I.J. Modern Education and Computer Science, **2017, 6, 40-47**
Published Online June 2017 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijmecs.2017.06.06

A Systematic Literature Review on Spell Checkers for Bangla Language

Prianka Mandal
Institute of Information Technology, University of Dhaka, Dhaka, 1000, Bangladesh
Email: prianka.iit.du@gmail.com

B M Mainul Hossain
Institute of Information Technology, University of Dhaka, Dhaka, 1000, Bangladesh
Email: raju@du.ac.bd

***Abstract*—**Spell checkers check whether a word is misspelled and provide suggestions to correct it. Detection and correction of spelling errors in Bangla language which is the seventh most spoken native language in the world, is very onerous because of the complex rules of Bangla spelling. There is no systematic literature review on this research topic. In this paper, we present a systematic literature review on checking and correcting spelling errors in Bangla language. We investigate the current methods used for spell checking and find out what challenges are addressed by those methods. We also report the limitations of those methods. Recent relevant studies are selected based on a set of significant criteria. Our results indicate that there are research gaps in this research topic and has a potential for further investigation.

***Index Terms*—**Systematic Literature Review, Spelling Errors, Spell Detecting, Spell Checking, Spell Checker, Bangla Language, Misspelled Word

I. Introduction

A Systematic Literature Review (SLR) is a process which identifies, evaluates and interprets all obtainable research relevant to a particular research area of interest. SLR can be exercised to summarize the existing evidence of a particular research topic, to identify any research gaps, to provide suggestions for further investigation and to provide a cooperation for generating new hypotheses. However, SLR requires more effort than traditional literature reviews [1]. SLR aims to detect as much as possible relevant information of a particular research domain. Conducting a systematic literature review on a particular topic is very supportive and beneficial.

Misspelled word is a word in a text which is not a valid word of a language and typically not found in a dictionary of the corresponding language. If it is in the corresponding dictionary then it is determined to be correctly spelled. Spell checking is a process of detecting spelling errors and provides most probable proper words to correct them. Spell checkers help users to improve their writing skill by reducing spelling errors. Detecting misspelled words and correcting those misspelled words automatically is a great research challenge. There are many well-established spell checkers for English and other western languages, but there is no well-established spell checker for Bangla. There are few research works on Bangla spell checking. Therefore, conducting a systematic literature review on checking and correcting spelling errors in Bangla language makes this topic more beneficial and would be very helpful for interested researchers to work on this engrossing research topic.

In this paper, we present our findings from a systematic literature review on checking and correcting spelling errors in Bangla language. Our approach looks into current methods on which Bangla spell checkers are developed, challenges which are addressed and limitations of existing works. By conducting an SLR on this topic, we make it easier for interested researchers to determine the present state of research on this topic. Our results make it possible for interested researchers to develop Bangla spell checker based on best knowledge and practice across many previous studies.

The remainder of the paper is organized as follows: Section II discusses the background of this study. Section III presents the methodology of the work. Section IV presents the validity of our review. Section V presents results of this SLR. Section VI provides a discussion about our findings. Finally, we summarize our conclusions in Section VII.

II. Background Study

In this section, we discuss about concepts that are relevant to this study. We provide an overview of Bangla language, since the main focus of this study is on spell checking task specifically for Bangla language. Different types of spelling errors along with spell checking techniques are also discussed here.

A. Bangla Language

Bangla or Bengali, a member of the Indo-Aryan languages, is the state language of Bangladesh and the second most spoken language in India. Over two thousand ten million people speak in Bangla, the majority of whom live in Bangladesh and in the Indian state of West Bengal. Bangla is the seventh most spoken native

language in the world. Even though it is easy to use Bangla verbally, due to its complex script nature, it is rather difficult when it comes to writing properly.

