IJESRT
INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A HYBRID APPROACH FOR SPELL CHECK AND ERROR CORRECTION FOR ENGLISH AND PUNJABI TEXT PARAGRAPHS

Komaldeep Kaur[*1] & Harpreet Kaur[2]
[*1]M.Tech Student, Department of Computer Science and Engineering, Lala Lajpat Rai Institute of Engineering & Technology, Moga
[2]Assistant Professor, Department of Computer Science and Engineering, Lala Lajpat Rai Institute of Engineering & Technology, Moga

DOI: 10.5281/zenodo.1299014

ABSTRACT

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. Every word from the text is looked up in the speller lexicon. The system is made to check the spellings and to correct them using various techniques for Punjabi and English 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.

Keywords: Spell Checker, Edit distance approach, Dictionary based approach, Statistical Approach (N-Gram Approach.

I. INTRODUCTION

Spell-checking is the way toward identifying and once in a while giving recommendations to erroneously spelled words in content. Spell checking framework can be made with the mix of high quality tenets by considering syntactic elements of the dialect for which spell checking framework is to be made and a lexicon which contain the precise spellings of different words in the objective dialect. Fundamentally, the better the handmade standard and bigger the lexicon of a spell-checker is, the higher is the mistake location rate; otherwise, misspellings would pass undetected. Alas, traditional dictionaries experience from out-of-vocabulary and data sparseness problems as they do not cover large vocabulary of words necessary to cover proper names, domain-specific terms, technical lexica, special acronyms, and terminologies. As a result, spell-checkers will experience less error correction and detection rate and will fail to encounter all errors in the text. All modern mercantile spelling error detection and correction tools work on word level and use a dictionary. Each word from the data or text is searched in the speller lexicon. When a word is not in the dictionary, it is detected as an error. In order to precise the error, a spell checker locates the dictionary for words that look like the wrong word most. These words are then encouraged to the client who chooses the word that was expected. Spelling checking is utilized as a part of different applications like machine interpretation, looks, data recovery and so on. There are two primary issue identified with spell checker. These are blunder location and mistake rectification. In creating upon the kind of mistake non word blunder and genuine word blunder. There are numerous frameworks accessible for recognizing and rectifying content. Spell checker can likewise be characterized as it is a supercomputer application that examination conceivable incorrect spelling in a content by alluding to the acknowledged spellings in a database. In the database different precise expressions of the objective dialect for which the spell – checker is to be made are put away which comprises of formal people, places or things for guys, females, nations, states, waterways, mountains and so forth. The framework is made to check the spellings and to right those utilizing different methods for Punjabi-English content. In this proposed framework contribution to type of a passage is given that can incorporate off base words and the framework will create the outcome which contain the exact content in the wake of dispensing with the mistakes. 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", "Statistical Approach", "Edit Distance approach" and use linguistic features of the Punjabi-English language.
These approaches can be explained in brief as follows:

1.1 Dictionary lookup approach

In this approach every word in the passage which will be given as information is checked for the database sections. On the off chance that the filtered word is found in the database then is thought to be right word i.e. spellings of the word are right yet in the event that the word is not present in the database table then it is considered as an erroneous word. In the wake of finding the word mistaken different high quality guidelines are connected to create the right spellings of the word by considering the etymological components of the Punjabi-English dialect, if approach produce the numerous passages for the single section then by utilizing factual investigation a more suitable word id picked by the framework and is supplanted with the erroneous word to create the outcome.

1.2 Edit Distance

Edit distance is a most straightforward system in spell rectification. This most straightforward technique depends on the announcement that the individual normally commits a few errors on the off chance that one, so consequently for every lexicon word the base number of the principal altering operations (inclusion, erasures, substitutions) required to change over a word reference word into the non-word. The lower the number the higher the likelihood, that the client has made such blunders. Through the operation of including, erasing and altering, Edit-Distance changes a word into the base working recurrence of another word.

1.3 Rule based Approach

In this methodology high quality guidelines are made by considering the components of the Punjabi-English dialect. These guidelines are connected on the words in the passage which are not found in the database. By the assistance of these standards the framework endeavors to create the precise spellings of the word which is under perception.

1.4 Statistical Analysis

This works when principle based methodology neglects to create the proper word for the off base words. In this methodology framework attempt to locate the precise word by considering its neighbor words by contrasting and the current section put away in the framework. This technique additionally recognizes the right word when more than two words are created by the guideline based methodology. At the point when a present day spell checker for Punjabi - English would be utilized to spell check blunders. All advanced trade spelling blunder location and amendment instruments chip away at word level and utilize a lexicon. Every word from the information is looked in the speller vocabulary. At the point when a word is not in the lexicon, it is recognized as a blunder.

II. LITERATURE SURVEY

Bhaire et al. (2015) tested a spell Checker project with added spell checking and correction functionality to the windows based application by using autosuggestion technique. It helped the user to reduce typing work, by identifying any spelling errors and making it easier to repeat searches .The main goal of the spell checker was to provide unified treatment of various spell correction. Firstly the spell checking and correcting problem would formally describe in order to provide a better understanding of these tasks. Spell checker and corrector was either stand-alone application capable of processing a string of words or a text or as an embedded tool which was part of a larger application such as a word processor. Various search and replace algorithms were adopted to fit into the domain of spell checker. Spell checking identified the words that were valid in the language as well as misspelled words in the language. Spell checking suggested one or more alternative words as the correct spelling when a misspelled word was identifies.

