



ISSN: 2350-0328

**International Journal of Advanced Research in Science,
Engineering and Technology**

Vol. 2, Issue 11 , November 2015

Spellchecking and Error Correcting System for text paragraphs written in Punjabi Language using Hybrid approach

Amanjot Kaur, Dr.Paramjeet Singh,Dr.Shaveta Rani

M.tech Student, CSE Department, GZSCCET, Bathinda
Associate Professor, CSE Department, GZSCCET, Bathinda
Associate Professor, CSE Department, GZSCCET, Bathinda

ABSTRACT: Spell-checking is the process of detecting and correcting incorrect spelled words in a paragraph. Spell checking system first detects the incorrect words and then provide the best possible solution of corrected words. Spell checking system is a combination of handcrafted rules of the language for which spell checking system is to be created and a dictionary which contain the accurate spellings of various words. Better rules and large dictionary of words is help to improve the rate of error detection otherwise all the errors cannot be detected. After detecting the wrong or misspelled words, the various spell correcting techniques are used to provide the best accurate correct words or alternate words which higher the rate of correction of the wrong words. There are many systems available for detecting and correcting text. 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.

KEYWORDS: Spellchecker, Punjabi, Error, dictionary.

I INTRODUCTION

Spell-checking is a very important part of any translation system in Natural Language Processing. It is the process of detecting errors in a paragraph of wrongly spelled words and provides the alternatives to correct them. 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 corpus of a spell-checker is, the system can check the wrong spelled words with higher rates; otherwise, wrong spelled words are not detected properly. Conventional corpuses experience from out-of-vocabulary and data sparseness problems as they do not cover large vocabulary of words which are needed to make available proper names, domain-specific terms, technical lexica, special acronyms, and terminologies etc. As a result, spell-checkers will have the less error correction and detection rate and will flop to encounter all misspelled words in the data or text. All latest mercantile spelling error detection and correction tools works on word level and used a corpus. Each word from the data or text is searched in the dictionary. When a word is not found in the dictionary lexicon, it is assumed as an error word. In order to precise the error, a spell checker system search words in the corpus which are mostly similar to the error word. These words then listed for user to select appropriate one. Spelling checking is used in various applications like machine translation, searches, information retrieval etc. There are two main terms which are related to spell checking and correction systems. First is error detection and second is error correction. In developing upon the type of error non word error and real word error. There are many systems available for detecting and correcting text. 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



ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 11 , November 2015

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. We will use hybrid approach to implement the Spelling checking and Correcting System. This hybrid approach is a combination of various approaches like “Dictionary look up approach” that can be used to locate the errors, “Rule based approach” can be used for preprocessing, “Statistical Approach” to select the best alternative among the generated suggestions, “Edit Distance approach” to generate the valid suggestions for the wrong spelled word and use linguistic features of the Punjabi language.

These approaches can be explained in brief as follows:

A. 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.

B. Edit Distance

Edit distance is almost effective technique to generate the alternates of wrongly spelled words. In this approach word containing the spelling mistake is compared to every word of the dictionary and various operations like insertions, deletions and updation are performed on the word corresponding to the every word of the dictionary. The total number of such operations is referred to as the distance. The minimum the distance the higher the possibility of the word to became target word.

C. 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. By the help of these rules the system attempts to generate the exact spellings of the word which is under observation.

D. Statistical Analysis

This works when rule based approach fails to generate the possible correct 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.

II. LITERATURE REVIEW

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 spell checker and also made 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.



ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 11 , November 2015

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 Engineering

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.