Bangla language has 49 letters in its alphabet and 10 digits in decimal number system [32]. Bangla alphabet comprises of 11 vowels and 39 consonant characters. Bangla alphabet has no concept of upper/lower case. Here, we discuss some challenges which are required to be addressed because of the complex rules of Bangla:

1. Phonetically similar characters: There are some characters in Bangla which are phonetically similar. Example: ন(n) and ণ(N), শ(sh),ষ(Sh) and স(s).
2. Consonant clusters or *Juktakkhors*: Consonant cluster consists of up to four consonants which are not separated by vowels. Example: ক্ক, ন্দ.
3. Use of *Phalas*: There are different types of phala such as *YA-phala, RA-phala* and *LA-phala*. Example: নব্য, প্রথম.
4. Use of *Matra*: Matra is a headline of many Bangla characters. Example: পর্ব, গর্ব.
5. Conjuncts with unusual pronunciations: Example: ক্ষ = ক + ◌্ + ষ. বক্ষ pronounced as বকখ.
6. Different pronunciations on different context. Example: ক্ষ = ক + ◌্ + ষ. ক্ষমা pronounced as খমা. বক্ষ pronounced as বকখ.
7. Multiple pronunciations of some letters in the same context.
8. Use of vowel diacritics: Every vowel has its diacritic. These vowel diacritics are used with consonants. These vowel diacritics are ◌া, ◌ি, ◌ী, ◌ু, ◌ূ, ◌ৃ, ◌ে, ◌ৈ, ◌োand ◌ৌ.
9. Use of modifier symbols. There are some modifier symbols in Bangla such as ৎ, ◌ং, ◌ঃ and ◌ঁ.

B. Types of Spelling Errors

Kukich [2] classified spelling errors into two types: non-word error and real-word error. Non-word error is word level error that occurs when a word is not a valid word. Example: "ভল" (vol) is a non-word error, because it is not a valid word. Real-word error is sentence level error that occurs when a word is a valid word but it is inappropriate in the context of that sentence. Example: "আমার আসা নেই" (amar asa nei). In this sentence, the word "আসা" (asa) is a valid word but it is inappropriate in the context of this sentence.

Kukich [2] also provided an alternative classification of spelling errors and divided them into two types, Cognitive error and Typographical error. Cognitive error occurs when user forgets the correct spelling during typing or does not know the correct spelling. Example: typing "রিদয়" (ridoy) instead of "হৃদয়" (hridoy). Typographical error occurs when user makes mistakes during typing. Example: typing "ভুমিকা" (vumika) instead of "ভূমিকা" (vumika).

There are many types of typographical spelling errors that can occur, such as insertion error, deletion error, substitution error and transposition error. Insertion error occurs when a user types an extra character in a word. For example, the word "পরিববার" (poribbar) contains an extra character "ব" . The correct word is "পরিবার" (poribar). Deletion error occurs when a user forgets to type a character in a word. For example, the word "সাধাণত" (sadhanata) is misspelled and the character "র" is missing. The correct word is "সাধারণত" (sadharanata). Substitution error occurs when a user types wrong character in any position of a word. For example, the word "প্রাথনিক" (prathonik) is misspelled. The word would be correct if replace the character "ন" by "ম" and the correct word is "প্রাথমিক" (prathomik). Transposition error occurs when user types a word in which characters exchange their place. For example, the word "পকল" (pokol) is misspelled. The correct word "পলক" (polok) can be obtained if characters "ক" and "ল" interchange their place.

Most of the misspellings occur because of

- phonetic similarity of Bangla characters,
- the difference between the grapheme representation and phonetic utterances, and
- lack of proper knowledge of spelling rules [3].

C. Spell Checker

A Spell checker is an application that is used to detect misspelled word and correct spelling error. The main tasks of a spell checker are

1. Check whether a word is correct or misspelled,
2. Generate candidate corrections if the word is misspelled and
3. Provide the most likely candidate corrections as suggestions to the user.

A spell checker may a stand-alone application which takes texts from users and provides suggestions if there are any misspelled word in that text. Spell checker can be implemented as a part of a large application such as email client, text editor and word processor.

There are various algorithms which are used when implementing spell checkers that are accompanied with word suggestions. One approach is to encode all words into its corresponding phonological code and then check spelling errors and generating suggestions. This phonetic similarity is generally measured by different encoding algorithms such as Soundex [4], Metaphone [5], Double Metaphone [6] and PHONIX [7]. Soundex is a phonetic algorithm that is used to group phonetically similar letters together and assign each group a numerical number. Soundex works on a letter-by-letter basis and cannot handle context-sensitive rules. Metaphone is another phontetic algorithm that is more accurate than Soundex because it considers the context-sensitive rules of English pronunciation. Double Metaphone is a new version of the phontetic algorithm that ables to handle the problem of Metaphone and produces more accurate results than the

Metaphone algorithm. PHONIX is an improved version of Soundex encoding. These algorithm are language-specific and typically designed for English language.