Kaur and Singh (2015) analyzed that a spell checker was an application program that banners words in a report that might not be spelled effectively. A spell checker was an essential need of a word processor of any dialect. Spell checker broke down the composed content keeping in mind the end goal to recognize any incorrect spellings and gave best right recommendations for those incorrect spellings. The majority of work had been done in English and Punjabi dialect. English was the third most talked dialect on the planet. In This paper the configuration, procedures and usage of the English spell checker was proposed. Mistake discovery, Error remedy by creating recommendations and substitution were the fundamental elements of this framework. The framework distinguishes around 83.2% of the mistakes and gave 77.9% of the right recommendations for the incorrectly spelled words.

Bhirud, Bhavsar and Pawar (2017) surveyed that Natural Language processing was an interdisciplinary branch of linguistic and computer science studied under the Artificial Intelligence (AI) that gave birth to an allied area called 'Computational Linguistics' which focuses on processing of natural languages on computational devices. A natural language consisted of many sentences which are meaningful linguistic units involving one or more words linked together in accordance with a set of predefined rules called 'grammar'. Grammar checking was fundamental task in the formal world that validated sentences syntactically as well as semantically. Grammar Checker tool was a prominent tool within language engineering. Our review drew on the till date development of various Natural Language grammar checkers to look at past, present and the future in the present context. Our review covered common grammatical errors overview of grammar checking process, grammar checkers of various languages with the aim of seeking their approaches, methodologies and performance evaluation, which would greatly help for developing new tool and system as a whole. The survey concluded with the discussion of different features included in existing grammar checkers of foreign languages as well as a few Indian Languages.

Kumar, Bala and Kumar (2018) described that Spell checker was one of the most important tools for any language to be digitized. A spell checker was a software tool / plug-in that identified and corrected any spelling mistakes in a text in a particular language. Spell checkers could be combined with other research areas focusing on linguistics like Machine translation, Information retrieval, natural language processing etc. or they could club with other like software compilers, word processor software. In this paper authors had made a study on the developmental approaches as well as roles of spell checker with respect to various applications based on Indian languages.

III. PROPOSED METHODOLOGY

We will use hybrid approach to deal with execute the Spelling checking and Correcting System. This hybrid approach is a combination of "Dictionary look up approach", "Rule based approach", "Statistical Approach (N-gram Approach), "Edit Distance approach" and use linguistic features of the Punjabi-English language. Dictionary Look up Approach and Edit Distance Approach is used in the research which is already implemented. The framework which is to created will utilize a half breed way to deal with check and to remedy the wrong spelled words. Presently in this proposal research I will utilize the Rule Based Approach, Edit distance Approach, Dictionary look up Approach and Statistical Approach with more accuracy.

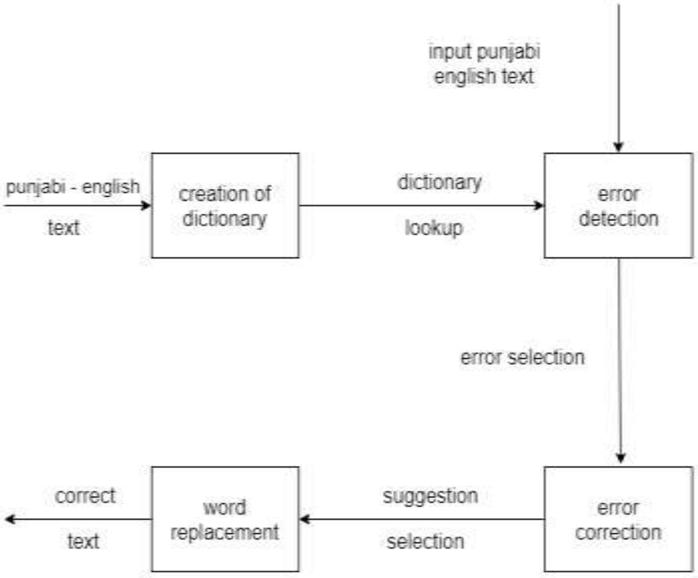

Figure: Architecture of Proposed system

3.1 Proposed Algorithm

Following are the steps of proposed algorithm:

Step I: Input the text paragraph containing the miss-spelled words.

Step II: divide the input collected from step 1 into words (Tokens).

Step III: Detect the language of the token and then compare every token with the corpus for the corresponding language.

Step IV: With the reference to the corpus check whether the particular token is correct or not. If it is correct, then go to Step III, otherwise apply Rule Bases Approach on that token**.**

Step V: After applying the rule based approach, search for the word in the corpus again. If word is found go to Step III, otherwise apply Edit Distance Approach.

Step VI: Now calculate the least distance from this particular Token to the closed related words in the Dictionary.

Step VII: Sort these generated words obtained in step VI in ascending order of their distance.

Step VIII: If the generated words have the same distance, then check that word by comparing previous and next word of the target word to obtain best possible suggestion among all generated suggestions.

Step IX: If after comparison the group of words the targeted words are found, then replace the top most word obtained in step VII with token otherwise go to step VII.

Step X: End.

