ISSN No. 0976-5697

Volume 7, No. 6(Special Issue), November 2016

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

A Combined Spell Checking and Error Correcting System for Punjabi -Hindi Language using Hybrid Approach

Amandeep Singh
M-Tech student, SSCET,Badhani,Pathankot,India
er.abbrar@gmail.com

Abstract: Spellchecker is software that analyzes possible misspellings in the text. It is the process of detecting and sometimes providing some suggestions for incorrectly spelled words in a text. If dictionary of spell checker is larger than higher is the error detection and error correction rate. Though considerable work has been done in English language but not much work has been done in regional languages of India including Punjabi and Hindi. Punjabi is the official language of Punjab state in India Punjabi is world's 12th most widely spoken language In Punjabi language, there is very small amount of work is completed in this region.. Hindi is the official language of 11 vowels and 33 consonants. Hindi is also the third most spoken language in the world. A few work is done in Punjabi-Hindi spell detection and correction field and it is not easy task to identify errors in Punjabi-Hindi text. The spell checker systems are online available but as not stand-alone applications. The only available spell checker for Punjabi is "Akhar" "Raftaar" and "Sudhaar". "Akhar" is paid software that is not available free for its use to everybody and "Sudhaar" spell checker is a desktop application Some paid Hindi spell checker software's are also online available. "Hinspell" & "Hinkhoj" are available spell checkers for Hindi language but a lot of improvement is needed. NLP (Natural language processing) is a field of computer science concerned with interaction between computer and human language. We have developed a combined spell checker and error correcting system for both Punjabi and Hindi Language. We used hybrid approach to implement the Spelling checking and Correcting System. 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, Error Detection, Error Correction, Punjabi, Hindi, Dictionary-lookup, Hybrid approach.

1. INTRODUCTION

A spell checker is a technique which identifies the incorrect or misspelled words and replaces them with the best possible combination of correct words. For find incorrect word firstly system checks the word in the dictionary. If the word is finding in the database then it assume to be correct word and if it is not present in the database then system assume this word incorrect and perform the required process to generate best possible combination of correct word. The ways in which the words can be meaningfully combined is defined by the language's syntax and grammar. The actual meaning of words and combinations of words is defined by the language's semantics. Hindi is the official language of India which consist 11 vowels and 33 consonants. Hindi is also the third most spoken language in the world .Spell checking is the process of detecting and providing correct suggestions for misspelled words in a written text. Spell correction is a one of the main functions of word processors, search engines, text editors, and optical character recognition. Error detection, suggestion generator, error correction are three main steps in a spell checker. Error Correction is a major issue in the language processing field. Much research has been done in this area over the years. Before studying about error detection and correction, it's very important to know how spelling errors occurs. 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-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.

Types of errors

1. Insertion error(IE): when at least one extra character is inserted in the desired word. Ifjo ^ Ijo gfotko^ gfotdko, ;kok ^ ;ko, आदमी ^ अदमी, ਸਤਾ - ਸੱਤਾ
2. Deletion errer(DE): when at least one character is deleted in the desired word.
   gfotko ^ gfoto, rb ^ rZb, अपमान ^ आपमान
3. Substitution error(SE): when at least one character is substituted by the other character.
   gfotko ^ gfotBo, भील ^ भलु, ਸ਼ੀਸ਼ਾ - ਸ਼ੀਸਾ

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

4. Transposition error(TE): when two adjacent characters are transposed. gfotko ^ gfotok ,oks ^ osk eow ^ ewo अपमान ^ अपनkम
5. Run-on error(ROE): when there is space missing between two or more valid words. gfotko tkd ^ gfotkotkd, पंजाब के ^ पंजाबके, मेरा घर ^ मेराघर, ਆਦਰ ਮਾਣ - ਆਦਰਮਾਣ
6. Split word error(SWE): This is Opposite of Run-on error when some extra space is inserted between parts of a word. The error can be removed by removing the extra space. gfotko ^ gf otko, महाराजा ^ महा राजा, dhtko ^ dh tko, ਦੁਖੀ - ਦੁ ਖੀ

2. LITERATURE SURVEY

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
In this paper we have surveyed the area of spell correction and error detection techniques. Existing work related with spell checkers in Punjabi-Hindi 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

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
This paper describes the development and working of online Raftaar Punjabi-Hindi 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-Hindi 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.

3. PROPOSED WORK

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-Hindi language. The system which is to developed will use a hybrid approach to check and to correct the wrong spelled words. Now in this project research I will use the Rule Based Approach and Statistical Approach with more accuracy.

