IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-7, Issue-6, June 2019 E-ISSN: 2347-2693

Detection and correcting the wrong words from Hindi, English and Punjabi Text Documents

Shaina[1*], Naresh Kumar[2]

[1,2]Department of Computer Science and Engineering, GZS Campus, Bathinda, India

Corresponding Author: cse2011shaina@gmail.com

DOI: https://doi.org/10.26438/ijcse/v7i6.314318 | Available online at: www.ijcseonline.org

Accepted: 11/Jun/2019, Published: 30/Jun/2019

***Abstract*—**Spell checking is a very important phase of any document processing system and Natural Language Processing. Spell Checking is a process to find the incorrect spells in a text document and to correct that particular incorrect spelling. There are various spell checking systems for various languages Like Hindi, Punjabi, English, French, Germen that can detect and correct the spell from a particular document. In this paper, we proposed a hybrid algorithm to detect and correct misspelled words from a text document written in three languages Hindi, English and Punjabi. Hybrid approach is a combination of various approaches like Dictionary lookup approach, Edit Distance Approach, Rule based approach and N-Gram approach. Proposed system can detect and correct the misspelled words from three given languages. A collision detection and correction system for alternates for misspell words has been also provided. Performance of proposed system is checked on various inputs collected from various books, websites etc. Results of the proposed system are evaluated on these outputs which have accuracy values higher than that of existing system.

Keywords— Spell Checking; Hybrid approach for Spell Checking; N-Gram Approach; Rule Based Approach; Edit distance approach.

I. INTRODUCTION

Spell-checking is the process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. Spell checking system can be created with the combination of handcrafted rules by considering grammatical features of the language for which spell checking system is to be created and a dictionary which contain the accurate spellings of various words in the target language. Basically, the better the handcrafted rule and larger the dictionary of a spell-checker, the higher is the error detection rate; otherwise, misspellings would pass undetected. Unfortunately, traditional dictionaries suffer from out-of-vocabulary and data sparseness problems as they do not encompass large vocabulary of words indispensable to cover proper names, domain-specific terms, technical jargons, special acronyms, and terminologies. As a result, spell-checkers will incur low error detection and correction rate and will fail to flag all errors in the text. All modern commercial spelling error detection and correction tools work on word level and use a dictionary. Every word from the text is looked up in the speller lexicon. When a word is not in the dictionary, it is detected as an error. In order to correct the error, a spell checker searches the dictionary for words that resemble the erroneous word most. These words are then suggested to the user who chooses the word that was intended. Spelling checking is used in various applications like machine translation, searches, information retrieval etc. There are two main issue related to spell checker. These are error detection and error correction. In developing, upon the type of error non word error and real word error. There are many techniques available for detection and correction. Spell checker can also be defined as it is a supercomputer application that analysis possible misspelling in a text by referring to the accepted spellings in a database. In the database various accurate words of the target language for which the spell – checker is to be made are stored which consists of proper nouns for males, females, countries, states, rivers, mountains etc. The system is made to check the spellings and to correct them using various techniques for Punjabi, English and Hindi text. In this proposed system input in form of a paragraph is given that can include incorrect words and the system will generate the result which contain the accurate text after eliminating the errors. We will use hybrid approach to implement the Spelling checking and Correcting System. This hybrid approach is a combination of "Dictionary look up approach", "Rule based approach" and "Edit Distance approach" and use linguistic features of the Punjabi, English and Hindi language.

II. LITERATURE SURVEY

R. Mishra[1], describes the development and working of online Raftaar Punjabi spell checker and also developed a proposed algorithm for the correction of wrong words, This System gives the result accuracy as 80% according to the research work for Punjabi words. It gives nearby result up to 80% of words tested in this thesis. It gives results for rest of 20% but not the best possible correct word was displayed on the top of the correct word list from the database.