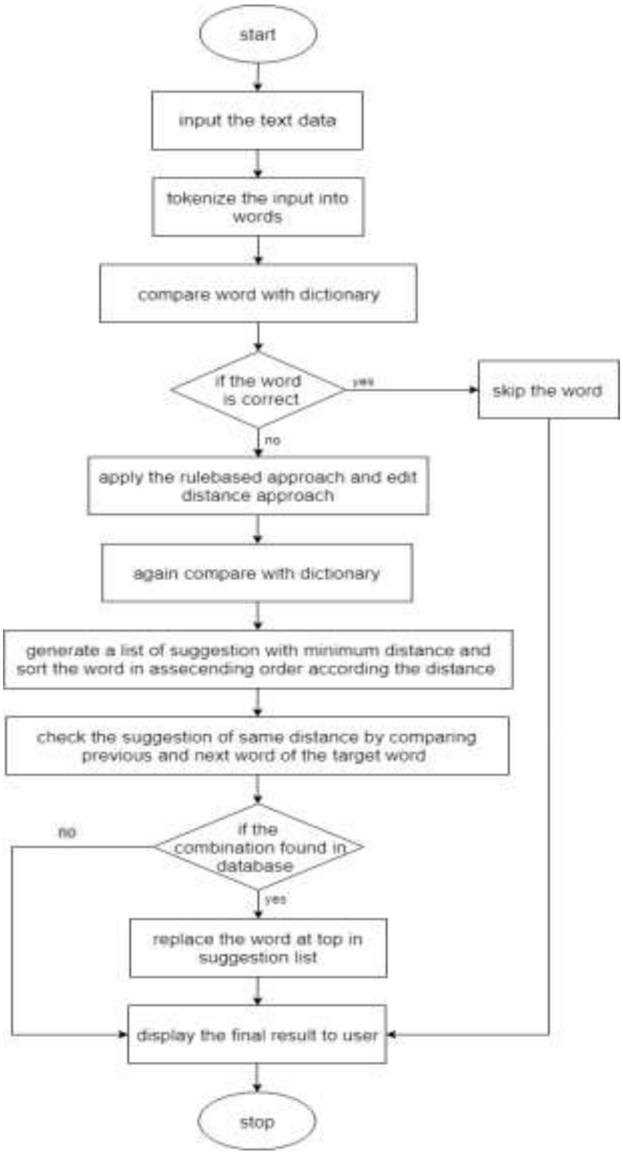
start
input the text data
tokenize the input into words
compare word with dictionary
if the word is correct
yes
skip the word
no
apply the rulebased approach and edit distance approach
again compare with dictionary
generate a list of suggestion with minimum distance and sort the word in assecending order according the distance
check the suggestion of same distance by comparing previous and next word of the target word
if the combination found in database
no
yes
replace the word at top in suggestion list
display the final result to user
stop

Figure: Flow chart of proposed system

3.2 Dictionary creation

Dictionary creation is a tool used in spell checker application to create the dictionary. This dictionary will be used as a database for the spell checker. Microsoft access 2007 is used to create a database for Punjabi-English spell checker. As Shown in figure 3.3 & 3.4 by clicking on insert data button, words will be added into database of the spell checker.

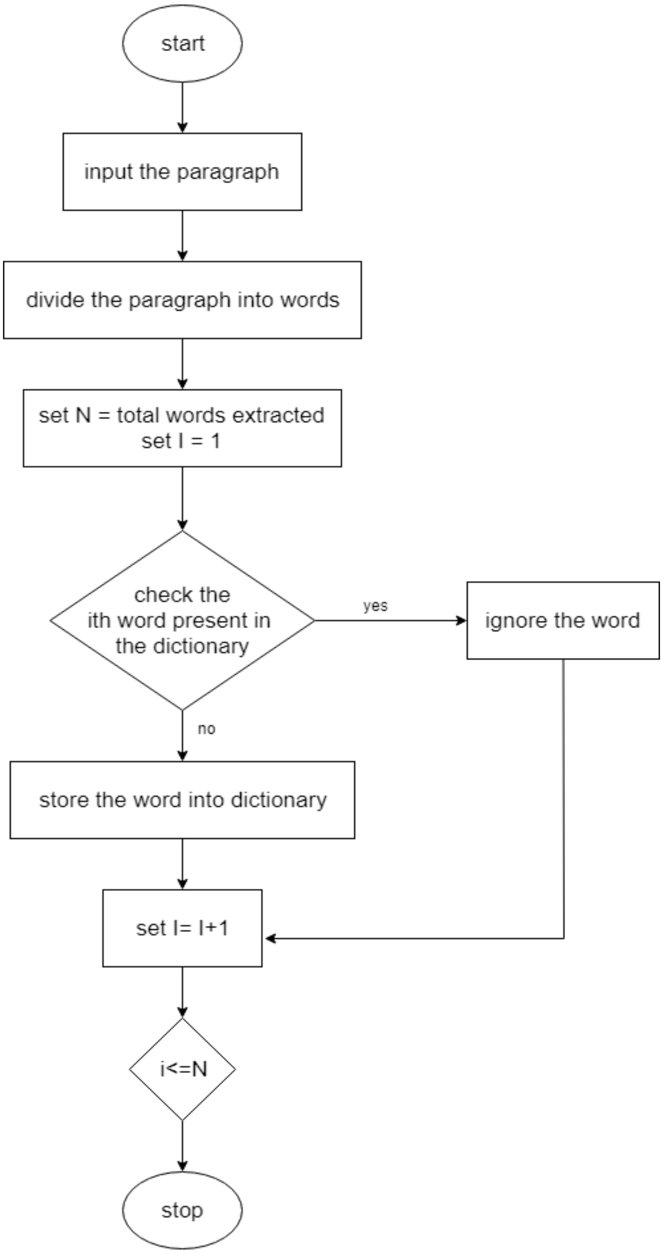

Figure: Flowchart of Dictionary Creation

3.3 Error detection

Error detection is the process of finding wrong spelled word in the text paragraphs. Dictionary lookup technique is the best method to find the errors in text paragraph.

