ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Spelling Error Detection and Correction System for Punjabi Unicode using Hybrid Approach

Rupinder Kaur, Er. Jasdeep Singh Mann

M.Tech Student, Dept. of C.S.E, BMSCE, Sri Muktsar Sahib, Punjab, India

Assistant Professor, Dept. of C.S.E, BMSCE, Sri Muktsar Sahib, Punjab, India

ABSTRACT: Spelling error detection and correction is an important tool for any language to remove the errors from a text. In this paper, we have developed and discussed a novel approach to find the errors and to correct them automatically from a Punjabi Unicode text. Proposed system use hybrid approach to perform the required task. This hybrid approach is a combination of Dictionary lookup approach, rule based approach, Improved edit distance approach and N-gram approach. Proposed system is evaluated on various inputs collected from various online sources and results are discussed accordingly. Proposed system gives the average accuracy of 97% which is very good as compared to existing systems.

KEYWORDS: Punjabi Unicode spell checking, Improved Edit distance approach, Rule based approach, N-Gram Approach, Spelling checking.

I. INTRODUCTION

Spell-error detection and correction is the process of detecting and correcting the misspelled words from a given text for a particular language. Spell checking framework can be made with the mix of carefully assembled leads by considering linguistic components of the language for which spell checking framework is to be made and a lexicon which contain the precise spellings of different words in the objective language. Essentially, the better the handmade rukes and bigger the word reference of a spell-checker is, the higher is the error identification rate; generally, incorrect spellings would pass undetected. Lamentably, conventional lexicons experience the ill effects of out-of-vocabulary and information inadequacy issues as they don't include expansive vocabulary of words key to cover legitimate names, area particular terms, specialized languages, uncommon acronyms, and wordings. As a result, spell-checkers will incur low error detection and correction rate and will fail to flag all errors in the text. All current business spelling error detection and correction instruments chip away at word level and utilize a lexicon. Each word from the content is gazed upward in the speller vocabulary. At the point when a word is not in the lexicon, it is distinguished as a error. Keeping in mind the end goal to remedy the mistake, a spell checker scans the lexicon for words that look like the mistaken word most. These words are then proposed to the client who picks the word that was planned. Spelling checking is utilized as a part of different applications like machine interpretation, seeks, data recovery and so on. There are two fundamental issue identified with spell checker. These are mistake recognition and error amendment 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 Unicode 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

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

approach", "Rule based approach", "N-Gram Approach", "Improved Edit Distance approach" and use linguistic features of the Punjabi language.

II. LITERATURE SURVEY

[1] Amanjot kaur et al., In this paper author describe that the system is made to check the spellings and to correct them using various techniques for Punjabi text. We used hybrid approach to implement the Spelling checking and Correcting System. This hybrid approach is a combination of "Dictionary look up approach", "Rule based approach", "Statistical Approach", "Edit Distance approach" and use linguistic features of the Punjabi language. This System gives the result accuracy as 91% according to the research work for Punjabi words. It gives approximate result up to 91% of words tested in the input data. It gives results for rest of 9% but not the best possible correct word was displayed on the top of the correct word list from the database.

[2] Neha Gupta et al.,In this paper author 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.

[3] Baljeet kaur et al., In this paper, author describes that A spell checker is an application program that flags words in a document that may not be spelled correctly. A spell checker is a basic need of a word processor of any language. Spell checker analyzes the written text in order to identify any misspellings and gives best correct suggestions for those misspellings. Most of work has been done in English and Punjabi language. Hindi is the third most spoken language in the world. In This paper the design, techniques and implementation of the Hindi spell checker is proposed. Error detection, Error correction by generating suggestions and replacement are the main features of this system. The system detects approximately 83.2% of the errors and provides 77.9% of the correct suggestions for the misspelled words.

[4] Jaspreet Kaur et al., Spell checker is a tool used for checking the spelling errors and also correcting those errors in the text or a document. Grammar checker is a program which is used to verify the grammatical errors in the text. Developing a spell checker and grammar checker for Indian languages such as Punjabi raises many new difficulties which are not in English, which makes the design of spell checker and grammar checker very difficult. The main difficulties faced are- there is no basic layout of Punjabi keyboard and also there is no approved format for Punjabi spellings. There are lots of differences in grammatical properties of Punjabi that makes it different from other languages. Punjabi is the world's 12th most widely spoken language. The very first requirement for developing any sp-ell checker is to have dictionary of different words of that language which will work as lexicon. The paper aims to develop a system which is hybrid combination of spell checker and grammar checker for Punjabi. Firstly the system checks for spelling errors, then checks for grammatical errors in the text. When some input text is given to system, it is passed through spelling checker and then grammar checker. A comparative, efficient algorithm has been proposed which is hybrid combination of spell checker and grammar checker for Punjabi which saves time and cost.

