% false positive detection
[18]	Complex orthographic rules	Edit distance, Stemming	13,000 words	for correction of single error misspellings 90.8% for correction of multiple error misspellings 67%
[19]	Substitution errors, insertion errors	Finite state automaton	291 words	for single character misspellings correction: 92% for multiple character misspellings correction:70%

[20]	Cognitive phonetic errors, Typographical errors	Recursive Simulation algorithm, Direct dictionary look up method	*Not mentioned*	*Not mentioned*
[21]	Non-word error	*N*-gram Model (character based)	50,000 incorrect words , 50,000 correct words	96.17%

VI. CONCLUSION

So, the goal of this research was finding a better way which can handle non-word Bangla spelling mistakes. The authors proposed a model based on edit distance method but with different accuracy calculation method than the previous papers on Bangla spelling correction. It is clear that the model proposed here is promising as a large dictionary is used and the accuracy of the model is 96.83%. This research work dealt with non word errors. The authors are working on another research which will deal with real word errors.

REFERENCES

[1] Bengali language, Available at: https://en.wikipedia.org/wiki/Bengali_language (Last Accessed: January, 10, 2019).

[2] Aye Myat Mon, "Spell checker for Myanmar language", 2012 International Conference on Information Retrieval & Knowledge Management, 2012

[3] Mohammad Bagher Dastgheib, Seyed Mostafa Fakhr ahmad, Mansoor Zolghadri Jahromi, "Perspell: A new Persian semantic-based spelling correction system", Digital Scholarship in the Humanities, Vol. 32, Pages 543–553, 1 September 2017.

[4] Amita Jain, Minni Jain, "Detection and correction of non-word spelling errors in Hindi language", 2014 International Conference on Data Mining and Intelligent Computing (ICDMIC), 2014.

[5] Behrang Qasemi Zadeh, Ali Ilkhani, Amir Ganjeii, "Adaptive Language Independent Spell Checking using Intelligent Traverse on a Tree", IEEE Conference on Cybernetics and Intelligent Systems, 2006.

[6] Shashi Pal Singh, Ajai Kumar, Lenali Singh, Mahesh Bhargava, KritikaGoyal, Bhanu Sharma, "Frequency based Spell Checking and Rule based Grammar Checking",in International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) 2016.IEEE,2017,pp 4435 – 4439.

[7] Andrew Carlson, Ian Fette, "Memory-Based Context-Dependent Spelling Correction at Web Scale",in Sixth International Conference on Machine Learning and Applications (ICMLA 2007), IEEE 2007.

[8] Ya Zhou, Shenghao Jing, Guimin Huang, Shaozhong Liu, Yan Zhang, "A Correcting Model Based on Tribayes for Real-word Errors in English Essays." 2012 Fifth International Symposium on Computational Intelligence and Design.

[9] Pratip Samanta, Bidyut B. Chaudhuri, "A simple real-word error detection and correction using local word bigram and trigram", Twenty-Fifth Conference on Computational Linguistics and Speech Processing (ROCLING 2013).

[10] Yizheng Cai, Kevin Roland Powell, Ravi Chandru Shahani, Lei Wang, "LINGUISTIC ERROR DETECTION", United States, US 8,855,997 B2, Oct. 7, 2014.

[11] DongHui Li, DeWei Peng. "Spelling Correction for Chinese Language Based on Pinyin- Soundex Algorithm." Internet Technology and Applications (iTAP), 2011.

[12] Majed M. Al-Jefri, Sabri A. Mahmoud, "Context-Dependent Arabic Spell Checker using Context Words and N-gram Language Models."2013 Taibah University International Conference on Advances in Information Technology for the Holy Quran and Its Sciences, 2013.

[13] Prianka Mandal, B M Mainul Hossain, "Clustering-based Bangla Spell Checker," in 2017 IEEE International Conference on Imaging, Vision & Pattern Recognition (icIVPR). IEEE 2017, pp 1-5.

[14] N. UzZaman and M. Khan, "A comprehensive bangla spelling checker," In the Proceeding of the International Conference onComputer Processing on Bengali (ICCPB), Dhaka, Bangladesh, 2006.

[15] N. UzZaman and M. Khan, "A bangla phonetic encoding for better spelling suggestions," in Proc. 7th International Conference on Computer and Information Technology, Dhaka, 2004.

[16] N. UzZaman and M. Khan, "A double metaphone encoding for bangla and its application in spelling checker," in 2005 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2005, pp. 705–710.

[17] Chaudhuri, Bidyut Baran, "Reversed Word Dictionary and Phonetically Similar Word Grouping based Spell-checker to Bangla Text." Proc. LESAL Workshop, Mumbai. 2001.

[18] Md. Zahurul Islam, Md. Nizam Uddin and Mumit Khan, "A Light Weight Stemmer for Bengali and its Use in Spelling Checker," Proc. 1st Intl. Conf. on Digital Comm and Computer Applications (DCCA07), Irbid, Jordan, March 19-23, 2007.

[19] Abdullah, Md Munshi, Md Zahurul Islam, and Mumit Khan. "Error-tolerant Finite-state Recognizer and String Pattern Similarity Based Spelling-Checker for Bangla", Proceeding of 5th International Conference on Natural Language Processing (ICON). 2007.

[20] Abdullah, A. B. A., and Ashfaq Rahman, "A Generic Spell Checker Engine for South Asian Languages." Conference on Software Engineering and Applications (SEA 2003). 2003.

[21] Khan, Nur Hossain, et al. "Checking the Correctness of Bangla Words using N -Gram." International Journal of Computer Application 89.11 (2014).

[22] K. Kukich, "Techniques for automatically correcting words in text," ACM Computing Surveys (CSUR), vol. 24, no. 4, pp. 377– 439, 1992.

[23] Daniel Jurafsky and James H. Martin, (2000), "Speech and Language processing", USA: Prentice-Hall, Inc.

[24] Md. Tarek Habib, Abdullah Al-Mamun, Md. Sadekur Rahman, Shah Md. Tanvir Siddiquee, Farruk Ahmed, "An Exploratory Approach to Find a Novel Metric Based Optimum Language Model for Automatic Bangla Word Prediction", International Journal of Intelligent Systems and Applications, Vol. 2, pp. 47-54, 2018.