3.3.1 Language Detection

In this proposed system, language detection for independent words is also been implemented according to which system check and correct the errors. The following is the flow chart of language detection system.

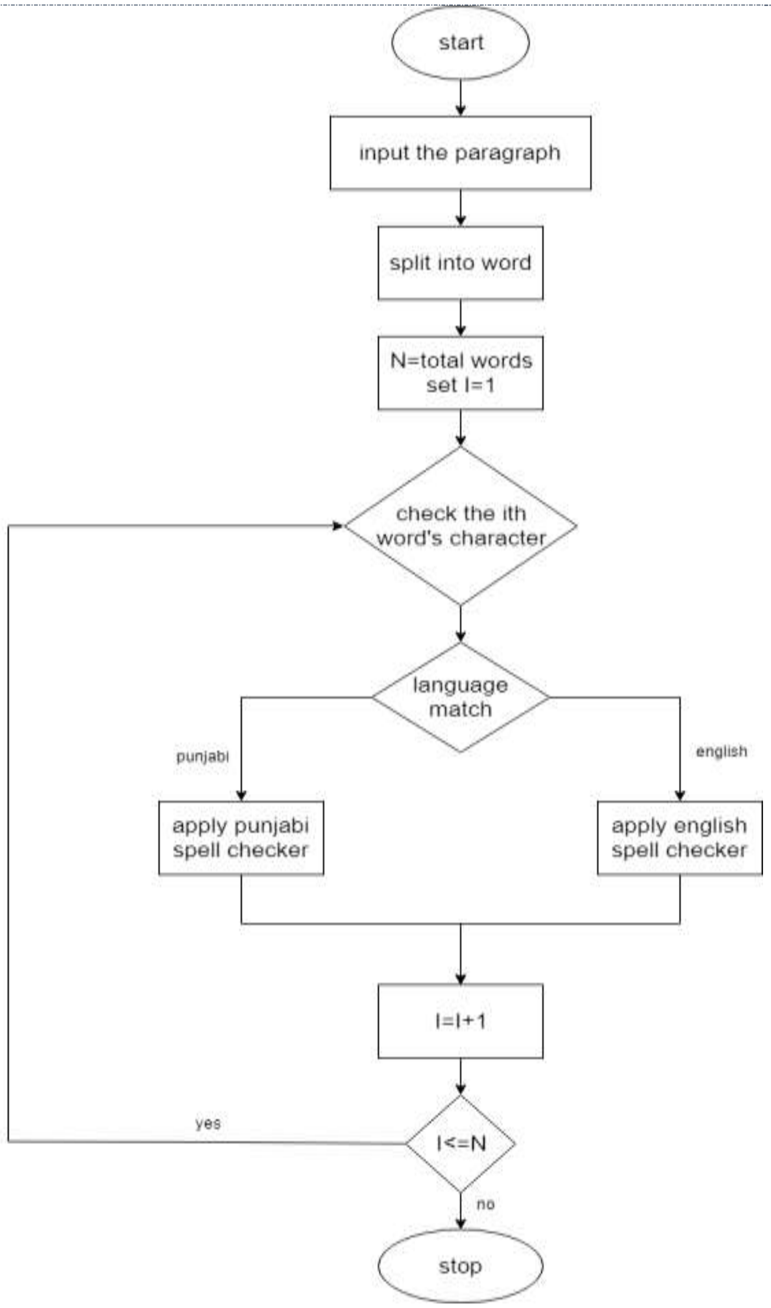

Figure: Flowchart of language detection

3.3.2 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 Punjabi-English words, various resources like Punjabi-English text books, online Punjabi-English 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: Enter the text data that contain the error words.
Step II: Divide this text data into tokens (words)
Step III: Compare it with the dictionary to check whether it is correct or not.
Step IV: End