Architecture of Proposed system

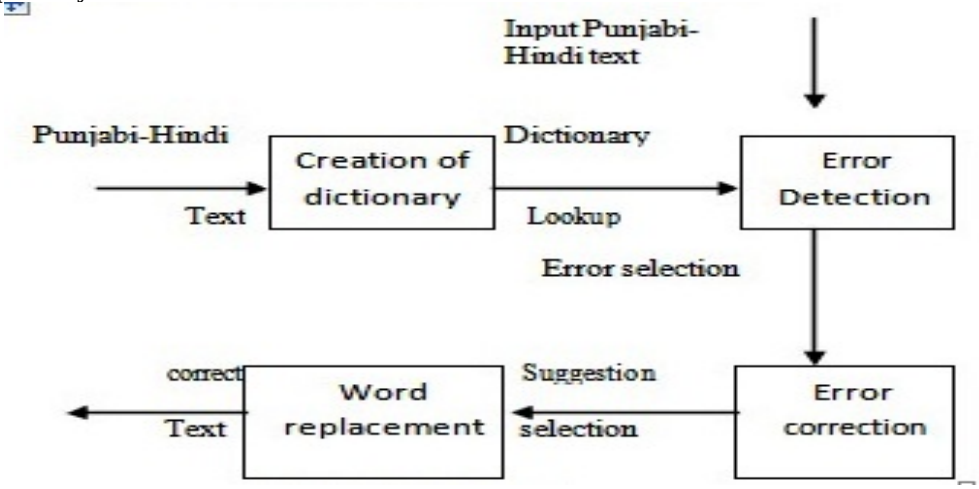

Fig. 1.1(*Error Detection)*

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

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.

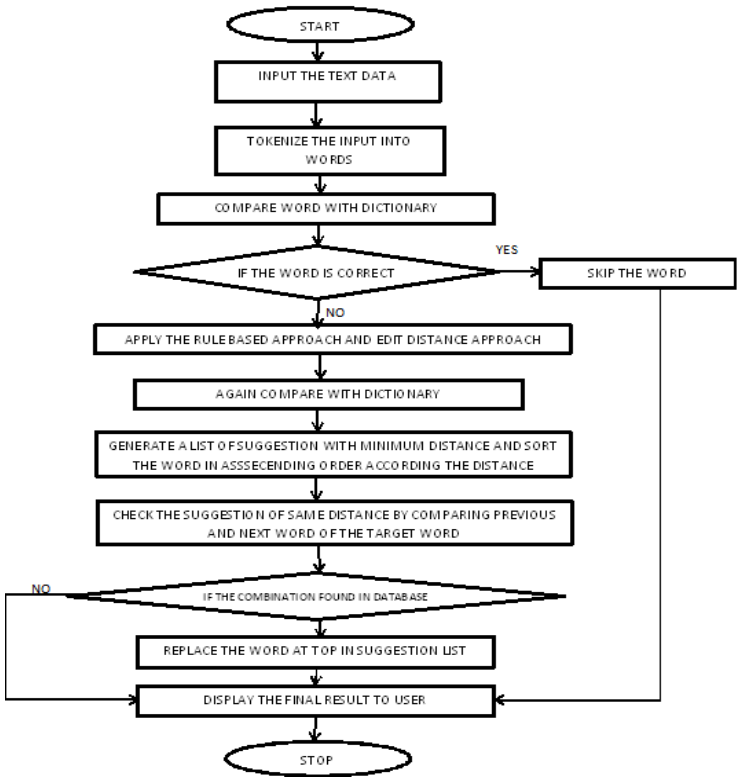

Fig. 1.2(Flow chart of hybrid approach)

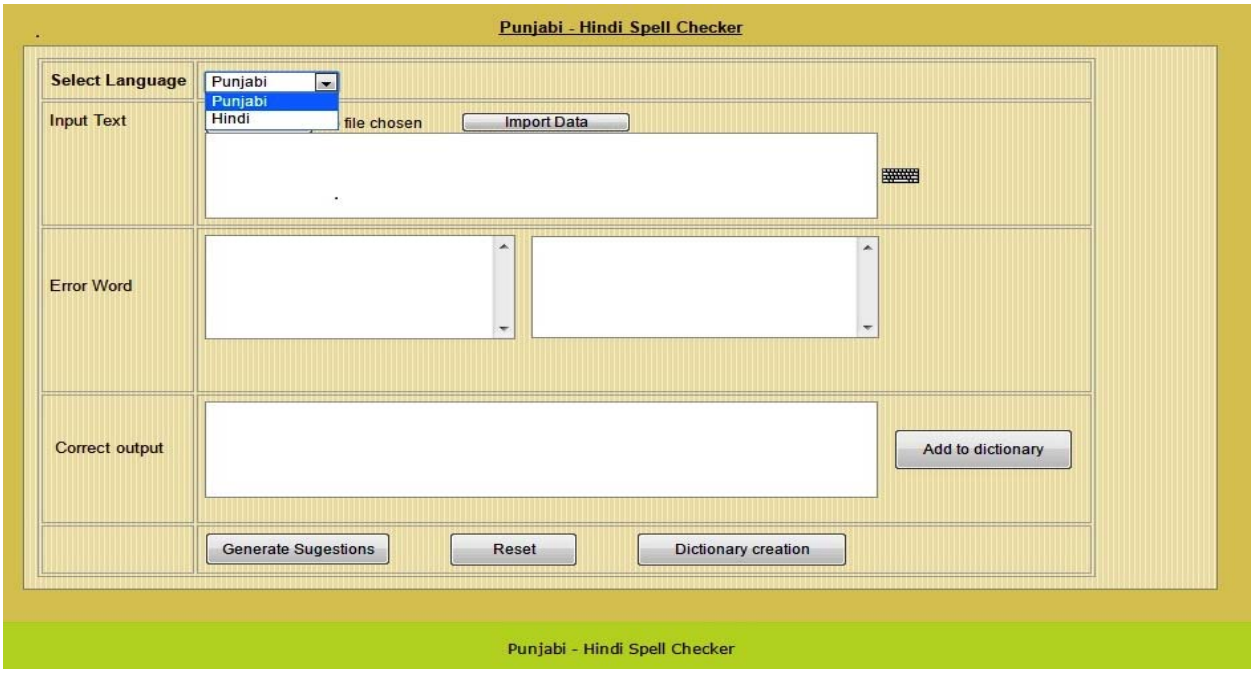

Fig:-1.3 (User Interface of Punjabi-Hindi Spell Checker)