Structural similarity can be used to detect and correct misspelled words. Edit distance [8] is used to estimate structural similarity between misspelled word and candidate corrections. Edit distance measures the minimum number of total operations required to transform one string into the other. Three different operations are applied when measuring edit distance: insert a new character into one of the strings, delete an existing character, and replace one character by another character. However, it is highly inefficient to evaluate the entire dictionary repeatedly.

Stemming is a process of splitting a word into stem and its affix. Stemming algorithm is used to improve the performance and effectiveness of spell checkers. Stemming can reduce dictionary size which is utilized as a part of different natural language processing applications, particularly for highly inflected languages. However, it is easy to extract root words by applying stemming algorithm for language like English [31]. The design of stemmers is language specific and requires some to significant linguistic expertise in the language, as well as the understanding of the needs for a spelling checker for that language [9]. The first published stemming algorithm is Lovins stemming algorithm [10]. Porter's algorithm [11] is the most common algorithm for stemming English. Porter's stemming algorithm is used for reducing derived words to their stems [34]. Porter stemmer applies a set of rules to iteratively remove suffixes from a word until none of the rules apply and the Lovins stemmer has a larger set of suffixes and does not apply its rules iteratively.

N-gram model is a statistical prediction technique that is also used to checking the correctness of a word. The idea of using n-grams in language processing was discussed first by Shannon [12]. An n-gram model is a type of probabilistic language model for predicting the next item in such a sequence in the form of a $(n-1)$ order Markov model. An n-gram of size one is referred to as a unigram, size two is a bigram, and size three is a trigram. Larger sizes are sometimes referred to by the value of n, such as four-gram, five-gram, and so on. One main advantage of the n-gram method is that it is language independent [13].

III. Methodology

In pursuance of systematic review guidelines [1], our systematic literature review was conducted and is comprised of few steps. Details of every steps is described in this section.

A. Identify the Need for a Systematic Literature Review

A lot of significant research works have been done in checking and correcting spelling errors for English language. Research works also have been done more or less for some other languages. However, some research works have been conducted on checking and correcting spelling errors in Bangla language.

Systematic literature review on checking and correcting spelling errors in Bangla language is necessary for those researchers who have worked on this research topic or are interested to work on this topic. However, there are no such papers based on systematic literature review on checking and correcting spelling errors in Bangla language to the best of our knowledge. Our motivation for this work is to take a preview of checking and correcting Bangla spelling errors.

In this paper, a systematic literature review on checking and correcting spelling errors in Bangla language is presented. Researchers can be come to know research works which have already been done on this topic, what are the limitation of those research works and what key challenges are addressed.

B. Research Questions

Identifying research questions is an important step in systematic literature review. Three research questions were considered when conducting this study. The research questions and their motivations are presented in Table 1.

Table 1. Research Questions and their Motivations

Research Question		Motivation
RQ1	What are the current methods used to develop Bangla spell checker?	To identify existing methods and algorithms to develop Bangla spell checker
RQ2	What key challenges are being focused when developing a Bangla spell checker?	To identify most of the challenges that captured researchers' attention when developing a Bangla spell checker
RQ3	What are the limitations of existing research?	To identify the limitations of existing research

C. Search for Studies

Keywords which were identified from research questions, were used to search for relevant papers. We used alternative spelling and synonyms for each search terms. These search keywords were linked then using Boolean "AND" and "OR" operators.

We used the following search string in our searches:

- check* AND ((spell* AND error*) OR (misspelled AND word*)) AND ((generate* AND suggestion*) OR correct*) AND (Bangla OR Bengali)
- (develop* OR implement*) AND (Bangla OR Bengali) AND ((spell* AND (check*) OR (misspelled AND word*))

D. Study Selection Criteria

The study selection criteria is based on the research questions. We included papers which focused on checking and correcting spelling errors in Bangla language and paper must be published as either a Journal paper or Conference proceedings.