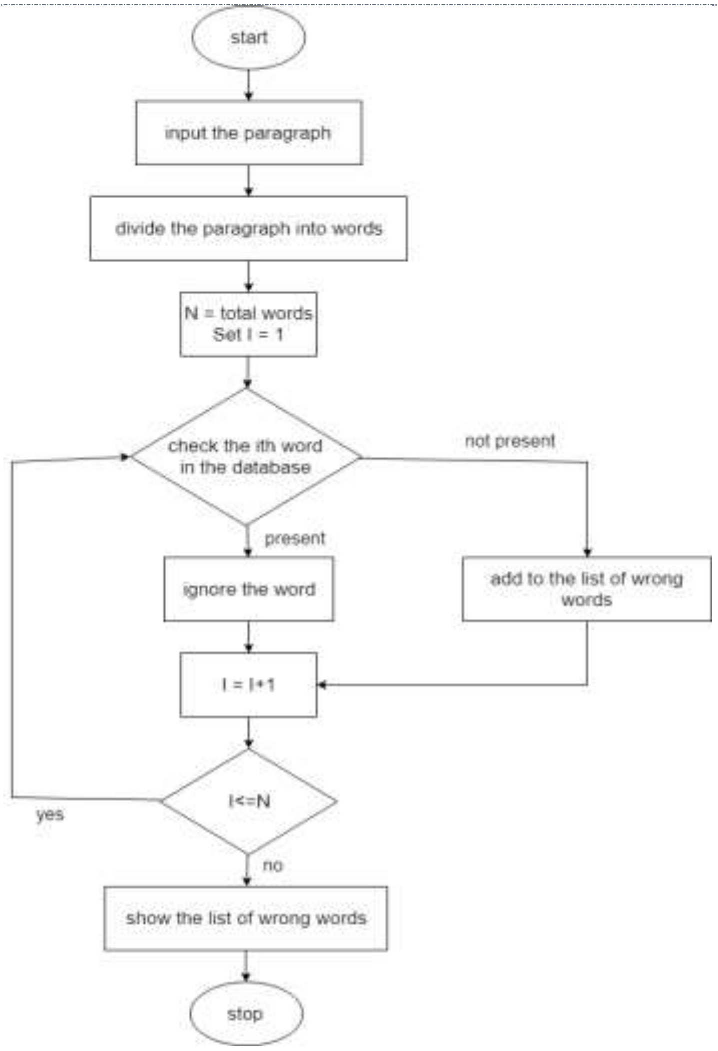

Figure: Flowchart of Dictionary Lookup approach

3.4 Error correction

It is the process of correcting the wrong spelled word in the text paragraphs. Rule based approach, edit distance approach and N-gram approach is the best method of error correction.

3.4.1 Rule Based Approach

Rule based approach is a set of language dependent rules with which the input tokens are compared and applied on the tokens. Rule based approach in the proposed system is used to correct the entries with the help of grammatical features of the particular language. The rule-based approach has successfully been used in developing many natural language processing systems. Systems that use rule-based transformations are based on a core of solid linguistic knowledge. The linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system. The advantage of the rule-based approach over the corpus-based approach is clear for: 1) less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable, and 2) for morphologically rich languages, which even with the availability of corpora suffer from data sparseness. These have motivated many researchers to fully or partially follow the rule-based approach in developing their Arabic natural processing tools and systems. The following are the steps of rule based approach:

Step I: Give input the particular input to correct the error.
Step II: Divide the input into number of independent words.
Step III: Compare every token generated in the step II with the corpus stored in database..

Step IV: If the corresponding token is found correct then start search with the next token otherwise apply rule based approach to correct that token.
Step V: Use this corrected output as an input to the next phase.
Step VI: end

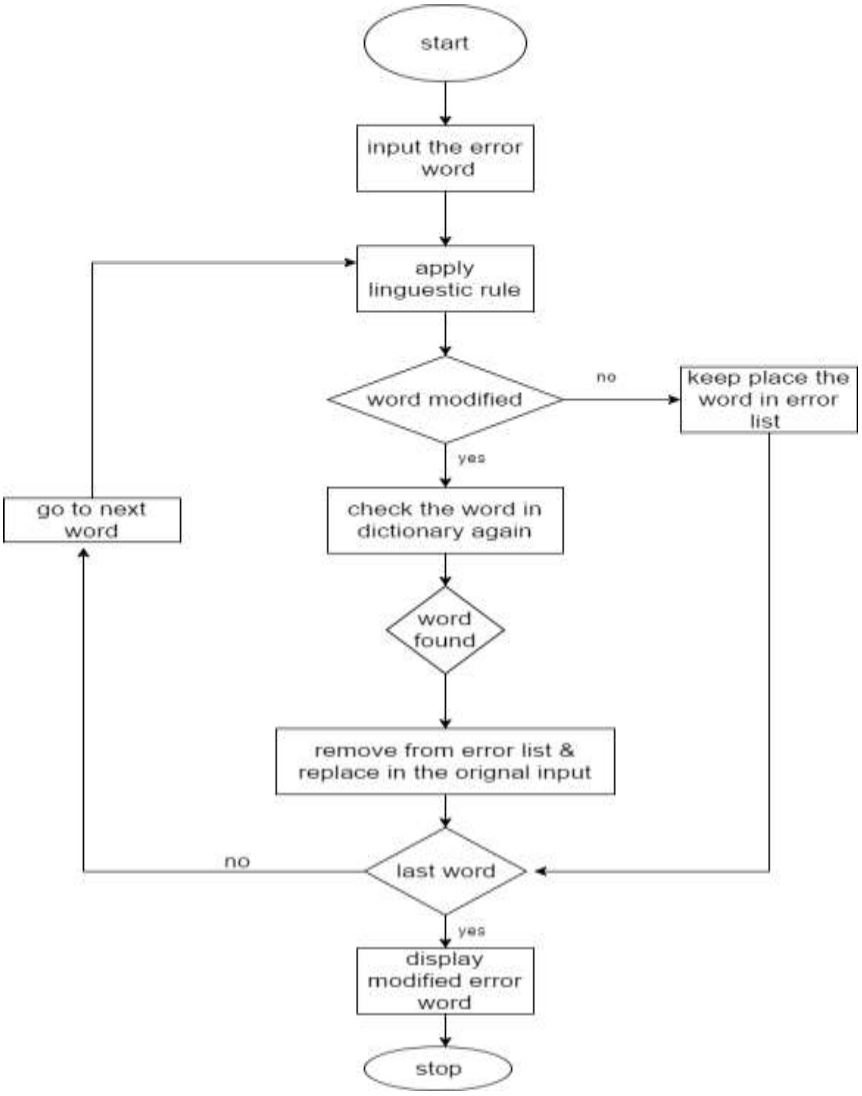
start
input the error word
apply linguestic rule
word modified
no
keep place the word in error list
yes
go to next word
check the word in dictionary again
word found
remove from error list & replace in the orignal input
no
last word
yes
display modified error word
stop