3.1 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-Hindi spell checker. As

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Shown in figure 1.4 & 1.5 by clicking on insert data button, words will be added into database of the spell checker.

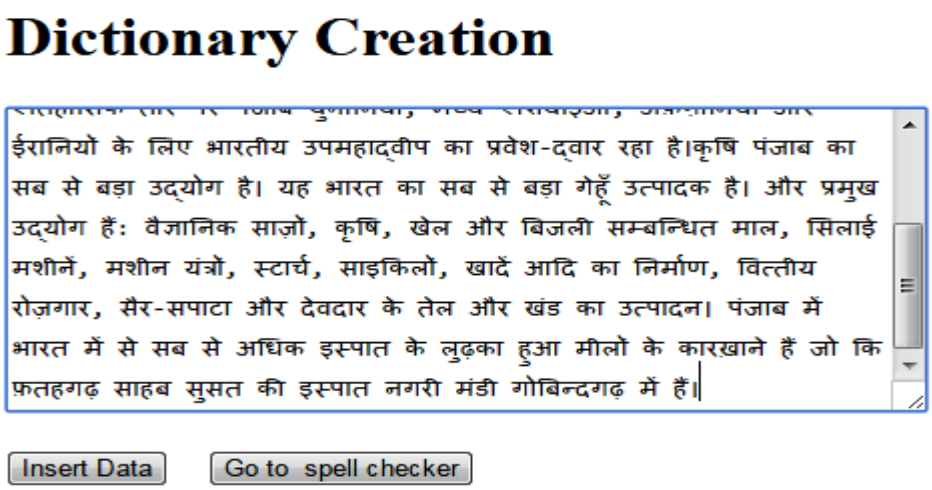

Fig:-1.4 (Dictionary creation tool For Hindi)

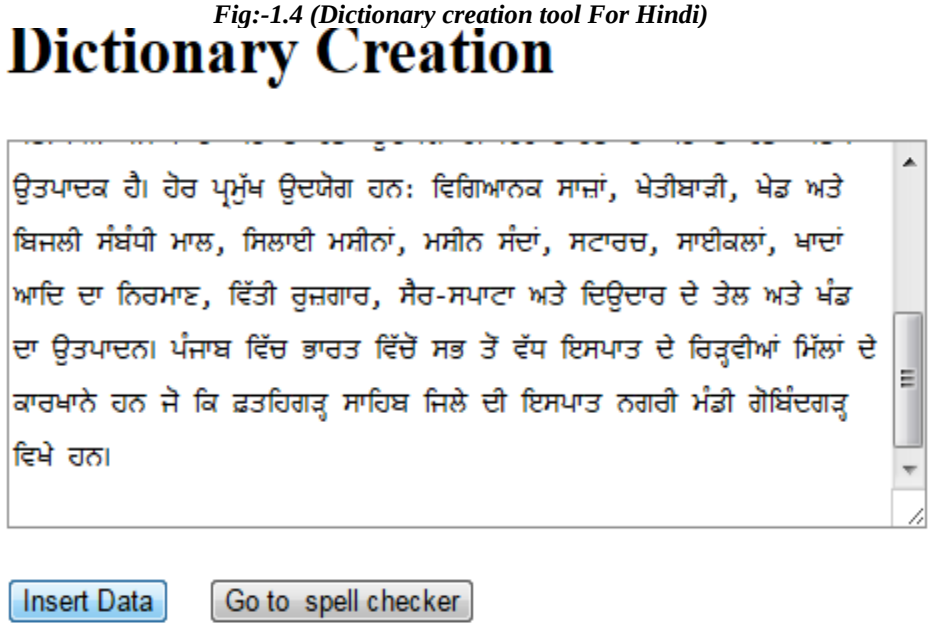

Fig:-1.5(Dictionary creation tool for Punjabi)