There are few research works on Bangla spell checker. Therefore, we included most of the paper which answered our research questions. We only excluded those which have repeated works. In that case, we only included the most recent ones.

E. Study Selection Process

A manual search process was applied for searching documents which provided answers of our research questions. Initially, we used following sources for our search process.

i. IEEExplore
ii. ACM Digital Library
iii. Google Scholar
iv. Springer
v. ResearchGate
vi. CiteSeerX
vii. Science Direct

These sources were chosen because these sources covered the most of the publications of the selected research topic. Then, we selected papers based on our selection criteria. Next, we checked the reference section of selected study for any relevant papers or journals or books. We also checked papers which cited these selected studies.

F. Data Extraction

The following extraction form, shown in Table 2 was used to record the information gathered from the primary studies.

Table 2. Data Collection Form

Data Item	Value
Study Identifier	S#
Paper Name	
Author Name	
Paper Type	Journal / Book / Thesis / Conference
Name of e-library	IEEE, ACM or any other
Publication Year	2001-2016
Research Question	RQ1, RQ2, RQ3
Motivation of Paper	
Method of the Paper	Approaches / Algorithms / Techniques
Limitation of the Paper	

IV. Validity of the Systematic Review

We performed this systematic literature review for investigating the techniques of detecting and correcting spelling errors in Bangla. For this investigation, we accumulated all available evidence. The main threats to the validity of our study are that our publication selection may be biased and there can be lack of sufficient information resources. However, we made an effort to get in touch with all possible and relevant resources. We found that there have a few publication in this research domain. The search process was manual rather than an automated search process. Therefore, lack of sufficient resources may be a possible threat to our study. This implies that we may have missed some relevant resources. However, we searched most of the sources using our search strings many times and our results were same.

V. Results Analysis

At the beginning of the search process, initial results returned many studies which cover many other related topics such as English spell checker, grammar checker, spell checker for other languages and other Bangla language processing related topics. We focused only Bangla spell checker related papers. Using our inclusion and exclusion criteria, we identified relevant papers. We identified 11 papers (S1-S11) by the search process. One of the papers (S8) was a short version of another paper (S1) therefore we selected only one when give answers to our research questions. Thus, the number of identified papers is 10. We investigated those papers to give answers of our research questions. We also found some papers (S12-S14) from references section of those papers. We also considered some university thesis (S15-S16). Since Bangla is the most spoken language in Bangladesh and India, almost all of authors of the selected papers are either Indian or Bangladeshi. The publication year of these selected papers is between 2001 and 2016.

These studies and their descriptions in terms of conference or journal name, publication year, where to access them and which research questions were answered are shown in Table 3.

A. RQ1: What are the current methods used to *develop Bangla spell checker?*

Authors of S1 [14] constructed a dictionary where phonetically similar characters were mapped into single units of character code and also a reversed dictionary where the characters of each word were kept in reverse order. They used string matching algorithm for detecting the phonetic errors and then found error zone length where the error occurred. Then misspelled words were corrected using both dictionaries.

Authors of S2 [15] presented phonetic encoding which was based on Soundex algorithm. They modified this encoding to match Bangla phonetics. They represented phonetically similar Bangla letters by a single code. Authors of S3 [16] proposed a Double Metaphone encoding for Bangla spell checker. Based on how the letters and conjuncts are pronounced in different contexts, Bangla letters were encoded. In this proposed encoding, there were a total of 107 transformations, which included vowels, consonants, and conjuncts in different contexts.

Authors of S4 [17] applied the stemming algorithm which was used to find and return a stem. This stemming algorithm only strips suffixes from words. If the stem is not found, then a list of suggestions are produced using the suggestion generation process. They used edit

distance algorithm to find the best match. Authors of S5 [18] used a direct dictionary look up method for detecting a misspelled word. For generating the suggestions for misspelled word, they considered the error patterns in usual typing and also the phonetic error patterns found in Bangla language. For generating suggestions for typographical errors, they calculated edit distance between misspelled word and candidate words. For generating suggestions for phonetic errors, they used Double Metaphone encoding. Finally, they considered both the scores which were found by phonetic error and typographical error for ranking the suggestion list.