Figure : Flowchart of Rule Based Approach

3.4.2 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

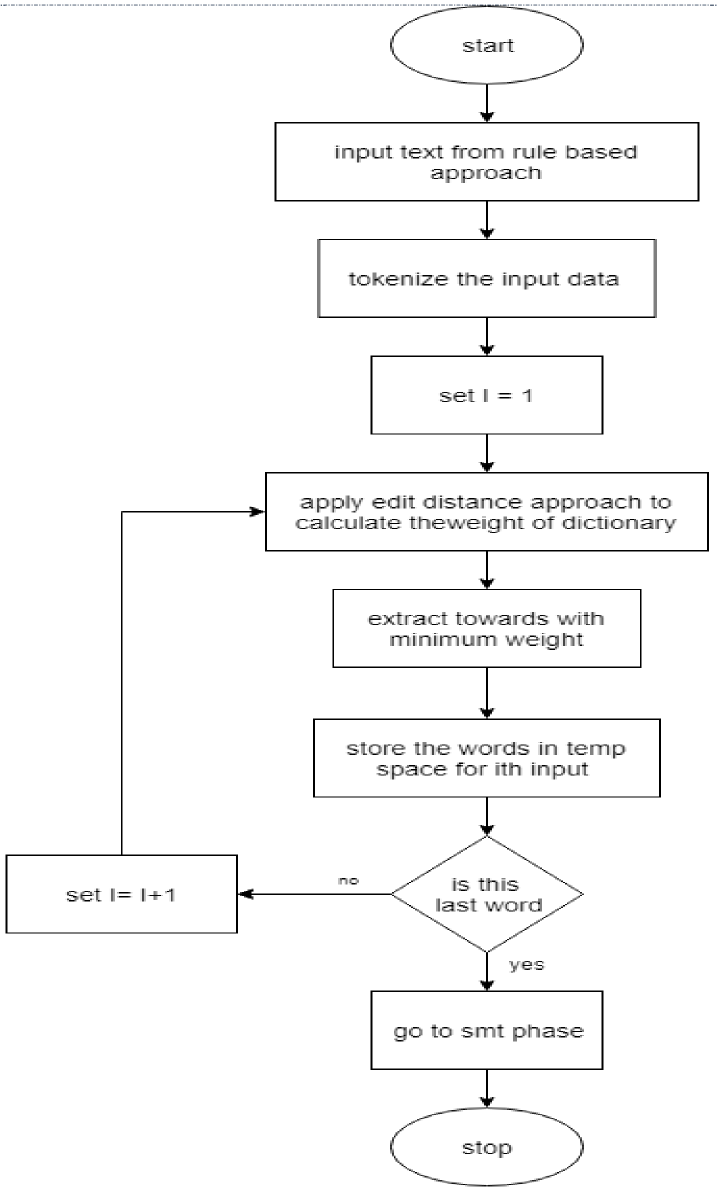
start
input text from rule based approach
tokenize the input data
set I = 1
apply edit distance approach to calculate theweight of dictionary
extract towards with minimum weight
store the words in temp space for ith input
is this last word
no
set I= I+1
yes
go to smt phase
stop

Figure: Flowchart of Edit Distance Technique

3.4.3 Statistical Approach (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. SMT 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.

Following are the steps of statistical 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

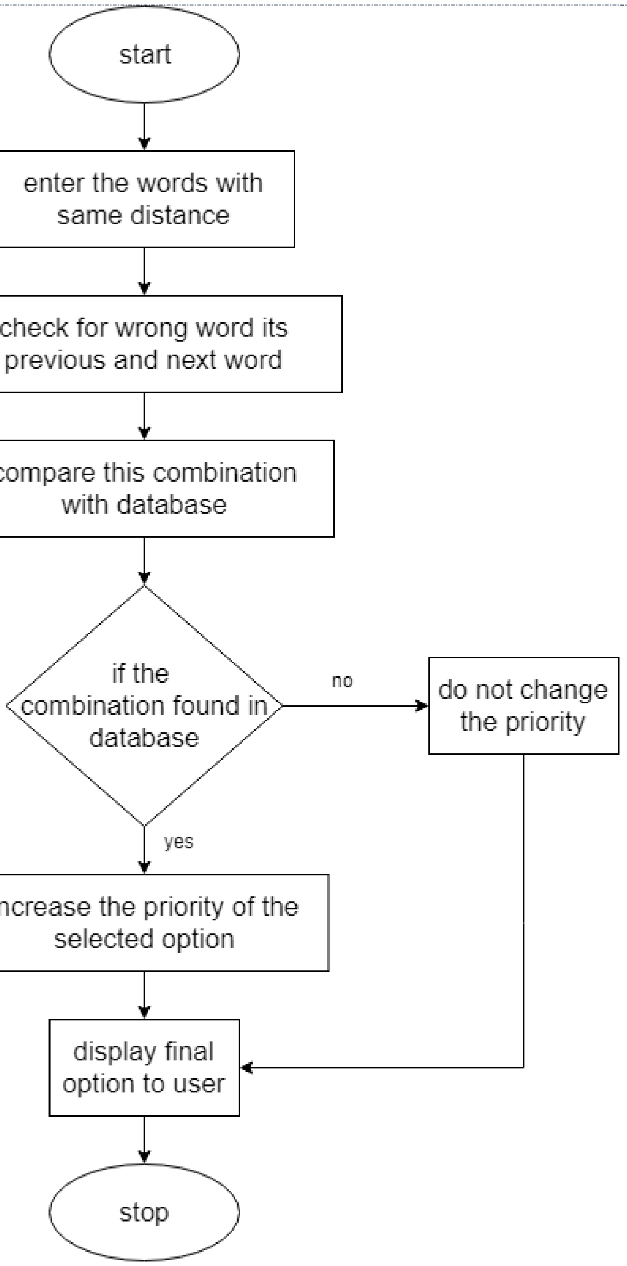
start
enter the words with same distance
check for wrong word its previous and next word
compare this combination with database
if the combination found in database
no
do not change the priority
yes
increase the priority of the selected option
display final option to user
stop