4. RESULTS

Input 1

ਭਾਰਤ ਹਿੰਦੀ ਪ੍ਰਾਚੀਨ ਜੰਬੂ ਦੀਪ ਆਧੁਨਿਕ ਦੱਖਣੀ ਏਸ਼ੀਆ ਵਿੱਚ ਸਥਿੱਤ ਭਾਰਤੀ **ਉਮਹਾਂਦੀਪ** ਦਾ ਸਭ ਤੋਂ ਵੱਡਾ ਦੇਸ਼ ਹੈ। ਭਾਰਤ ਭੂਗੋਲਕ ਨਜ਼ਰ ਵਲੋਂ **ਸਸਾਰ** ਵਿੱਚ ਸੱਤਵਾਂ ਸਭ ਤੋਂ ਵੱਡਾ ਅਤੇ ਪੱਖੋਂ ਦੂਜਾ ਸਭ ਤੋਂ ਵੱਡਾ ਦੇਸ਼ ਹੈ। ਭਾਰਤ ਦੇ ਪੱਛਮ ਵਿੱਚ ਪਾਕਿਸਤਾਨ ਉੱਤਰ ਪੂਰਬ ਵਿੱਚ ਚੀਨ, **ਨੇਪਲ** ਅਤੇ ਭੂਟਾਨ ਅਤੇ ਪੂਰਬ ਵਿੱਚ ਬੰਗਲਾਦੇਸ਼ ਤੇ ਮਿਆਂਮਾਰ ਦੇਸ਼ ਸਥਿਤ ਹਨ। ਇਸਦੇ ਉੱਤਰ ਵਿੱਚ ਹਿਮਾਲਾ ਪਹਾੜ ਹਨ ਅਤੇ ਦੱਖਣ ਵਿੱਚ ਹਿੰਦ ਮਹਾਂਸਾਗਰ ਹੈ। ਭਾਰਤ ਵਿੱਚ ਕਈ ਵੱਡੀਆਂ ਨਦੀਆਂ ਹਨ। ਗੰਗਾ ਨਦੀ ਭਾਰਤੀ ਸੰਸਕ੍ਰਿਤੀ ਵਿੱਚ ਬਹੁਤ ਪਵਿੱਤਰ ਮੰਨੀ ਜਾਂਦੀ ਹੈ। ਭਾਰਤ ਸੰਸਾਰ ਦਾ ਸਭ ਤੋਂ ਵੱਡਾ ਲੋਕਤੰਤਰ ਹੈ। ਇੱਥੇ ਤੋਂ ਜ਼ਿਆਦਾ ਭਾਸ਼ਾਵਾਂ ਬੋਲੀਆਂ ਜਾਂਦੀਆਂ ਹਨ। ਸੰਸਾਰ ਦੇ ਚਾਰ ਧਰਮ ਹਿੰਦੂ, ਬੁੱਧ, ਜੈਨ ਅਤੇ ਸਿੱਖ ਦਾ ਜਨਮ ਅਤੇ ਵਿਕਾਸ ਭਾਰਤ ਵਿੱਚ ਹੀ ਹੋਇਆ। ਭਾਰਤ ਭੂਗੋਲਕ ਖੇਤਰਫਲ ਦੇ ਅਧਾਰ ਤੇ ਸੰਸਾਰ ਦਾ ਸੱਤਵਾਂ ਸਭ ਤੋਂ ਵੱਡਾ ਰਾਸ਼ਟਰ ਹੈ