N. Gupta [2], describes the various techniques for spell checking and error correction. This paper also provides information about various available spell checking systems developed for various Indian language. In this paper two techniques for spell checking are described which are (1) N Gram Analysis based on statistical technique and (2) is Dictionary lookups. This paper describes the properties of various spell checker and spell Corrector, these systems includes Bangla spell Checker, Oriya Spell Checker , Tamil spell Checker, Marathi spell checker, Punjabi spell checker etc. Techniques described in this paper for spelling error correction includes "Edit distance", "similarity keys", "Rule Based Techniques", "N-Gram based techniques", "Neural Network based techniques etc.

B. Kaur[3], have surveyed the area of spell correction and error detection techniques. Existing work related with spell checkers in Punjabi and Punjabi language is also discussed. In this paper the author will implement a Punjabi spell-checker by using dictionary lookup and edit-distance based technique with more accuracy. In this paper techniques for Error Correction are used (1) N Gram Analysis (2) Rule Based Approach and (3) Edit Distance.

N. Gupta[4], Spell checkers in Indian languages are the basic tools that need to be developed. A spell checker is a software tool that identifies and corrects any spelling mistakes in a text. Spell checkers can be combined with other applications or they can be distributed individually. In this paper the authors are discussing both the approaches and their roles in various applications. In this paper they have surveyed the area of Spell checking techniques. They have discussed various detection and correction techniques that are useful in finding the text with error. In future they will implement algorithm that is based on dictionary lookup techniques for detection and minimum edit distance techniques for correction of result in the area of Indian language spell checking.

Y. Bassil[5], Spell-checking is the process of detecting and sometimes providing suggestions for incorrectly spelled words in a text. Basically, the larger the dictionary of a spell-checker is, the higher is the error detection rate; otherwise, misspellings would pass undetected. Alas, traditional dictionaries experience from out-of-vocabulary and data sparseness problems as they do not include large vocabulary of words essential to cover proper names, domain-specific terms, technical jumbos, special acronyms, and terminologies. As a result, spell-checkers will encounter less error detection and correction rate and will fail to flag all errors in the text. This paper proposes a new parallel shared-memory spell-checking algorithm that uses rich real-world word statistics from Yahoo! N-Grams Dataset to correct non-word and real-word errors in computer text. Essentially, the proposed algorithm can be divided into three sub-algorithms that run in a parallel fashion: The error detection algorithm that detects misspellings, the candidates generation algorithm that generates correction suggestions, and the error correction algorithm that performs contextual error correction. Experiments conducted on a set of text articles containing misspellings, showed a remarkable spelling error correction rate that resulted in a radical reduction of both non-word and real-word errors in electronic text. In an additional study, the planned algorithm is to be improved for message-passing systems so as to become more supple and inexpensive to scale over distributed machines. This paper presented a novel shared-memory parallel spell-checking algorithm for detecting and correcting spelling errors in computer text. The proposed algorithm is based on Yahoo! N-Grams Dataset that comprises trillions of word sequences and n-grams, originally extracted from the World Wide Web. When experimented to correct misspellings in 300,000-word articles, the proposed algorithm outclassed other existing spell-checkers as it effectively corrected 94% of the total errors, distributed as 99% non-word errors and 65% real-word errors. On the other hand, the Hunspell spell-checker managed to correct 66% of total errors; while, the Ginger spell-checker was able of 78% of total errors. In sum, the error correction rate for the proposed algorithm was 16% higher than Ginger and 28% higher than Hunspell. The major reason behind these outstanding results is the use of Yahoo! N-Grams Dataset as a dictionary model which delivers wide-ranging set of words and n-gram statistics that cover domain-specific terms, technical jargons, proper names, special acronyms, and most of the words that possibly can occur in a text.

III. EXISTING TECHNIQUES

A. Dictionary lookup approach

Dictionary lookup approach is used to check whether a particular word is spelled correctly or not just by comparing that word with the database. In this approach a text paragraph is tokenized into words and each word in the Document which will be given as an input is checked for the database entries. If the scanned word is found in the database then is considered to be correct word i.e. spellings of the word are correct but in case the word is not present in the database table then it is considered as an incorrect word. After finding the word incorrect various handcrafted rules are applied to generate the correct spellings of the word by considering the linguistic features of the Punjabi language, if approach generate the multiple entries for the single entry then by using statistical analysis a more appropriate word id chosen by the system and is replaced with the incorrect word to generate the result.