Authors of S7 [20] proposed an approach to checking the spelling errors of Bangla that employed a finite state automaton to probabilistically generate the suggestion list for a misspelled word. Finite state automata represent finite languages and are therefore useful for storing words for spell checking [33]. Authors of S9 [22] used a direct dictionary look up method, binary search for the detection of spelling errors in word. They proposed RecursiveSimulation method for generating the appropriate suggestions for the misspelled word. Authors of S11 [24] used n-gram model for checking the correctness of a word in Bangla.

We tabulated our summarized results in Table 4. From our results, methods used for developing Bangla spell checker which we found are: 1) SoundEx encoding, 2) Double Metaphone encoding, 3) Stemming algorithm, 4) Edit distance, 5) Direct dictionary lookup method such as Binary search and 6) N-gram model. Also some other methods are used in spell checking approach such as Finite State Automaton. Moreover, some studies (S6 [19], S10 [23] and S14 [27]) are based on ad hoc method.

B. RQ2: What key challenges are being focused when developing a Bangla spell checker?

Table 5 demonstrates our findings for RQ2. We summarized challenges which were focused on existing approaches. Our results indicated that most of the studies tried to focus on phonetic similarity problem. Some of these studies tried to solve different types of spelling errors such as typographical errors and cognitive phonetic errors. They can correct single spelling error accurately. Most of these approaches focused on non-word errors.

Table 3. Descriptions of Relevant Studies

Study ID	Conference Name/Journal Name	Publication Year	E-library	Contributed Research Question
S1 [14]	LESAL Workshop	2001	Google Scholar	RQ1, RQ2, RQ3
S2 [15]	International Conference on Computer and Information Technology	2004	IEEExplore	RQ1, RQ2, RQ3
S3 [16]	International Conference on Natural Language Processing and Knowledge Engineering	2005	IEEExplore	RQ1, RQ2
S4 [17]	International Conference on Digital Comm. and Computer Applications	2007	Google Scholar	RQ1, RQ2, RQ3
S5 [18]	International Conference on Computer Processing on Bengali	2006	Google Scholar	RQ1, RQ2, RQ3
S6 [19]	International Conference on Computer and Information Technology	2002	ResearchGate	RQ2
S7 [20]	International Conference on Natural Language Processing	2007	Google Scholar	RQ1, RQ2, RQ3
S8 [21]	Language Engineering Conference	2002	IEEExplore	RQ1, RQ2, RQ3
S9 [22]	Software Engineering and Applications	2003	Google Scholar	RQ1, RQ2, RQ3
S10 [23]	Interactive multimedia systems	2002	ResearchGate	
S11 [24]	International Journal of Computer Application	2014	CiteSeerX	RQ1
S12 [25]	International Conference on Computer and Information Technology	2002		RQ3
S13 [26]	International Conference on Information Technology for Applications	2004		RQ3
S14 [27]	International Conference on Computer and Information Technology	2003		RQ1
S15 [28]	Undergraduate thesis, BRAC University	2005	Google Scholar	
S16 [29]	Undergraduate thesis, Department of Computer Science and Engineering, Military Institute of Science and Technology	2014	Google Scholar	
S17 [30]	Undergraduate thesis, BRAC University	2007	Google Scholar	

C. RQ3: What are the limitations of existing research?

Our findings for RQ3 are summarized in Table 6. From this results, we found that most of the studies did not handle all complex rules of Bangla such as different pronunciation of letters or conjuncts in different context. Some studies have lack of efficiency and effectiveness. Some studies focused only phonetic similarity and some

studies focused structural similarity. Some studies considered these two but still failed to provide proper suggestions for misspelled word. Moreover, some of these studies provide absurd results in some cases.