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Output 1

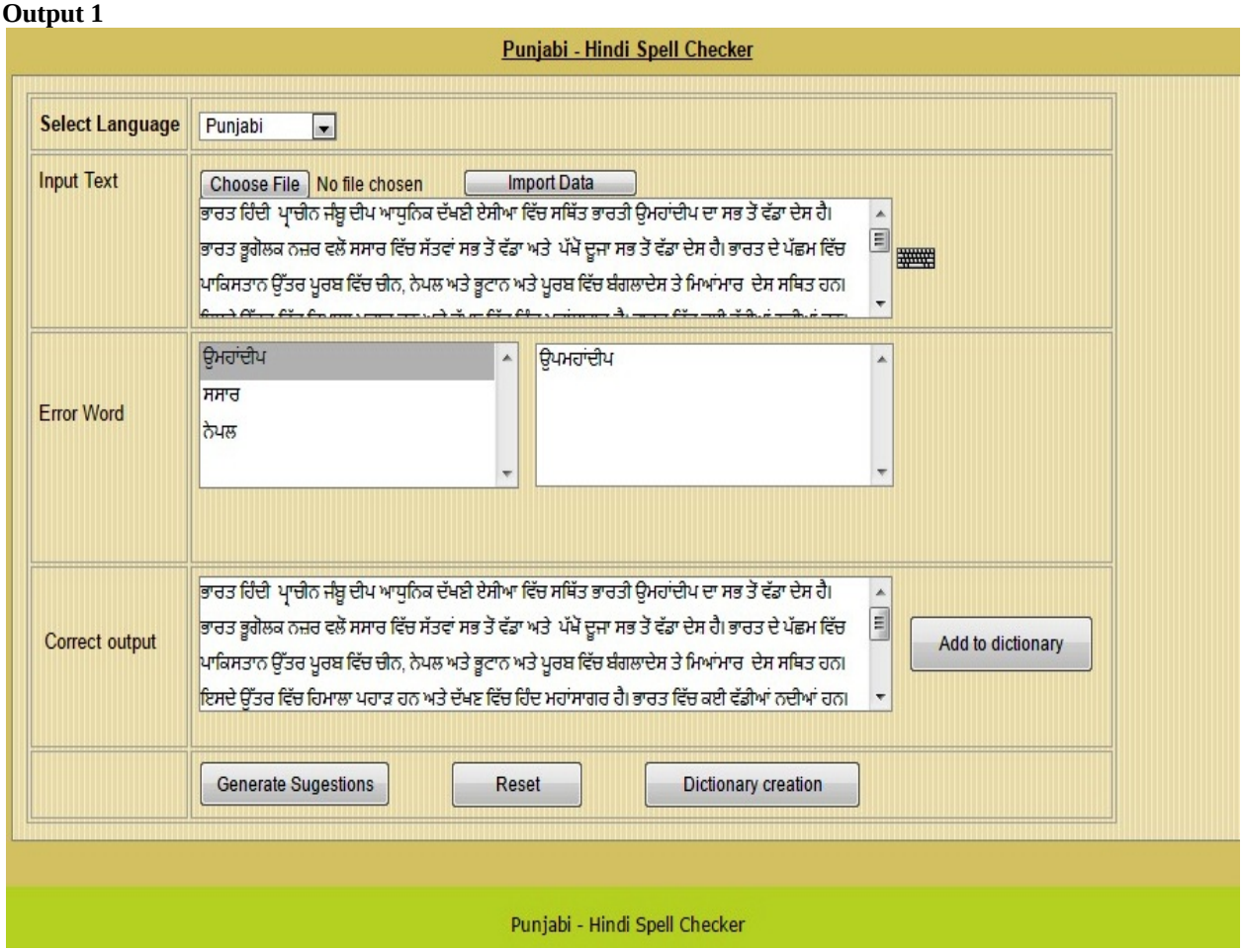

Fig:-1.6

Input 2

ਪੂਰਬ ਵਿੱਚ ਬੰਗਾਲ ਦੀ ਖਾੜੀ ਹੈ ਅਤੇ ਪੱਛਮ ਵਿੱਚ ਅਰਬ **ਸਗਰ** ਹੈ। ਭਾਰਤ ਵਿੱਚ ਕਈ ਵੱਡੀਆਂ ਨਦੀਆਂ ਹਨ। **ਭਰਤ** ਦੇ ਦੋ ਅਧਿਕਾਰਕ ਨਾਮ ਹਨ ਹਿੰਦੀ ਵਿੱਚ ਭਾਰਤ ਅਤੇ **ਅੰਗਰਜ਼ੀ** ਵਿੱਚ **ਇਡੀਆ** ,ਇਸਨੂੰ ਹਿੰਦੁਸਤਾਨ ਵੀ ਆਖਦੇ ਹਨ।ਭਾਰਤ ਵਿੱਚ ਕਈ ਵੱਡੀਆਂ ਨਦੀਆਂ ਹਨ। ਗੰਗਾ ਨਦੀ ਭਾਰਤੀ ਸੰਸਕ੍ਰਿਤੀ ਵਿੱਚ ਬਹੁਤ ਪਵਿੱਤਰ ਮੰਨੀ ਜਾਂਦੀ ਹੈ। ਭਾਰਤ ਸੰਸਾਰ ਦਾ ਸਭ ਤੋਂ ਵੱਡਾ ਲੋਕਤੰਤਰ ਹੈ। ਇੱਥੇ ਤੋਂ ਜ਼ਿਆਦਾ ਭਾਸ਼ਾਵਾਂ ਬੋਲੀਆਂ ਜਾਂਦੀਆਂ ਹਨ। ਸੰਸਾਰ ਦੇ ਚਾਰ ਧਰਮ ਹਿੰਦੂ, ਬੁੱਧ, ਜੈਨ ਅਤੇ ਸਿੱਖ ਦਾ ਜਨਮ ਅਤੇ ਵਿਕਾਸ ਭਾਰਤ ਵਿੱਚ ਹੀ ਹੋਇਆ। ਭਾਰਤ ਭੂਗੋਲਕ ਖੇਤਰਫਲ ਦੇ ਅਧਾਰ ਤੇ ਸੰਸਾਰ ਦਾ ਸੱਤਵਾਂ ਸਭ ਤੋਂ ਵੱਡਾ ਰਾਸ਼ਟਰ ਹੈ