B. Edit Distance

Edit distance is a simple technique in spell correction. This Simplest method is based on the assumption that the person usually makes few errors if ones, therefore for each

dictionary word .the minimal number of the basic editing operations (insertion, deletions, substitutions) necessary to convert a dictionary word in to the non-word .the lower, the number ,the higher the probability that the user has made such errors. Through the operation of adding, deleting and modifying, Edit-Distance changes a word into the minimum operating frequency of another word.

C. Rule based Approach

In this approach handcrafted rules are made by considering the features of the type of input language. These rules are applied on the words in the paragraph which are not found in the database. With the help of these rules the system tries to generate the correct spellings of the word which is under observation.

IV. PROPOSED METHODOLOGY

Proposed system use hybrid approach to implement the Spelling checking and Correcting System. This hybrid approach is a combination of "Dictionary look up approach", "Rule based approach" and "Edit Distance approach" and use linguistic features of all Punjabi, English and Hindi language. Dictionary Look-Up Approach and Edit Distance Approach is used in the research are already implemented.

Following are the steps of proposed algorithm:
Step I: Input the source string.
Step II: Tokenize the input of first step into words.
Step III: Convert the tokenized word into the array of characters.
Step IV: Evaluate the language of the tokenized word among the three languages.
Step V: Compare the tokenized word with the corresponding dictionary.
Step VI: Check whether it is correct or not. If it is correct, then go to Step IV, otherwise apply Rule Bases Approach.
Step VII: Again find the word from dictionary. If word is found go to Step IV, otherwise apply Edit Distance Approach.
Step VIII: Find the minimum distance from this Token to the word in the Dictionary.
Step IX: Sort these words in ascending order of their distance.
Step X: If distance of topmost words is same then assign the maximum weight to the word having close semantic meaning according to the line using N-Gram Approach.
Step XI: Replace the top most word obtained in step VII with token
Step XII: End.

Rule Based Approach

In this approach a set of rules are developed by which the input token is compared. Every rule created in the rule based is applied on the token and corresponding results is produced after applying the rule based approach.
The following are the steps of rule based approach:
Step I : input the Text to find the errors
Step II: Tokenize the input of first step into words.
Step III For each Token compare it with the Dictionary.
Step IV Check whether it is correct or not. If it is correct, then go to Step III, otherwise apply Rule Based Approach.
Step v: use output of rule based system into next phase
Step VI: end

Dictionary lookup technique

This approach is mainly used to check whether the particular token is correct or not by comparing the token with the dictionary values. It is assumed that the word which is being checked is correct if it is available in the dictionary. To create dictionary for various input words, various resources like text books, online websites are being used. The accuracy of the system is highly depends upon this phase. If the required word is correct but not in the dictionary then it will give wrong output.
Steps for dictionary look up technique are as follows:
Step I : input the String data to be checked
Step II : Tokanize this input into words
Step III Compare it with the dictionary to check whether it is correct or not.
Step IV End
Edit Distance Technique
This technique will work if rule based approach becomes unable to generate the accurate word. This technique is used to find the nearest possible word from the dictionary to obtain the result. With the help of this technique various suggestions are generated with respect to the token which is being checked in the ascending order of their distances. In this approach, the word distance means the minimum number of operations required to equate the wrong word with the word in dictionary.

The steps to implement this technique are as follows:
Step I: Input the text string
Step[II: Tokenize the input string
Step III : for each word in the dictionary perform following steps IV and V
Step IV : calculate the distance of word from step III with input token
Step V Store the word and token in the temp location and ignore if distance is more than 3.
Step VI : Sort the words obtained in step V in ascending order and display it to the user.
Step VII End

N-Gram Approach

This approach is used to remove the ambiguity between the generated words by the other approaches. When top most generated words have the same distance than that of original word then it is said that ambiguity occurs. N-Gram approach is used to remove this ambiguity by comparing the words along with their previous and next words with the paragraph stored into the database. If the combination of these words

found in the stored paragraph then max weight is assigned to the word that occurs in that combination and that word is moved onto top.