Table 4. Current Methods Used to Develop Bangla Spell Checker

Study ID	Method/Algorithm
S1 [14]	String matching algorithm
S2 [15]	SoundEx encoding
S3 [16]	Double Metaphone encoding
S4 [17]	Stemming algorithm and Edit distance
S5 [18]	Direct dictionary look up method, Double Metaphone encoding and Edit distance
S7 [20]	Finite state automaton
S9 [22]	Direct dictionary look up method and RecursiveSimulation algorithm
S11 [24]	N-gram model
S14 [27]	Edit distance

D. Performance Analysis

Most of the papers did not provide the performance results of their proposed approaches. Table 7 shows the evaluation results of existing approaches. S1 [14] argued that their system works with high accuracy when they used a corpus of 250,000 words as well as it also makes 5% false error detection. S3 [16] used 1607 misspelled words for their evaluation and the result showed that the system can detect 1473 out of 1607 words, therefore accuracy is 91.37%. S4 [17] used 600 root word and 100 suffixes and tested their spelling checker with 13,000 single and multiple spelling error words. Their experiment results showed that correction accuracy for single error misspellings and multiple error misspellings are 90.8% and 67% respectively. S5 [18] evaluated their spelling checker on 1607 misspelled words of Bangla. They claimed that correction accuracy of their system for single error misspellings is 98% and 100% accuracy for correcting 2-error misspellings words for this sample. S7 [20] mentioned that the accuracy for single character error and multiple character errors are 92% and 70% respectively. For evaluating the system, study S11 [24] used 50,000 correct words and another 50,000 incorrect word. The results show that the average performance of the system is 96.17% where the correct words detect as correct at the rate of 95.19% and incorrect words detect as incorrect at the rate of 97.14%. Other studies only provided their proposed approaches and did not provide proper evaluation results. From these above analysis, it can be concluded that most of these approaches provide better results for single spelling errors and still unable to provide better results for multiple spelling errors.

Table 5. Challenges are Focused When Developing Bangla Spell Checker

Study ID	Challenges are Addressed
S1 [14]	Addresses the phonetic problem and solved this by representing phonetically similar vowels and consonants by a single code Can handle phonetically similar characters in Bangla language Can detect any error and correct single error
S2 [15]	Mentions the phonetic problem and solves only the first two challenge we mentioned in section II(A)
S3 [16]	Focuses on the most of the challenges of the complex spelling rules for Bangla Concentrates on context-sensitive rules of Bangla
S4 [17]	Aims at handling inflection related word Follows complex orthographic rules of Bangla
S5 [18]	Concentrates on detection of typographical errors and cognitive phonetic errors
S6 [19]	Focuses on proper coding system to handle replacement errors, deletion errors, insertion errors and transposition errors
S7 [20]	Touches the detection of non-word errors Focused on detection of substitution errors, insertion errors and errors
S9 [22]	Focuses on detection of typographical errors and cognitive phonetic errors
S12 [25]	Addresses phonetic problem and deals with phonetic similar characters and some trivial cases of Bangla
S13 [26]	Concentrates on detection of typographical errors and cognitive phonetic errors Deals with the first two challenge we mentioned in Section II(A)

VI. Discussion

This study covers gist of existing works, the novelty and shortcomings of proposed approaches and scope of further works. To answer the RQ1, it has been shown that there are many current methods used in developing spell checker for Bangla. However, selecting appropriate methods is very important for achieving high performance and effectiveness. Results for RQ2 and RQ3 show that there are few challenges that have been addressed and there are many limitations of existing works. Although we found various approaches used for spell checking, no approach considered all of the complex rules of Bangla. The date format of Bangla is also not mentioned by any existing approach. None of these existing approach can handle real word errors. Moreover, context-sensitive rules of Bangla are not focused in most of the studies.

It is clear from the literature that there is a need for a comprehensive approach for Bangla spell checking that will address all of these challenges and ensure high performance as well as effectiveness. Besides, almost all

of the studies did not discuss about data structures in which they stored the dictionary. Storing all dictionary words in a proper way helps to reduce search space and time. Therefore, it is actually very important to develop a proper data structure for storing all words of dictionary.

Table 6. Limitations of Existing Research

Study ID	Limitations
S1 [14]	Always needs double amount of memory space because of one dictionary and its reverse dictionary Does not deal with challenges discussed in section II(A) apart from the first two
S2 [15]	Does not handle challenges we discussed in section II(A) except the first two
S4 [17]	Unable to handle derivational suffixes, for which a proper morphological analyzer is required Does not consider phonetic, typographical, or other types of errors
S5 [18]	Fails frequently to suggest for multiple errors in both root and postfix of a word Generates large number of suggestions for detectable multiple error circumstances
S7 [20]	Fails to handle transposition errors as a single edit distance errors Does not handle the more unusual cases
S9 [22]	Algorithm is not fast and inefficient and has many bugs
S12 [25]	Cannot handle unusual pronunciation of many clusters or conjuncts, different uses of Phalaa's, different pronunciation of letters or conjuncts in different context and multiple pronunciations of some letters in the same context
S13 [26]	Does not consider the full phonetic complexity of Bangla orthographic rules