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Output 2

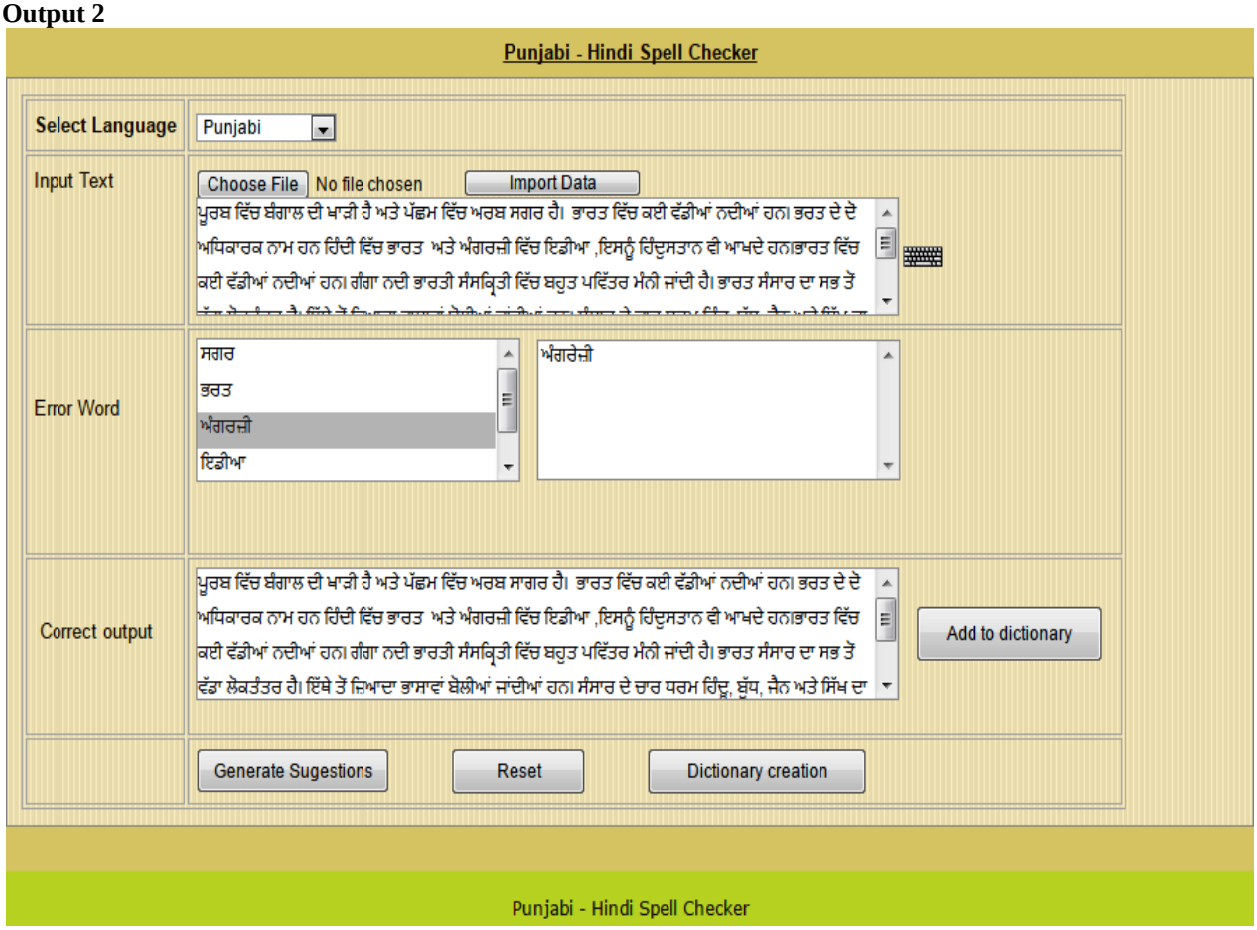

Fig:-1.7

Input 3

प्राचीन ओलंपिक्स से वर्तमान सदी तक खेल आयोजित किये जाते रहे हैं और उनका विनियमन भी होता रहा है। औद्योगिकीकरण की वजह से विकसित और विकासशील देशों के नागरिकों के अवकाश का समय भी बढ़ा है, जिससे नागरिकों को खेल समारोहों में भाग लेने और दर्शक के रूप में मैदानों तक पहुंचने, एथलेटिक गतिविधियों में अधिक से अधिक भागीदार करने और उनकी पहुंच बढ़ी है। प्राचीन ओलंपिक्स से वर्तमान सदी तक खेल आयोजित किये जाते रहे हैं और उनका विनियमन भी होता रहा है। औद्योगिकीकरण की वजह से विकसित और विकासशील देशों के नागरिकों के अवकाश का समय भी बढ़ा है, जिससे नागरिकों को खेल समारोहों में भाग लेने और दर्शक के रूप में मैदानों तक पहुंचने, एथलेटिक गतिविधियों में अधिक से अधिक **भागीदरी** करने और उनकी पहुंच बढ़ी है। मास मीडिया और वैश्विक संचार माध्यमों के प्रसार से ये प्रवृत्तियां जारी रहीं. व्यवसायिकता की प्रधान हुई, जिससे खेलों की लोकप्रियता में वृद्धि हुई, क्योंकि खेल **प्रशसकों** ने रेडियो, टेलीविजन और इंटरनेट के माध्यम से व्यावसायिक खिलाड़ियों के खेल का बेहतरीन आनंद लेना शुरू किया।

