BAUSTJ
BAUST Journal

Bangla Spell Checker: A State of the Art

Naima Islam Nodi[1*], Nabila Anwar[2], Iftakharul Islam[3], S. M. Raihan[4], Md. Moazzem Hossain[5], Md. Mamunur Rashid[6] and Syed Akhter Hossain[7]

[*1, 2, 3, 4, 5, 6]Department of CSE, Bangladesh Army University of Science and Technology, Saidpur, Bangladesh
[7]Department of CSE, Daffodil International University, Dhaka, Bangladesh

emails: [*1]*naimanodi@gmail.com;* [2]*nabilaanwar43@gmail.com;* [3]*iftakharopu27@gmail.com;* [4]*raihanrajon42@gmail.com;* [5]*moazzem.cse10@gmail.com;* [6] *mamunst2003@yahoo.com and* [7]*aktarhossain@daffodilvarsity.edu.bd*

ARTICLE INFO

Article History:

Received: 9th September 2019
Revised: 25th July 2020
Accepted: 27th July 2020
Published online: 25th August 2020

Keywords:

Spell-checker
Misspelled
Edit-distance
Phonetic
Bangla

ABSTRACT

Spell checker is a type of process that checks words in a stream of text and determines if it is misspelled or not. The purpose of a Spell checker is to determine the wrong word(s) and provide suggestions. It works in a way that gives a rank wise suggestion. In this paper, we will find out the methods that have been used over the years till now along with their limitations, challenges, and yet discover works in this field. This paper consists of the overview of all the research related works that have been done from the early works and a pathway to start contributing in this field.

1. INTRODUCTION

This paper demonstrates an overview and of the state of the art of the Bangla Spell checker. Bangla is the 7th spoken language all around the world and spoken by approximately 300 million people. Bangla is spoken in the south Asian region. Regarded as the official language of Bangladesh and some province of India (e.g.: West Bengal, Tripura, and Barak Valley). This is 2nd most used language in India. But there are not substantial research works done with Bangla Language. Bangla language is a descendant of Sanskrit. It has a complex orthographic structure and word construction. Spelling mistakes is one common problem for any language. It is relevant to have a well-developed spell checker in Bangla for better typing and processing Bangla word more comfortable. The works done for developing the Bangla spell checker is not efficient till now. In this paper, we aim to point out the type of researches have done for Bangla spell checker, their limitations, accuracy, and the challenges. In the following discussion we'll try to discuss these topics:

- The present status of the Bangla Spell Checker.
- Avoid reinventing the same work, give a milestone for the works that have been done so far.
- Ensure that any researchers who are working under the current status on Bangla Spell-checker would be benefited from this article and get to know about the current status as a whole.

2. BACKGROUND

Mainly three steps are followed by a working spell checker (Naushad UzZaman & Mumit Khan, 2006). These three things determine the theoretical solution of any spell checker. In the following sections we will describe the steps in the process of checking the spelling of a word:

- Find out if the word is misspelled or not.
- Provide suggestions if the word is misspelled.
- Most appropriate suggestions should come earlier.

The problems that are faced related to the spell checker work are:

- Phonetic similarity of Bangla characters.
- The difference between grapheme representation and phonetic utterances.
- Bangla is a language with complex orthographic rules, in result, there exists a large gap between the spelling and pronunciation of a word.
- A large number of words in Bangla is originated from Sanskrit, an ancestral predecessor of Bangla. However, these words have either been modified in terms of pronunciation or both in terms of spelling and pronunciation. Thus, there exists a gap between spelling and pronunciation requiring complex orthographic rules.

The typical errors that are often seen are several. These are:

- Typographic errors.
- Cognitive or phonetic errors.
- Visual errors.
- Space-related errors.
- Homonym errors.

