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 is 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 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. 
   \( \text{If}^2 \text{Jo} \) \( \text{Ijo gfo} \text{tko} \) \( \text{gfo} \text{tko} \)

2. Deletion error(DE): when at least one character is deleted in the desired word.
   \( g\text{fotko} \) \( \text{gfo} \) \( \text{rB} \) \( rZB \), \( \text{अपमान} \) \( \text{आपमान} \)

3. Substitution error(SE): when at least one character is substituted by the other character.
   \( g\text{fotko} \) \( g\text{fotBo} \), \( \text{शीत} \) \( \text{भारु} \), \( \text{शीमा} \) \( \text{शीमा} \)

978-93-85670-72-5 © 2016 (RTCSIT)
4. Transposition error (TE): when two adjacent characters are transposed. 
gfotko ^ gfotok , oks ^ osk  eow ^ ewo
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, अच्छा अच्छा - अच्छा अच्छा
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 ^ gfo tko, महाराजा ^ महाराजा, dhtko ^ dh 
tko, दुखी - दुखी

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)](image-url)
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.

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

**Dictionary Creation**

![Dictionary Creation](image1.png)

Fig:-1.4 (Dictionary creation tool For Hindi)

![Dictionary Creation](image2.png)

Fig:-1.5(Dictionary creation tool for Punjabi)

4. RESULTS

Input 1

Input the words from Table 1.1 to Table 1.2. The words will be added to the database of the spell checker.
Output 1

**Punjabi - Hindi Spell Checker**

<table>
<thead>
<tr>
<th>Select Language</th>
<th>Punjabi</th>
</tr>
</thead>
</table>

**Input Text**

Fig: 1.6

**Error Word**

| ਪੂਰਬ | ਬੰਗਾਲ ਦੀ ਖਾੜੀ ਹੈ ਅਤੇ ਪੰਛਮ ਵਿਚ ਅਰਬ ਸਗਰ ਹੈ। ਭਾਰਤ ਵਿਚ ਕੇਵਲ ਵਾਲੀਆਂ ਨਦੀਆਂ ਹਨ। ਭਾਰਤ ਦੇ ਦੋ ਅਧਿਕਾਰਕ ਨਾਮ ਹਨ ਹੰਦੀ ਵਿਚ ਭਾਰਤ ਅਤੇ ਅੰਗਰੇਜੀ ਵਿਚ ਇਡੀਆ ਹਨ। ਭਾਰਤ ਵਿਚ ਲਾਗੂ ਕੀਤੀਆਂ நੋਂ ਮੰਨੀ ਜਾਂਦੀ ਹੈ। ਭਾਰਤ ਸੰਸਾਰ ਦਾ ਸਬ ਵੱਡਾ ਲੋਕਤੰਤਰ ਹੈ। ਇੱਥੇ ਜ਼ਿਆਦਾ ਭਾਵਾਂ ਬੋਲ ਜਾਂਦੀਆਂ ਹਨ। ਸੰਸਾਰ ਦੇ ਚਾਰ ਧਰਮ ਹੰਦੂ, ਬੁੱਧ, ਜੈਨ ਅਤੇ ਇਸੱਖ ਦਾ ਜਨਮ ਅਤੇ ਵਿਕਾਸ ਭਾਰਤ ਵਿਚ ਹੋਇਆ। ਭਾਰਤ ਭੂਗੋਲਕ ਖੇਤਰਫਲ ਦੇ ਅਧਾਰ ਤੇ ਸੰਸਾਰ ਦਾ ਤਹਾਂ ਸਭ ਵੱਡਾ ਰਾਸਤਰ ਹੈ।
Output 2

Punjabi - Hindi Spell Checker

Input 3

प्राचीन ओलंपिक्स से वर्तमान सदी तक खेल आयोजित किये जाते रहे हैं और उनका विनियमन भी होता रहा है। ओलम्पिकीकरण की वजह से विकसित और विकाशशील 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 पहुँचने, एथलेटिक गतिविधियों में अधिक से अधिक भागीदार बनने और उनकी पहुंच बढ़ी है। प्राचीन ओलंपिक्स से वर्तमान सदी तक खेल आयोजित किये जाते रहे हैं और उनका विनियमन भी होता रहा है। ओलम्पिकीकरण की वजह से विकसित और विकाशशील 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 पहुँचने, एथलेटिक गतिविधियों में अधिक से अधिक भागीदार बनने और उनकी पहुंच बढ़ी है। मास मीडिया और वैश्विक संचार माध्यमों के प्रसार से ये प्रवृत्तियां जारी रहीं। व्यवसायिक का प्रयास हुई, जिससे खेलों की ओलिम्पियाट में वृद्धि हुई, क्योंकि खेल प्रशस्ती के रूप में टेलीविजन और इंटरनेट के माध्यम से व्यावसायिक खिलाड़ियों के खेल का बेहतरीन आनंद लेना शुरू किया।
भारत का विभाजन के बाद बैंक, भारत और पाकिस्तान दरमियान विभाजन दिया गया था। 1947 में भारतीय पंजाब का विभाजन फिर से गई और लोग भी। नतीजे के पीछे हार्दिक और बिहार विभाजन प्रदेश हो गए और पंजाब का मौजूदा राज बना। यह भारत का अकेला सूबा है जहाँ केवल बहुमत है।

पंजाब का सबसे बड़ा उदयोग है, यह भारत का सब से बड़ा गेहूँ उत्पादक है। और प्रमुख उदयोग है। वैज्ञानिक सातों, कृषि, खेत और विज्ञान सम्बन्धित माल, सिलाई मशीनों, मशीन यंत्रों, स्टर्थ, साइकिलों, खाद आदि का निर्माण, वित्तीय रोजगार, सैर-सपाटा और टेक्स्चर के तेल और खड़ा का उत्पादन।
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

of Punjabi Spell Checker”, International Journal of Systemic, Cybemetics and Informatics, pp.70-75.