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Output 3

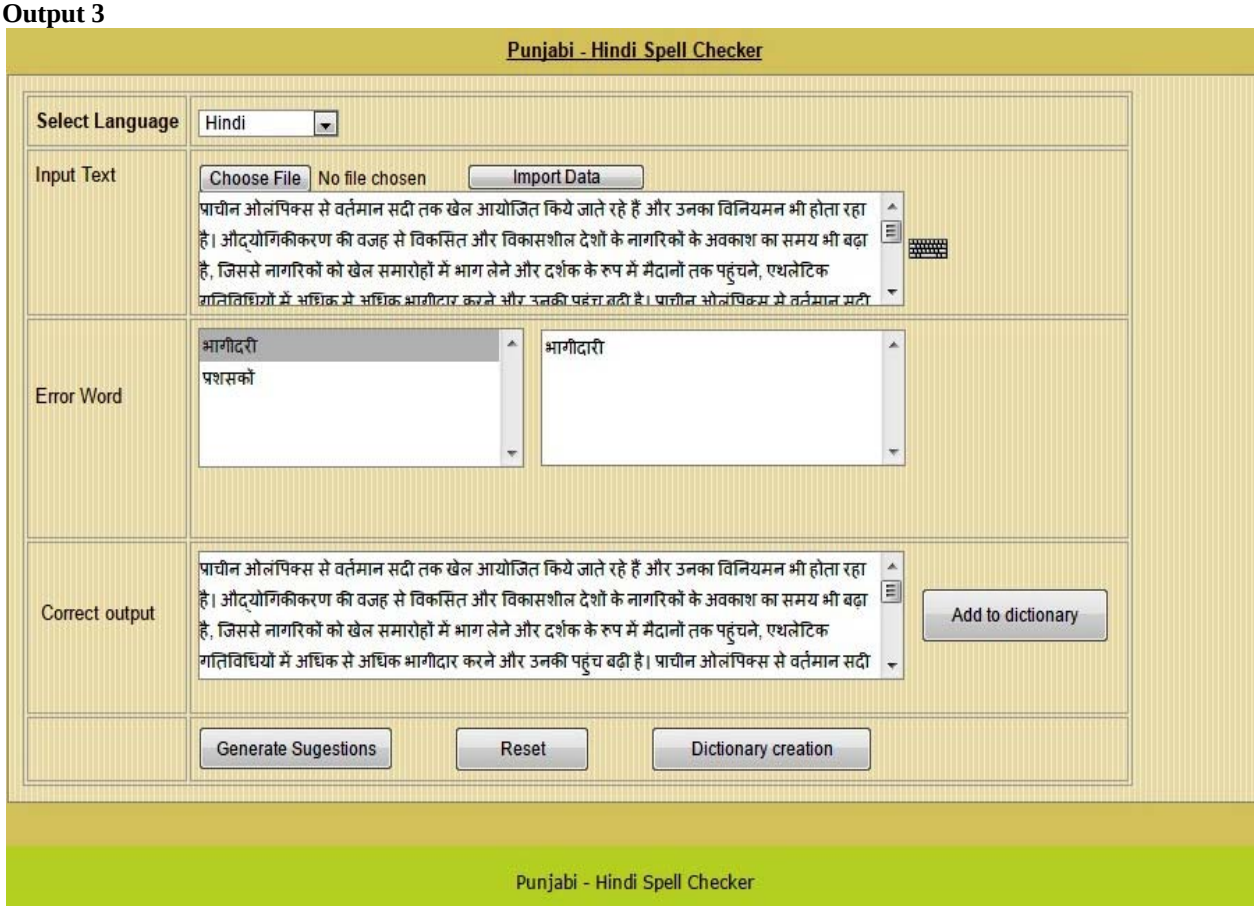

Fig:-1.8

Input 4

1947 भारत का विभाजन के बाद बर्तानवी भारत के पंजाब सूबे को भारत और पाकिस्तान दरमियान विभाजन दिया गया था। 1966 में **भरतीय** पंजाब का विभाजन फिर से गो गई और नतीजे के तौर पर **हरियणा** और हिमाचल प्रदेश होंद में आए और **पंजब** का मौजूदा राज बना। यह भारत का अकेला सूबा है जहाँ सिख बहुमत में हैं। कृषि पंजाब का सब से बड़ा उद्योग है, यह भारत का सब से बड़ा गेहूँ उत्पादक है। और प्रमख उद्योग हैं। वैज्ञानिक साज़ों, कृषि, खेल और बिजली सम्बन्धित माल, सिलाई मशीनें, मशीन यंत्रों, स्टर्च, साइकिलों, खादें आदि का निर्माण, वित्तीय रोज़गार, सैर-सपाटा और देवदार के तेल और खंड का उत्पादन।