Figure: Flowchart of Statistical Approach (N-Gram Approach)

3.5 Word Replacement

Word replacement is the process of replacing the wrong word with the correct word. Spell checking system also generate the suggestion for wrong word. 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. At the last user replaced the wrong word with the correct word and the system will generate the result which contain the accurate text after eliminating the errors.

IV. EXPERIMENTAL RESULTS

To evaluate the performance of the system various inputs from various text books and online websites with some errors in their words are input to the system and results are analyzed. It is calculated that the overall accuracy of the system is 91%.

Table: Experimental Results

Sr. No.	Total No. of words in paragraph	Errors in Paragraphs	Correction by Existing System by Edit Distance Technique	Accuracy of Existing system by 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	250	40	36	90%	38	95%

This above table shows the values of total no. of words in paragraphs, errors in paragraphs, correction and accuracy of existing system and proposed system. The table clearly shows that proposed system corrected more errors as compared to existing system and has more accuracy than existing system.

4.1 Comparison based on correctness

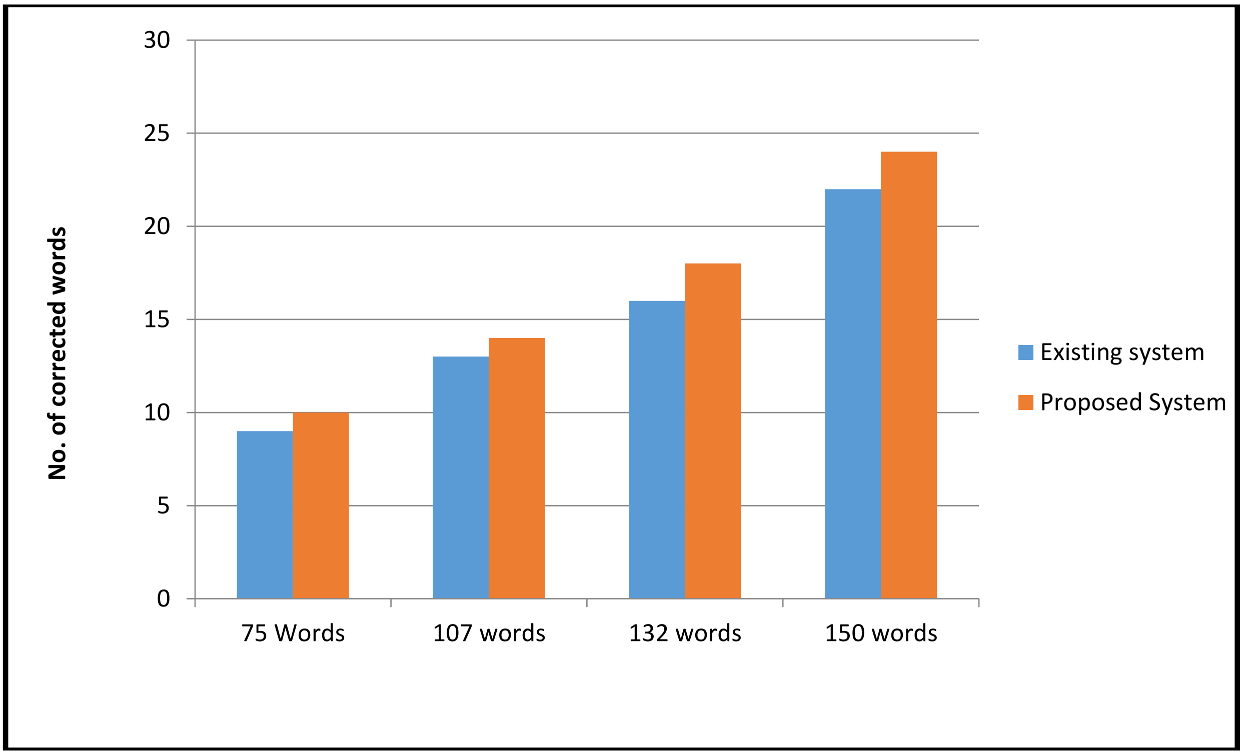

Figure: Correctness comparison of existing and proposed system

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.