Table 7. Evaluation Results of Existing Approaches

Study ID	Input Data	Accuracy
S1 [14]	25,000 words	high accuracy with 5% false positive detection
S3 [16]	1607 words	91.37%
S4 [17]	13,000 words	90.8% for correcting single error misspellings and 67% for correcting multiple error misspellings
S5 [18]	1607 words	98% for correcting single error misspellings and 100% for correcting 2-error misspellings
S7 [20]	291 words	92% for correcting single character misspellings and 70% for correcting multiple character misspellings
S11 [24]	50,000 correct words and 50,000 incorrect words	96.17% where correct words detection rate 95.19% and incorrect words detection rate 97.14%

VII. Conclusion

This paper shows the findings of a systematic literature review on checking and detecting Bangla spelling errors. Since there are few works on Bangla spell checker, we investigated a little amount of studies. We presented a complete description of our systematic literature review process with findings and discussion about these findings. We discussed the current status of this research topic so that interested researchers can be benefited from this work. Our findings indicate that develop an approach for Bangla spell checker which would address all challenges of complex rules of Bangla is very essential.

As future work, we will further investigate how we can achieve high performance and effectiveness for spell checking in Bangla. In future, we will also design and implement a new spell checker for Bangla language.

References

[1] Keele, Staffs. "Guidelines for Performing Systematic Literature Reviews in Software Engineering." Technical report, Ver. 2.3 EBSE Technical Report. EBSE. 2007.

[2] Kukich, Karen. "Techniques for Automatically Correcting Words in Text." ACM Computing Surveys (CSUR) 24.4 (1992): 377-439.

[3] P. Kundu and B.B. Chaudhuri (1999) "Error Pattern in Bangla Text". International Journal of Dravidian Linguistics. 28(2): 49-88.

[4] D. E. Knuth, The Art of Computer Programming, Vol. 3, Addison-Wesley Publishing Company, Reading, Massachusetts, 2nd edition, 1982.

[5] Lawrence Phillips, "Hanging on the Metaphone", Computer Language, 7(12), 1990.

[6] Lawrence Phillips, "The Double Metaphone Search Algorithm", C/C++ Users Journal, 18(6), June, 2000.

[7] T. N. Gadd, "PHONIX: The Algorithm", Program, 24(4), pp. 363-366, 1990.

[8] Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions, and Reversals. 10(8), 707–710.

[9] W. Kraaij and R. Pohlman, "Viewing Stemming as Recall Enhancement", In the Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1996, pp. 40–48.

[10] Lovins, Julie Beth (1968). "Development of a Stemming Algorithm". Mechanical Translation and Computational Linguistics 11: 22–31.

[11] Porter, Martin F.1980. An Algorithm for Suffix Stripping. Program 14 (3): 130-137.

[12] C. E. Shannon, "Prediction and Entropy of Printed English," Bell Sys. Tec. J. (30):50–64, 1951.

[13] Farag Ahmed, Ernesto William De Luca, and Andreas Nürnberger, "Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness", August 22, 2009.

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

[15] Naushad UzZaman and Mumit Khan, "A Bangla Phonetic Encoding for Better Spelling Suggestions", Proc. 7th International Conference on Computer and Information Technology, Dhaka, December, 2004.

[16] UzZaman, Naushad, and Mumit Khan. "A Double Metaphone Encoding for Bangla and its Application in Spelling Checker." 2005 International Conference on Natural Language Processing and Knowledge Engineering. IEEE, 2005.