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

Output 4

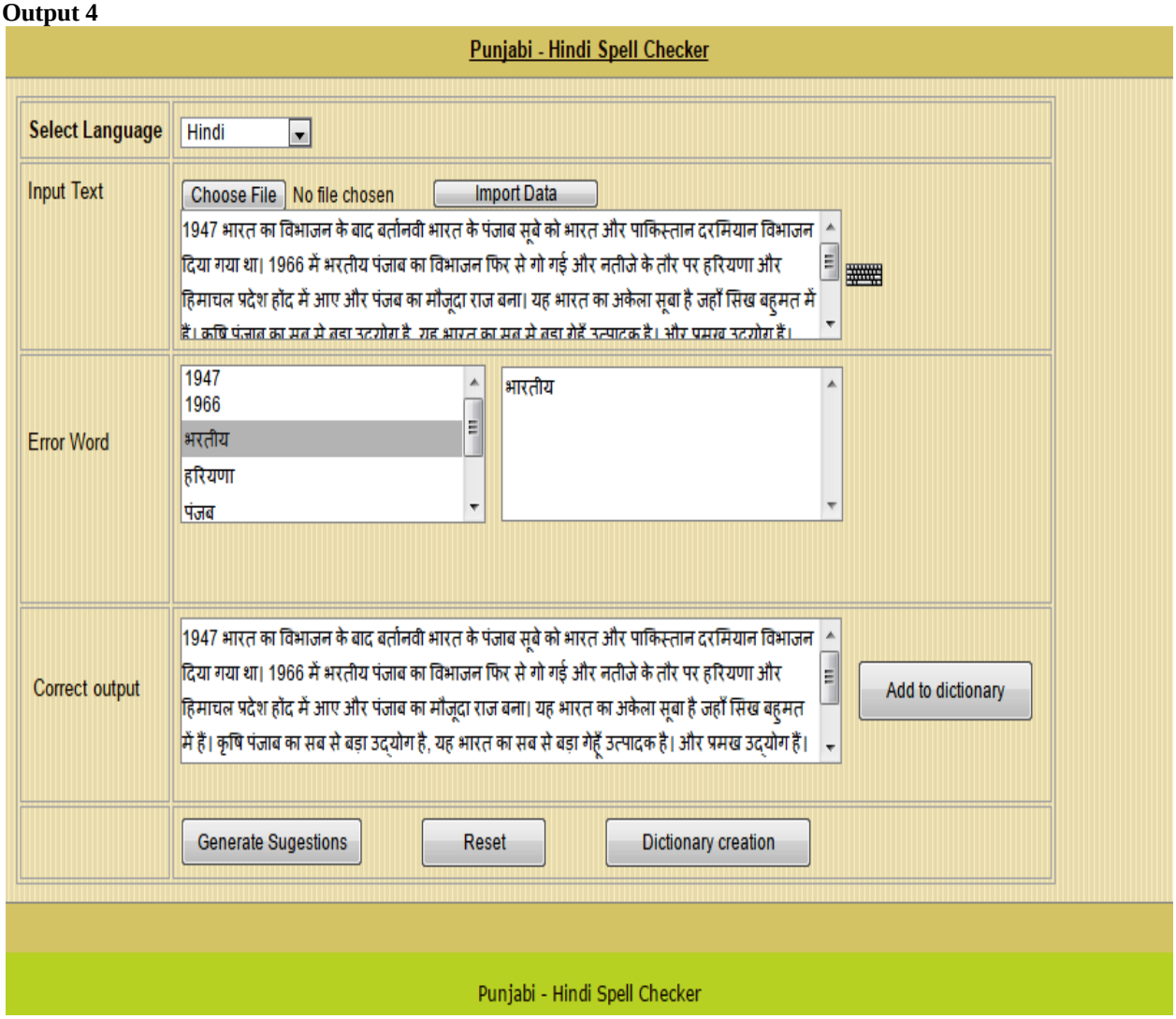

Fig:-1.9

5. CONCLUSION AND FUTURE SCOPE

In this Research work, I have developed an online Punjabi-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 three approaches which are rule based approach, dictionary look up approach and edit distance approaches are used into one. The main features of Punjabi-Hindi spell checker are large database, online application, easy to operate, email and printing options. 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. For future scope, this system can be enhanced for the complex sentences. The accuracy of the system can also be improved. For further research, some grammatical rules like the combinations of noun, verb, and adverb may be added. This system can be used for also other languages with modification of dictionary and keyboard.

7. REFERENCES

[1] Ritika Mishra, Navjot Kaur, August (2013), "Design and Implementation of Online Punjabi Spell Checker Based on Dynamic Programming", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 8.

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

[3]Gurpreet Singh Lehal, (2007) "Design and Implementation

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India

of Punjabi Spell Checker", International Journal of Systemic, Cybemetics and Informatics, pp.70-75.
[4]G S Lehal & Meenu Bhagat, "Spelling Error Pattern Analysis of Punjabi Typed Text", In Proceedings of the 2007 International Symposum on Machine Translation, NLP and TSS, pp. 128-141.
[5]Youssef Bassil & Mohammad Alwani May (2012), "Context-sensitive Spelling Correction using Google Web IT 5-Gram Information," Department of Computer and Information Science, Vol. 5, No.3.
[6]Ritu aggrawal, September (2007), "Hindi editor with spell checker", Vinayaka Mission University, Salem.
[7]Hindi spell checker available at https://addons.mozilla.org/en-US/firefox/addon/hindi-spell-checker/
[8]Hinkhoj spell checker available at http://dict.hinkhoj.com/spell-checker/check-spelling.php
[9] Spell guru available at http://bhashagiri.com/
[10]. Meenu Bhagat, (2007), "Spelling Error Pattern Analysis of Punjabi Typed Text", Thesis report, Thapar University, Patiala.
[11]Amit Sharma &Pulkit Jain, "Hindi Spell Checker", Indian Institute of Technolog Kanpur, April 17, 2013
[12] 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.
[13] R.E. Gorin (1971), "SPELL: A spelling checking and correction program", Online documentation for the DEC-10 computer.
[14] 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.

978-93-85670-72-5 © 2016 (RTCSIT)

CONFERENCE PAPER
International Conference on
Recent Trends in Computer Science & Information Technology (RTCSIT-2016)
21st August 2016
Guru Nanak College Budhlada, Punjab India