The following are the steps of the N-Gram approach:

Step I: Check the top most words from the options generated.

Step II : Compare the distances of these topmost words.

Step III : if the distance is different then go to step VIII.

Step IV: generate the combination of previous, current, and next word.

Step V : Find this combination in the database of paragraphs.

Step VI: If combination is found then increase the weight of the current word.

Step VII: Display the word at the top having maximum weight.

Step VIII: End

V. RESULTS AND DISCUSSION

Proposed system is implemented using ASP.Net and C#.Net with MS-ACCESS Database. In the proposed system a language selection option is provided from which user can select whether to perform the spell checking in Punjabi, English or Hindi language.

An independent dictionary is created for Punjabi, English and Hindi language that containing more than 50,000 words for both the language.

We have performed various test cases on the system with Hindi, English and Punjabi Language documents collected from various online websites and E-Books. Results are calculated in form of error corrected by the proposed system. The following are the tables and graphs calculated from our proposed system on various inputs representing the accuracy of the overall system.

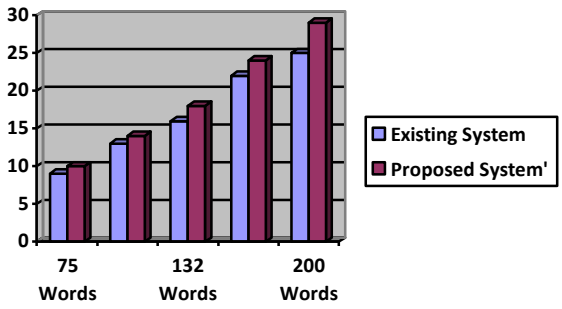
Our System gives more accuracy than that of existing system because our system works independently on every word of the input. So in case the input contains the words more than one language, then existing system can't correct them, whereas our system will correct it.

Also our purposed system has large database then the existing system. And we know that the accuracy of spell checker system is highly dependent on the size of database. Hence our purposed system gives more accuracy than that of existing system.

The following table is presented that represents the comparison on the basis of accuracy of proposed and existing system:

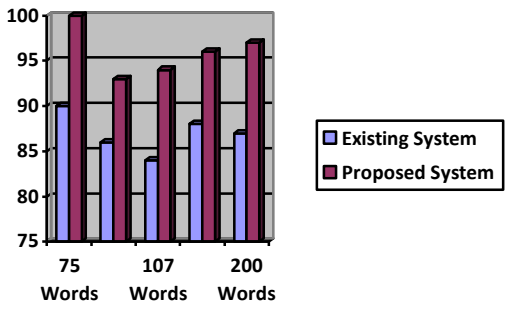
S. No.	Total No. of words in paragraph	Errors in Paragraphs	Correction by Edit Distance Technique	Accuracy of Edit Distance Technique	Errors Corrected by Proposed System	Accuracy of Proposed System
1	75	10	9	90%	10	100%
2	107	15	13	86%	14	93%
3	132	19	16	84%	18	94%
4	150	25	22	88%	24	96%
5	200	30	25	87%	29	97%

The following graph is showing the comparison of existing and proposed system:



Here this graph represents the corrected error by existing system and proposed system. It shows from 75 words system detects 10 error words and existing system corrects 9 and proposed system corrects 10 words. In next input from 107 words there are 15 error words, existing system corrects the 13 words whereas proposed system corrects 14 words and so on.

Accuracy Comparison Graph of existing and proposed system



This graph shows the accuracy of existing system and proposed system. From the input of 75 words existing system gives 90 % accuracy whereas proposed system gives 100% accuracy and for 107 words existing system gives 86% and proposed system gives 95% accuracy and so on.

