IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

SPELL CHECKING AND ERROR CORRECTING SYSTEM FOR TEXT PARAGRAPHS WRITTEN IN PUNJABI LANGUAGE USING HYBRID APPROACH

Harpreet Kaur*, Navroop Kaur
*M.Tech Student, Assistant Professor, CSE Department, YCOE, Talwandi Sabo.

DOI: 10.5281/zenodo.46488

ABSTRACT

Spell checking and correcting them is very important phase of any word processing system. There are various word processing software are available for various languages like English, Hindi. We have developed a spell checker and error correcting system for Punjabi language. The proposed system use hybrid approach which s a combination of various approaches like rule based approach, dictionary lookup approach , edit distance approach and N-Gram approach. Proposed system is tested with various inputs collected from different sources and results are found very accurate than that of existing system.

KEYWORDS: Spell Checking, Punjabi Language, Edit Distance approach, Hybrid approach, Word processing.

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 is, 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 in used in various applications like machine translation, searches, information retrieval and 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 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.

LITERATURE SERVEY

Ritika Mishra[1],This paper 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.

Youssef Bassil [2],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 a further study, the proposed algorithm is to be optimized for message-passing systems so as to become more flexible and less costly to scale over distributed machines.

Neha[3], 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.

PROPOSED METHODOLOGY

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" , "N-Gram Approach , "Edit Distance approach" and use linguistic features of the Punjabi language. These approaches can be explained in brief as follows

Dictionary lookup approach

In this approach each word in the paragraph 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.

Rule based Approach

In this approach handcrafted rules are made by considering the features of the Punjabi 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.

N-Gram Approach

This works when rule based approach fails to generate the appropriate word for the incorrect words. In this approach system try to find the accurate word by considering its neighbor words by comparing with the existing paragraph stored in the system. This method also helps to identify the correct word when more than two words are generated by the rule based approach.

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 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 Bases Approach**.**
Step V Again find the word from dictionary. If word is found go to Step III, otherwise apply Edit Distance Approach.
Step VI Find the minimum distance from this Token to the word in the Dictionary.

Step VII Sort these words in ascending order of their distance.
Step VIII Check the words obtained with same distance by comparing previous and next word of the target word to obtain best possible suggestion.
Step IX If the combination available in the database then replace the top most word obtained in step VII with token otherwise go to step VII.
Step IX End.

RESULTS AND DISCUSSION

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%

Following graph is showing the comparison of existing and proposed system

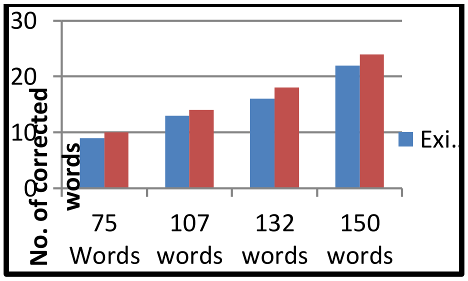

Figure 1.9

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 of existing and proposed system

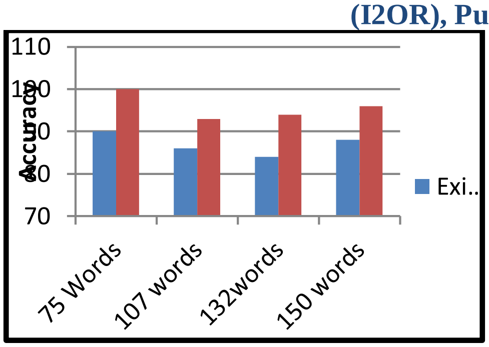

Figure 1.10

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 93% accuracy and so on.

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 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 91% according to the research work for Punjabi words. It gives nearby result up to 91% of words tested in this minor project. 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. 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. Ritika Mishra, Navjot Kaur, Design and Implementation of Online Punjabi Spell CheckeBased on Dynamic Programming, Volume 3, Issue 8, August 2013, ISSN: 2277 128X ,International Journal of Advanced Research in Computer Science and Software Engineering
2. Youssef Bassil ,Parallel Spell-Checking Algorithm Based on Yahoo! N-Grams Dataset ,International Journal of Research and Reviews in Computer Science (IJRRCS), ISSN: 2079-2557, Vol. 3, No. 1, February 201
3. 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 Engineerin
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. Gurpreet Singh Lehal, "Design and Implementation of Punjabi Spell Checker", International Journal of Systemics, Cybemetics and Infomatics, 2007.
7. 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.
8. Jesus Vilares& Manuel Vilares, "Managing Misspelled Queries in IR Application," Issue 8, October 2010.
9. Dr. R.K Sharma ,"The Bilingual Punjabi English spell checker ," Resource centre for Indain language Technology Solution ,TDIL newsletter

10. Mukand Roy ,Gaur Mohan,Karunesh K arora,"Compartive study of spell checker algorithm for building a generic spell checkers For Indian language C-DAC NODIA ,India.
11. P.Kundra and B.B Charudhari (1999),"Error pattern in Bangla text ","international Jounral of Dravidian Linguistics
12. Amit Sharma &Pulkit Jain, "Hindi Spell Checker", Indian Institute of Technology Kanpur, April 17, 2013
13. MeenuBhagat, (2007), "Spelling Error Pattern Analysis of Punjabi Typed Text", Thesis Report, Thapar University, Patiala
14. F.J. Damerau (1964), "A Technique for Error Detection and Correction of Spelling Errors", Communication ACM, pp. 171-176.