[17] Islam, Md, Md 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.
[18] N. UzZaman and M. Khan, “A Comprehensive Bangla Spelling Checker”, In the Proceeding of the International Conference on Computer Processing on Bengali (ICCPB), Dhaka, Bangladesh, 2006.
[19] Hoque, Md Tamjidul, and Md Kaykobad. "Coding System for Bangla Spell Checker." 5th International Conference on Computer and Information Technology. 2002.
[20] 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.
[21] Chaudhuri, Bidyut Baran. "Towards Indian Language Spell-checker Design." Language Engineering Conference, 2002. Proceedings. IEEE, 2002.
[22] 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.
[23] Murshed, M. Manzur, Mahbubur Rahman Syed, and M. Kaykobad. "A Linguistically Sortable Bengali Coding System and its Application in Spell Checking: A Case Study of Multilingual Applications." Interactive multimedia systems (2002): 251.
[24] Khan, Nur Hossain, et al. "Checking the Correctness of Bangla Words using N-Gram." International Journal of Computer Application 89.11 (2014).
[25] Haque, Md Tamjidul, and M. Kaykobad. "Use of Phonetic Similarity for Bangla Spell Checker." Proc. 5th International Conference on Computer and Information Technology. 2002.
[26] Abdullah, A. B. A., and Ashfaq Rahman. "A Different Approach in Spell Checking for South Asian Languages." Proc. 2nd International Conference on Information Technology for Applications (ICITA), China. 2004.
[27] Abdullah, Arif Billah Al-Mahmud, and Ashfaq Rahman. "Spell Checker for Bangla Language: An Implementation Perspective." Proc. 6th International Conference on Computer and Information Technology, Dhaka, Bangladesh. 2003.
[28] UzZaman, Naushad. "Phonetic Encoding for Bangla and its Application to Spelling Checker, Name Searching, Transliteration and Cross Language Information Retrieval." Undergraduate thesis (Computer Science), BRAC University (2005).
[29] Bhowmik, Kowshik, Afsana Zarin Chowdhury, and Sushmita Mondal. Development of A Word Based Spell Checker for Bangla Language. Diss. Department of Computer Science and Engineering, Military Institute of Science and Technology, 2014.
[30] Asadullah, Munshi. Finite State Recognizer and String Similarity based Spelling Checker for Bangla. Diss. BRAC University, 2007.
[31] Govilkar, Sharvari S., J. W. Bakal, and Sagar R. Kulkarni. "Extraction of Root Words using Morphological Analyzer for Devanagari Script." International Journal of Information Technology and Computer Science (IJITCS) 8.1 (2016): 33.
[32] Aktaruzzaman, Md, and Md Farukuzzaman Khan. "A New Technique for Segmentation of Handwritten Numerical Strings of Bangla Language." International Journal of Information Technology and Computer Science (IJITCS) 5.5 (2013): 38.
[33] Doumi, Noureddine, et al. "A Semi-Automatic and Low Cost Approach to Build Scalable Lemma-based Lexical Resources for Arabic Verbs." International Journal of Information Technology and Computer Science (IJITCS) 8.2 (2016): 1.
[34] Divya, K. S., R. Subha, and S. Palaniswami. "Similar Words Identification Using Naive and TF-IDF Method." International Journal of Information Technology and Computer Science (IJITCS) 6.11 (2014): 42.

Authors’ Profiles

Prianka Mandal is a graduate student at the Institute of Information Technology (IIT), University of Dhaka, Bangladesh. Currently, she is pursuing her Master of Science in Software Engineering (MSSE). She earned her Bachelor of Science in Software Engineering (BSSE) from the same institution. Her core areas of interest are software engineering and natural language processing.

Dr. B. M. Mainul Hossain is Assistant Professor at the Institute of Information Technology (IIT), University of Dhaka, Bangladesh. He received his Ph.D. degree in computer science from University of Illinois at Chicago, USA. Before that, he earned his Bachelor of Science and Master degrees from the department of Computer Science & Engineering, University of Dhaka, Bangladesh. He has the experiences of working both in industry and academia. He worked as a Software Engineer in Microsoft Corporation (Redmond, USA) & Accenture Technology Lab (Chicago & California). His core areas of interest are software engineering, security, data mining and machine learning.

How to cite this paper: Prianka Mandal, B M Mainul Hossain,"A Systematic Literature Review on Spell Checkers for Bangla Language", International Journal of Modern Education and Computer Science(IJMECS), Vol.9, No.6, pp.40-47, 2017.DOI: 10.5815/ijmecs.2017.06.06