VI. CONLUSION AND FUTURE SCOPE

In our Research work, we have developed an online Punjabi, English and Hindi spell checker and also developed a new proposed algorithm for the correction of wrong words according to the dictionary. Proposed System is based on hybrid approach in which four approaches which are rule based approach, dictionary look up approach, edit distance and N-Gram approaches are used. Various test cases have been performed to measure the performance of the proposed system. Overall accuracy of the proposed system in terms of misspelled words correction is 98% which is much higher than that of existing systems.

In future, more unique words can be added to the dictionary to improve the overall performance of the system. Further new rules which are both specific to Punjabi, English and Hindi are to be added to increase the performance of the proposed system.

REFERENCES

[1] Ritika Mishra, Navjot Kaur, Design and Implementation of Online Punjabi Spell Checker Based on Dynamic Programming, Volume 3, Issue 8, August 2013, ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering

[2] Neha Gupta, Pratistha Mathur, Spell Checking Techniques in NLP: A Survey, Volume 2, Issue 12, December 2012 , ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering

[3] Baljeet Kaur, Review On Error Detection and Error Correction Techniques in NLP: Volume 4, Issue 6, June 2014 ISSN: 2277 128X, International Journal of Advanced Research in Computer Science and Software Engineering.

[4] Rupinderdeep Kaur and Parteek Bhatia, "Design and Implementation of SUDHAAR-Punjabi Spell Checker," International Journal of Information and Telecommunication Technology, Vol. 1, Issue 15 May, 2010.

[5] S. Dasgupta, C.H. Papadimitriou, and U.V. Vazirani, 'Algorithms', p173, available at http:/ / www.cs.berkeley.edu/ ~vazirani/ algorithms.html.

[6] Neha Gupta &PratisthaMathur,"Spell Checking Techniques in NLP: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 12, December 2012.

[7] Gurpreet Singh Lehal, "Design and Implementation of Punjabi Spell Checker", International Journal of Systemics, Cybemetics and Infomatics, 2007.

[8] Amit Sharma & Pulkit Jain, "Hindi Spell Checker", Indian Institute of Technology Kanpur, April 17, 2013.

[9] MeenuBhagat, (2007), "Spelling Error Pattern Analysis of Punjabi Typed Text", Thesis Report, Thapar University, Patiala.

[10] F.J. Damerau (1964), "A Technique for Error Detection and Correction of Spelling Errors", Communication ACM, pp. 171-176.

[11] Monisha Das, S. Borgohain, JuliGogoi, S. B. Nair (2002), "Design and Implementation of a Spell Checker for Assamese",lec, pp. 156, Language Engineering Conference (LEC'02).

[12] Morris, Robert & Cherry, Lorinda L, "Computer Detection of typographic errors", IEEE Trans Professional Communications, vol. PC-18, no. 1, pp 54-64, March 1975.

[13] R.E. Gorin (1971), "SPELL: A spelling checking and correction program", Online documentation for the DEC-10 computer.

[14] K. Kukich (1992) "Techniques for automatically correcting words in text". ACM Computing Surveys. 24(4): 377-439.

[15] Peterson James (1980), "Computer Programs for Detecting and Correcting Spelling Errors", Computing Practices, Communications of the ACM.

[16] G S Lehal & MeenuBhagat, "Spelling Error Pattern Analysis of Punjabi Typed Text", In Proceedings of International Symposum on Machine Translation, NLP and TSS, pp. 128-141, 2007.

[17] Jesus Vilares& Manuel Vilares, "Managing Misspelled Queries in IR Application," Issue 8, October 2010.

[18] Youssef Bassil& Mohammad Alwani, "Context-sensitive Spelling Correction using Google Web IT 5-Gram Information," Department of Computer and Information Science, Vol. 5,No.3, May 2012.G. Eason, B. Noble, and I.N. Sneddon, "On certain integrals of Lipschitz-Hankel type involving products of Bessel functions," Phil. Trans. Roy. Soc. London, vol. A247, pp. 529-551, April 1955.