III. RESEARCH METHODOLOGY

Proposed Methodology for spelling error detection and correction uses various approaches like dictionary lookup approach (for detecting the misspelled words), Improved Edit Distance Approach (to generate the suggestions for misspelled words) and N-Gram Approach (to select the best suggestion among all the generated suggestions).

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

A. *Algorithm Steps for the proposed system are as below:*
Step I: Input the Punjabi Unicode text that can contain the miss-spelled words.
Step II: Tokenize the Punjabi Unicode text into words.
Step III For each Tokenized word of Punjabi Unicode Check if the word is spelled correct or not by comparing it with the available corpus of Punjabi Unicode.
Step IV If the scanned word is correct then go to the next Punjabi Unicode word otherwise apply Rule based approach to make it correct.
Step V: After correction is made again check that Unicode word that if it is correct or not by comparing it with the available corpus of Punjabi Unicode. If it is found correct then go to next word otherwise apply Improved Edit Distance technique to generate the suggestions and go to next step.
Step VI Sort these suggestions in ascending order of their distance.
Step VII If distance of topmost words are same then assign the maximum weight to the word having close semantic meaning according to the line using N-Gram Approach.
Step VIII. Replace the top most word obtained in step VII with mis-spelled word.
Step IX End.

B. *Dictionary Look Up Approach:*
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 Punjabi words, various resources like Punjabi text books, online Punjabi 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.

C. *Improved Edit Distance Technique:*
Improved edit distance technique is used to generate the suggestions for mis-spelled words. This technique is different from existing technique in a way that it generate the distance (that represents the difference between the two words) by not only comparing the contents of the words but also comparing the length and other features such as first and last characters of the word to generate the appropriate output whereas exiting technique generate the output only based on contents of the words.
The steps to implement this technique are as follows:
Step I: Input the misspelled word extracted from the previous phase.
Step II: for each word in the corpus calculate the distance of mis-spelled word from step I by comparing the contents, length and other features of the words.
Step III Store the word and token in the temp location and ignore if distance is more than 3.
Step IV: Sort the words obtained in step V in ascending order and display it to the user.
Step V: End

D. *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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

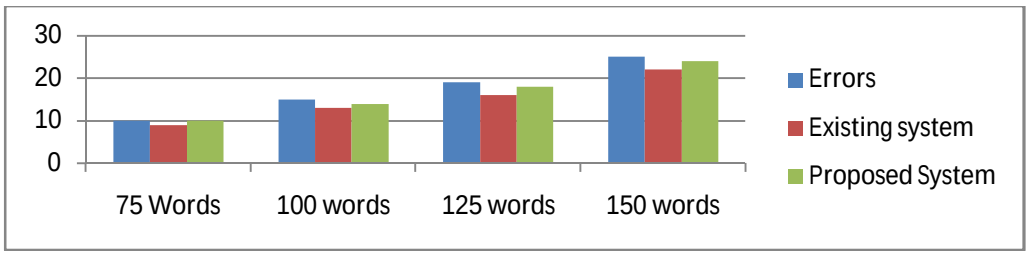
Step VIII: End

IV. Results and Discussion

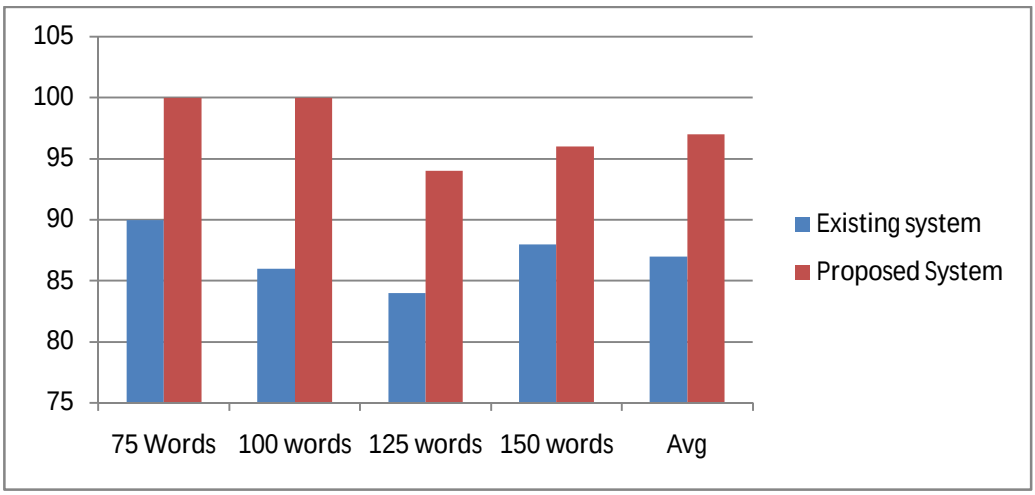
Proposed system is tested on more than 25 Punjabi Unicode text paragraphs statistics of which are represented as follows:

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	100	15	13	86%	14	95%
3	125	19	16	84%	18	96%
4	150	25	22	88%	24	96%
	Avg.	69	60	87%	66	97%

Following graph is showing the comparison of existing and proposed system:



Accuracy Comparison of existing and proposed system:



ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

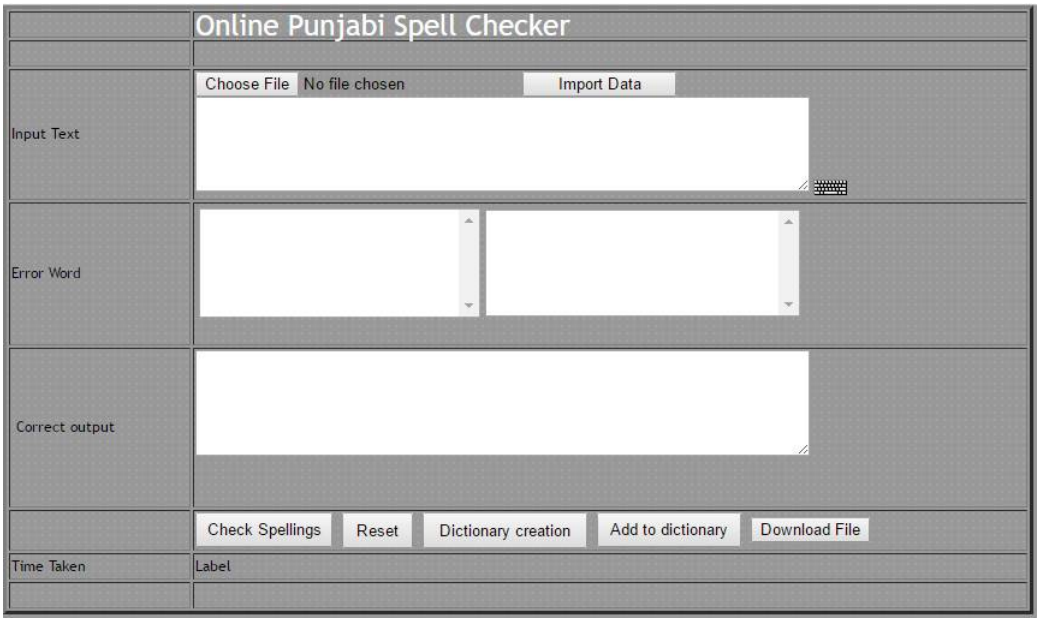
International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

Snapshot of the proposed system:



V. CONCLUSION AND FUTURE SCOPE

In our Research work, we have developed an online Punjabi 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 three approaches which are rule based approach, dictionary look up approach and Improved edit distance approaches are used into one. The main features of Punjabi spell checker are large database, online application, easy to operate, email and printing options. This System gives the result accuracy as 97% according to the research work for Punjabi words. It gives nearby result up to 97% of words tested in this minor project. It gives results for rest of 3% but not the best possible correct word was displayed on the top of the correct word list from the database. In this Research work, the word is not given the highlighter for wrong words. The future scope for this project as the words highlighted with red highlighter which are not correct according to the dictionary. For further research, some grammatical rules like the combinations of noun, verb, and adverb may be added. In future more databases can be added to the system to improve overall accuracy.

REFERENCES

1. Amanjot Kaur, Dr. Paramjeet Singh, Dr. Shaveta Rani,"Spellchecking and Error Correcting System for text paragraphs written in Punjabi Language using Hybrid approach",International Journal of Advanced Research in Science, Engineering and Technology, Vol. 2, Issue 11 , November 2015.
2. Neha Gupta, Pratistha Mathur, "Spell Checking Techniques in NLP: A Survey",International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 12, December 2012 , ISSN: 2277 128X.
3. Baljeet kaur, Harsharndeep Singh, "Design and Implementation of HINSPELL -Hindi Spell Checker using Hybrid approach",International Journal of scientific research and management (IJSRM), Volume-3,Issue-2.
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.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 9, September 2017

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.