ISSN: 2277-9655
Impact Factor: 5.164
CODEN: IJESS7

4.2 Comparison based on accuracy

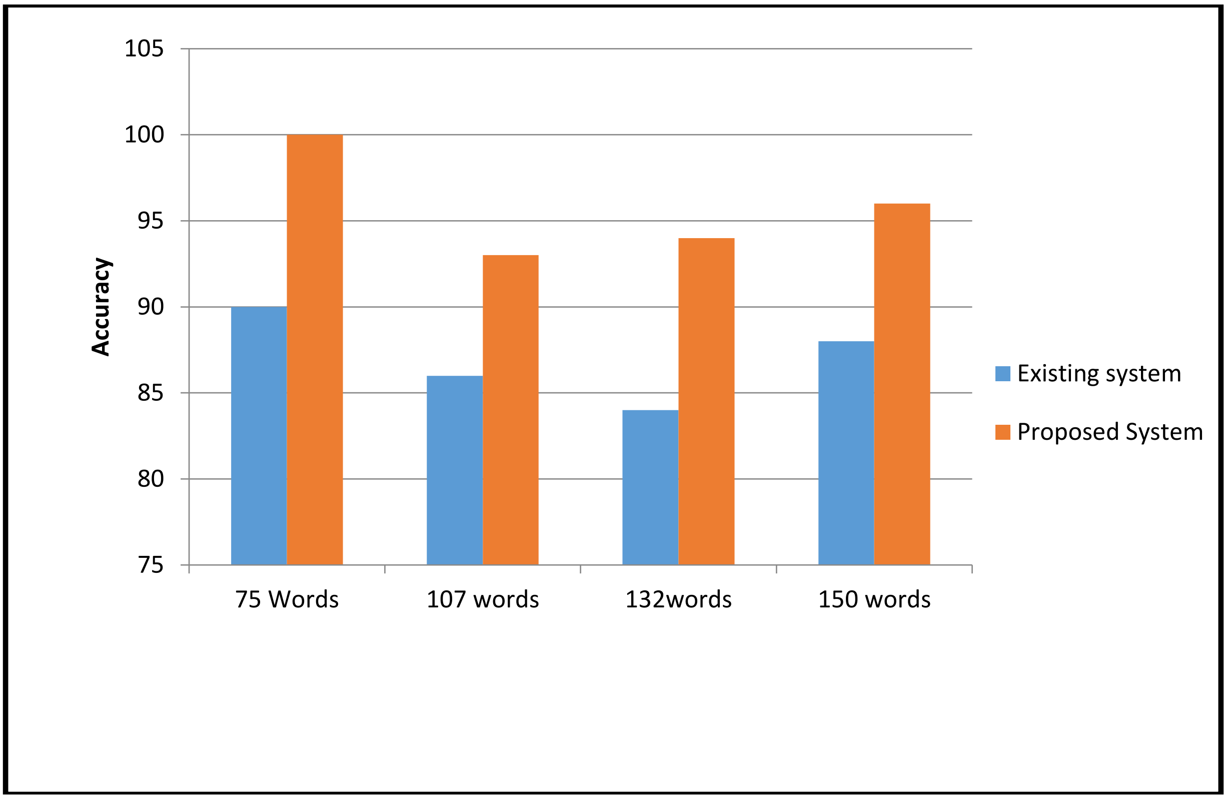

Figure: Accuracy Comparison of existing and proposed system

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 93% accuracy and so on.

V. CONCLUSION

In this Research work, a new online Punjabi-English spell checker is developed along with a noval hybrid based approach for the correction of wrongly spelled words according to the corpus stored in the database. Proposed system is based on hybrid approach in which these approaches which are dictionary look up approach, edit distance approach, rule based approach and Statistical approach (N-gram approach or SMT) are used into one. The main features of Punjabi-English spell checker are large database, online application, easy to operate and automatic language detection system. This System gives the result accuracy as 91% according to the research work for Punjabi-English words. 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.

VI. FUTURE SCOPE

In this Research work, instead of giving the highlighter for wrong spelled words, a list of wrong words has been provided. This limitation of the proposed work can be overcome in the future by providing the wrongly spelled words with red highlighter which are not stored the corpus for the particular language. In future to extend the proposed work, some grammatical rules for both of the languages may also be added. In future, more words related to the targeted languages can be added to the corpus of the system to improve overall accuracy.

REFERENCES

[1] Harpreet Kaur,Gurpreet Kaur, Manpreet Kaur,"Punjabi Spell Checker Using Dictionary Clustering", International Journal of Science, Engineering and Technology Research (IJSETR), Volume 4, Issue 7, July 2015

[2] Jaspreet Kaur,Kamaldeep Garg,"Hybrid Approach for Spell Checker and Grammar Checker for Punjabi", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 6, June 2014

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

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

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

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

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

[8] Neha Gupta and Pratistha Mathur,"Spell Checking Techniques in NLP: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, No. 12, December 2012.

[9] Neha Gupta and Pratistha Mathur,"Spell Checking Techniques in NLP: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 12, Dec. 2012.

[10] Nivedita S. Bhirud, R.P. Bhavsar, B.V. Pawar,"GRAMMAR CHECKERS FOR NATURAL LANGUAGES: A REVIEW", International Journal on Natural Language Computing (IJNLC) Vol. 6, No.4, August 2017

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

[12] Rakesh Kumar, Minu Bala, Kumar Sourabh,"A study of spell checking techniques for Indian Languages ",JK Research Journal in Mathematics and Computer Sciences,Vol. (1) No. (1) March 2018.