3. LITERATURE REVIEW

The early work on a spell checker for Bangla is done in Reversed Word dictionary and phonetically similar word grouping based spell-checker to Bangla text (Bidyut Baran Chaudhuri, 2001). This paper introduced the term modified dictionary that has been used for phonetically similar words. And a reversed dictionary has been used for unmatched words and by comparing these two dictionaries, the unmatched area and error position has been identified and this correct result is being provided. Using conventional and reversed dictionary error localization is done like:

ৈে →য়া →ক→র →ন [conventional]

মূ←ম ←ষূর্ [Reversed-dictionary]

a ← b ← c ← d ← e ← f

So, c is the error point. A coding system for Bangla Spell Checker (Md Tamjidul Hoque & Md Kaykobad., 2002) was recommended to implement. The paper showed the common error patterns as “Keyboard Adjacency error”. It directed the adjacent keys to create an ambiguous word. Moreover, SQL queries to match in a database with the written word. If the word is matched with the result then assures that the word is correct, otherwise misspelled. This matching procedure is of constant complexity. Moreover, matching is faster and garbage-free as mentioned. A provided ranking suggestion based on adjacent co-ordination of keys keyboard letters. Sometimes there are valid words that are typed wrong and indicated as misspelled they are being recognized as BREAKING UP WORDS. Concurrently, a development proposal was suggested for Bangla as a subset of many Indian Language contemplating the phonetic similarities, error detection, and position finding and solving on issues by implementing a simplified morphological analyser (Chaudhuri., 2002). This early work consists of some commend for creating a corpus containing about 3 million words. The system worked fine for one-fourth data of its corpus and gave high accuracy. Though error detection failed several times on account of conjugate words. The binary dictionary has been used as a standard database from which words are sorted according to their hashed weights (ABA Abdullah & Ashfaq Rahman, 2003). A recursive simulation algorithm has been used to provide suggestion dialog circularly. Phonetic encoding for better spellchecking in Bangla suggests using an algorithm (Naushad UzZaman & Mumit Khan, 2004). This paper refers to use the “Soundex” algorithm as the basis of spell checking and also provided their own defined Unicode for Bangla letters. They have partitioned the letters with phonetic similarity to produce a single code for each division. By constructing an encoded dictionary that will be searched first after encoding the written word and checked either is it the original one or been misspelled. If the entry exists then either the original exists in the list of words corresponding to this code, or the word is misspelled and the list is offered as the set of alternatives for the original word. If the entry does not exist, then the alternatives must be suggested using the edit-distance algorithm. They have also provided suggestions for the “juktakkors” by encoding “”. But other conjuncts letters have not been encoded, also if any word with “juktakkors” at first letter does not provide a good suggestion. Some example of correcting errors using Phonetic Matching:

Input: আসণ suggestion: আসন

Input: মুগধ suggestion: মুগ

On Double Metaphone Encoding for Bangla, the main challenge was complex orthographic rules of Bangla grammar (Naushad UzZaman & Mumit Khan, 2005). The rationale for mapping rules are given here, assumed that Bangla text is encoded using Unicode Normalization Form C(NFC). Including vowels, consonants, and conjuncts and all different contexts. The paper proposed 107 transformation, which is, the letters are encoded based on how the letters and conjuncts are pronounced in a different context. To map logically, each Bangla letter is followed by the name of that letter used in the Unicode chart. Finally, the pronunciation of the letter in the International Phonetic Alphabet (IPA) is addressed by the Unicode code point of that letter. They have experimented with the procedure on 1607 misspelled words and resulting accuracy with 91.37%. Some example of error correction using a Double Metaphone Encoding System:

"পাচ" →"পাঁচ" Edit distance: 1

"নসট" →"নষ্ট" Edit distance: 2

"পছন্দিনয়" →"পছন্দনীয়" Edit distance: 3

In (Naushad UzZaman & Mumit Khan, 2006) a subset called “short list” has been used to provide a suggestion list of the lexicon. Solves the problem of phonetically similar.

Conjugate and simple words. To correct the words that include "'s has been coded differently. Along with the "phonetic encoding", approximate string matching algorithm has been used to provide a better suggestion of the correct words and also to rank the correct words. For example, "ক্ষ" is sounded like "খ" at the beginning of word so "ক্ষ" has given the same code as "খ". ক্ষয় -> খয়. But at the end of the word "ক্ষ" pronounced as "কখ". পক্ষ -> পকখ. The very contemporary work that has been done in the province of Bangla spell checker is clustering based (Prianka Mandal & BM Mainul Hossain, 2017). Here, through congregate words and time and space complexity are considered. The data set is separated into several distinct groups. Similar observations belong to a similar group. Partitioning Around Medoids (PAM) algorithm was used to partition the data set into a specified number of disjoint clusters. Moreover, the accuracy of this technique has been told to be 99.8%. The clustering of the dictionary is done in several steps. Those are:

- Separating phonetic similar word.
- Creating clusters.
- Filtering clusters.
- Selection of best method.
- Detecting errors.
- Correcting errors.

The following Table compares the work is done and limitations of the re-known papers done on Bangla spell checking. This table may help researchers to get knowledge about the progress and improvement in this sector so that they can continue working for further development in the path of digitizing Bangla.

Table 1

Research on Bangla Spell Checker

Title	Authors	Work Done	Limitations
A Comprehensive Bangla Spelling Checker	(Naushad UzZaman & Mumit Khan, 2006)	Solves the misspelled word with a maximum edit distance of two, using phonetic encoding.	Unable to solve Nth character error, space-related errors, homonym errors, conjuncts with unusual punctuation, and different pronunciation in a different context.
Reversed word dictionary and phonetically similar word grouping based spellchecker to Bangla text	(Bidyut Baran Chaudhuri, 2001)	A modified dictionary has been used for phonetically similar words. A reversed dictionary is used for unmatched words and by comparing these two dictionaries, the unmatched area and error position is identified and thus the correct result is being provided.	There is no discussion of conjugate words known as "juktakkors" in bangla.
Coding System For Bangla Spell Checker	(Md Tamjidul Hoque & Md Kaykobad., 2002)	SQL queries to detect frequent errors, devising coding and it's utilization, a priority of minor errors, improving the suggestion list and its' presentation, coding for detecting the minor or delta errors, and breaking-up words.	Words having similar sounding based suggestions will come later.
A Generic Spell Checker Engine for South Asian Languages	(ABA Abdullah & Ashfaq Rahman, 2003)	Add-in approached has been used to detect highly misspelled words and provides a suggestion for it.	Used recursive simulation algorithm has lots of bugs and fails to provide correct suggestions.

A Bangla Phonetic Encoding for Better Spelling Suggestions	(Naushad UzZaman & Mumit Khan, 2004)	Phonetically error correction and provided Unicode for Bangla letters are phonetically the same to spell.	Does not provide Unicode for all the vowels. Some of the English letters more different letters in bangla that may cause an ambiguous situation.
A double metaphone encoding for bangla and its application in spelling checker	(Naushad UzZaman & Mumit Khan, 2005)	Solves error with edit distance one using double metaphone encoding.	Remains some ambiguity in case of multiple pronunciations of the same letter.
Clustering based Bangla Spell Checker	(Prianka Mandal & BM Mainul Hossain, 2017)	Reduces time and space complexity by Clustering the dictionary. The Clustered dictionary is constructed on structural Similarity and phonetic similarity of words.	Does not solve space-related errors and homonym errors.
A Spell Checker and Corrector for the Native South African Language South Sotho	(LA Grobbelaar & JDM Kinyua, 2009)	A step has been taken to make a multithreaded spell checker searched in the dictionary to give the correct result.	Word wise correction is being provided by separating the prefect Translation of more than one phrase at a time is difficult in this approach. Excluding “Full stop”, other punctuations are not been considered.
A Non-Learning Approach to Spelling Correction in Web Queries	(Jason Soo, 2013)	Adverse environment spelling correction algorithm, Segments has been used which is a hybrid approach that uses N-gram and a series of substring generation rules, does not require training data, also language and domain-independent.	Algorithm complexity is higher. Web crawling can provide misspelled words sometimes. In this case, result in accuracy decreases.
Spell Checker	(V. V. Bhaire,, A. A. Jadhav,, & P. A. Pashte, 2015)	A spell has been checked using edit distance and comparing with the corrected dictionary. Digital tree data structured has been used to store dynamic set or associative array of English words. An effort has been made to omit space related errors.	It does not solve homonym errors, name (noun) errors, unable to solve multiple errors in a word. It’s not a language-independent solution.
Automated word prediction in Bangla language using Stochastic language models	(Md. Masudul Haque, Md. Tarek Habib, & Md. Mokhles, 2015)	N-gram language model, backoff and deleted interpolation techniques have been used to predict bangla words in a sentence.	The Solution provides less accuracy for a large data set.

4. RESEARCH CHALLENGES

The first requirement in Spell Checking is an enriched dataset. But there is no centrally developed dataset and also our “Bangla Academy Dictionary” has not been yet published in digital platforms. Online Bangla websites are filled with misspelled words. So that would be a great challenge to develop an error-free dataset for comparing misspelled words. Microsoft office had also stopped to provide the CSAPI till 2000, which was used in spell checking of local languages. Selecting training examples as well as test examples will be a great concern. There exist a

lot of approaches for spell checking of multilingual languages. It is also difficult to choose the appropriate one to solve conjugate word errors, space-related errors, homonyms errors, etc. The use of both a modified encoded dictionary and a reversed dictionary can give a better result. But building a reversed dictionary will consequence a greater challenge to meet. The amount of research that has been done over the years was based on a simple word spell checker. There is no notable work that was based on the word that started with a conjugate word (known as "juktakkors"). Apart from this, there is no research done for the homonym words that have the same phonetic utterances but differ by meanings. An area of research was mentioned earlier as "Breaking Up words" which is if there is any space in between a correct word. Any word that makes no sense or grammatically inappropriate is not enlisted for any correction. Hence, we can say that there are a lot of scopes for researching in Bangla Language Processing. The Spell checker is one of them. Though there exist some researches on Spell-checking, we should watch the possibilities that are yet to make.

5. CONCLUSION

In this paper, we've tried to provide a brief history of works on Bangla linguistic and acknowledge the valuable works done in this field. This would help the new researchers to find the way from where and how to start their work on Bangla spell checking to improvise it. We've reviewed the major works which have the most impact on Bangla Spell Checker and also compared with works on other languages to get the knowledge of differences and similarities to meet the challenges. Further works can be guided towards an easy approach to select a research area and contribute in this sector.

ACKNOWLEDGEMENTS

We would like to acknowledge our indebtedness and render our warmest thanks to our Supervisor, Professor Dr. Syed Akhter Hossain, and Professor Dr. Md Mamunur Rashid. Without their guideline, this would not be possible. We would also like to thank Mr. Md. Moazzem Hossain, who directed us to develop the survey from scratch. Our sincere gratitude towards Mr. Md. Shajalal and all respected researchers in Bangla Language Processing field who made it possible to work with this topic.

REFERENCES

ABA Abdullah, & Ashfaq Rahman. (2003). A generic spell checker engine for south asian languages. *Conference on Software Engineering and Applications.*

LA Grobbelaar , & JDM Kinyua. (2009). A spell checker and corrector for the native south african language, south sotho. *Annual Conference of the Southern African Computer Lecturers.*

Naushad UzZaman, & Mumit Khan. (2006). A comprehensive bangla spelling checker.

Bidyut Baran Chaudhuri. (2001). Reversed word dictionary and phonetically similar word grouping based spell-checker to. *LESAL Workshop.* Mumbai.

Chaudhuri., B. B. (2002). Towards indian language spell-checker design. *Language Engineering Conference.*

Jason Soo. (2013). A non-learning approach to spelling correction in web queries. *22nd International Conference on World Wide Web.*

Md Islam, M. U. (2007.). A light weight stemmer for bengali and its use in spelling checker.

Md Tamjidul Hoque, & Md Kaykobad. (2002). Coding system for bangla spell checker. *5th International Conference on Computer and Information Technology*,.

Md. Masudul Haque, Md. Tarek Habib, & Md. Mokhles. (2015). Automated word prediction in bangla language using stochastic language models. *International Journal in Foundations of Computer Science & Technology*.

Naushad UzZaman, & Mumit Khan. (2005). A double metaphone encoding for bangla and its application in spelling checker. *International Conference on Natural Language Processing and Knowledge Engineering.* IEEE.

Prianka Mandal, & BM Mainul Hossain. (2017). Clustering-based bangla spell checker. *IEEE International Conference on Imaging, Vision & Pattern Recognition.*

Naushad UzZaman, & Mumit Khan. (2004). *A bangla phonetic encoding for better spelling suggestions.* BRAC University.

V. V. Bhaire,, A. A. Jadhav,, & P. A. Pashte. (2015). Spell checker. *International Journal of Scientific and Analisys Publication.*