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

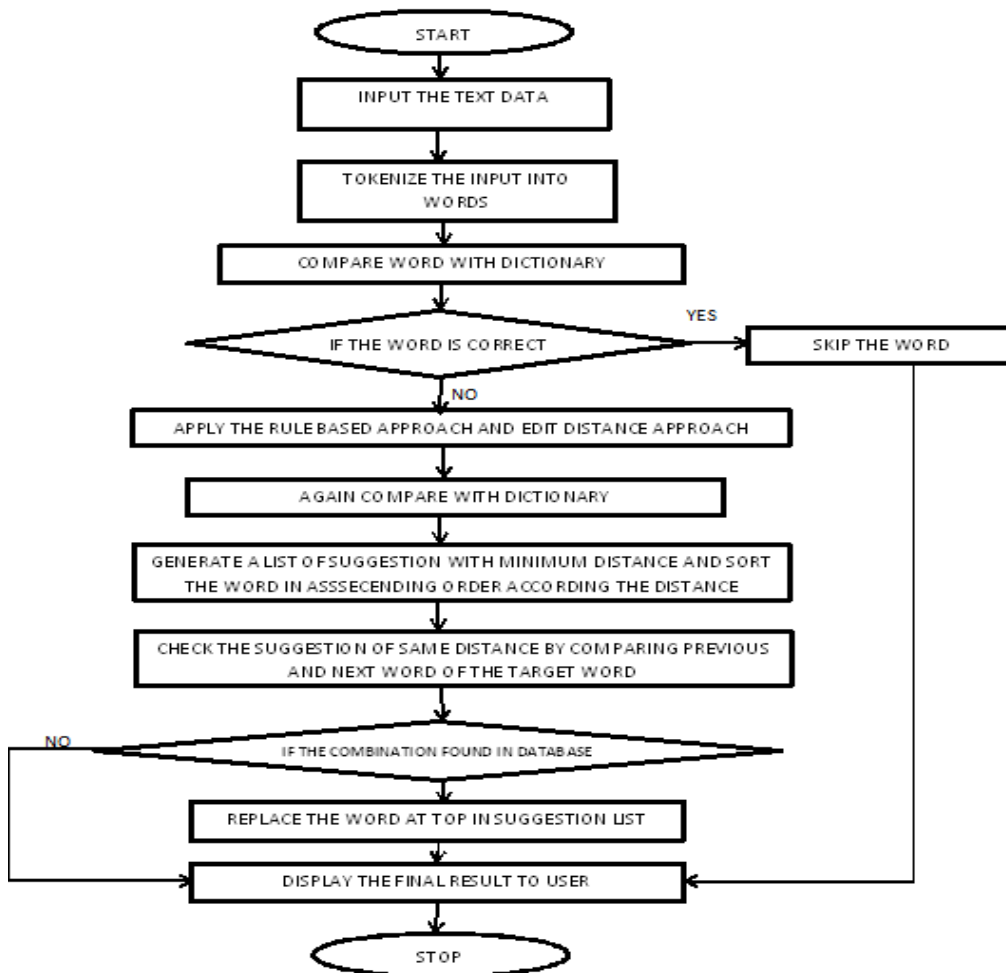
In this paper we have surveyed the area of spell correction and error detection techniques. Existed work related with spell checkers in Punjabi and Punjabi language is also discussed. In this paper the author will implement a Punjabi spell-checker by using dictionary lookup and edit-distance based technique with more accuracy. In this paper techniques for Error Correction are used (1) N Gram Analysis (2) Rule Based Approach and (3) Edit Distance.

III. RESEARCH 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”, “Statistical Approach”, “Edit Distance Technique” and use linguistic features of the Punjabi language. Dictionary Look Up Approach and Edit Distance Approach is used in the research which is already implemented. 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

A. 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.
- StepVIII 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 goto step VII.
- Step IX End.



B. Rule Based Approach

In this approach a set of rules are developed by which the input token is compared. Every rule created in the rule based is applied on the token and corresponding results is produced after applying the rule based approach.

The following are the steps of rule based approach:

- Step I : input the Text to find the errors
- 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 Based Approach.
- Step v: use output of rule based system into next phase
- Step VI: End

C. 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.

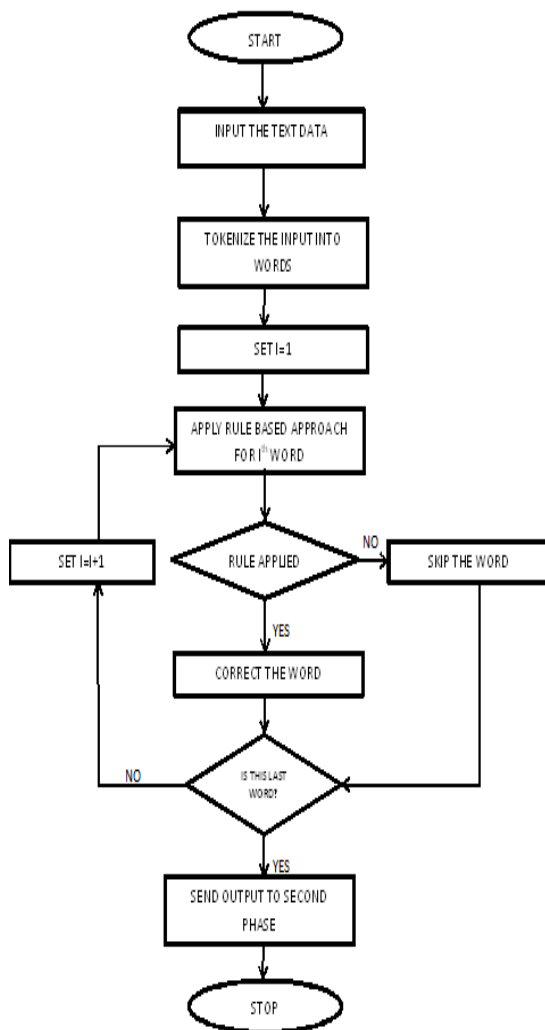
Steps for dictionary look up technique are as follows:

- Step I : input the String data to be checked

Step II :Tokenize this input into words

Step III Compare it with the dictionary to check whether it is correct or not.

Step IV End



FLOW CHART FOR RULE BASED APPROACH

D. 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.

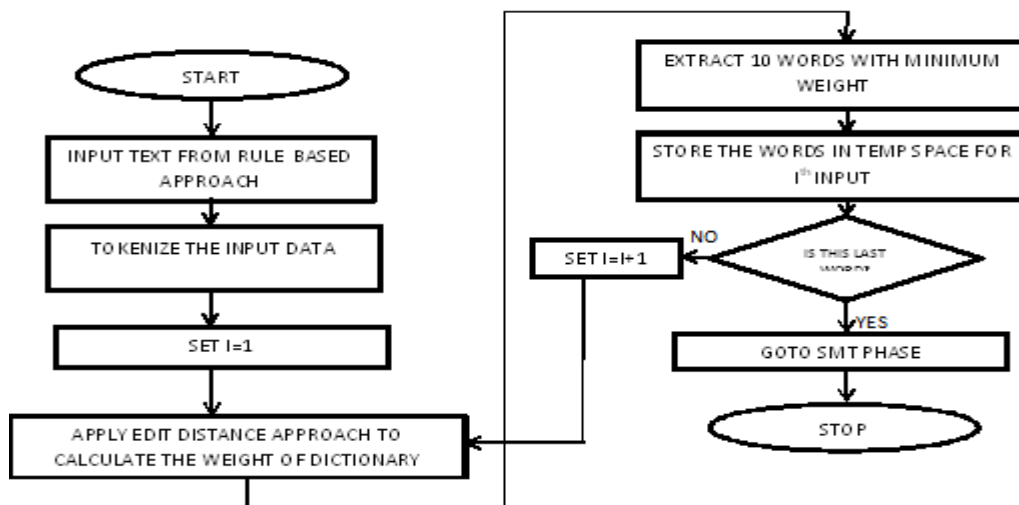
The steps to implement this technique are as follows:

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



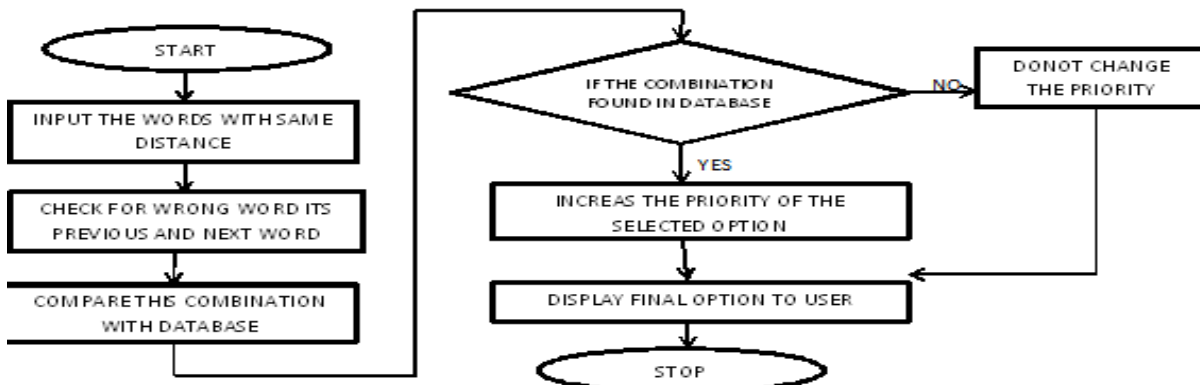
FLOW CHART FOR EDIT DISTANCE APPROACH

E. Statistical Machine Translation:

This works when rule based approach fails to generate the best possible correct 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.

The steps to implement this technique are as follows:

- Step 1: Input the word from edit distance phase.
- Step 2: Compare i-1, ith and i+1 word.
- Step 3: if the combination found in database then change the priority.
- Step 4: Otherwise goto the next word.
- Step 5: End



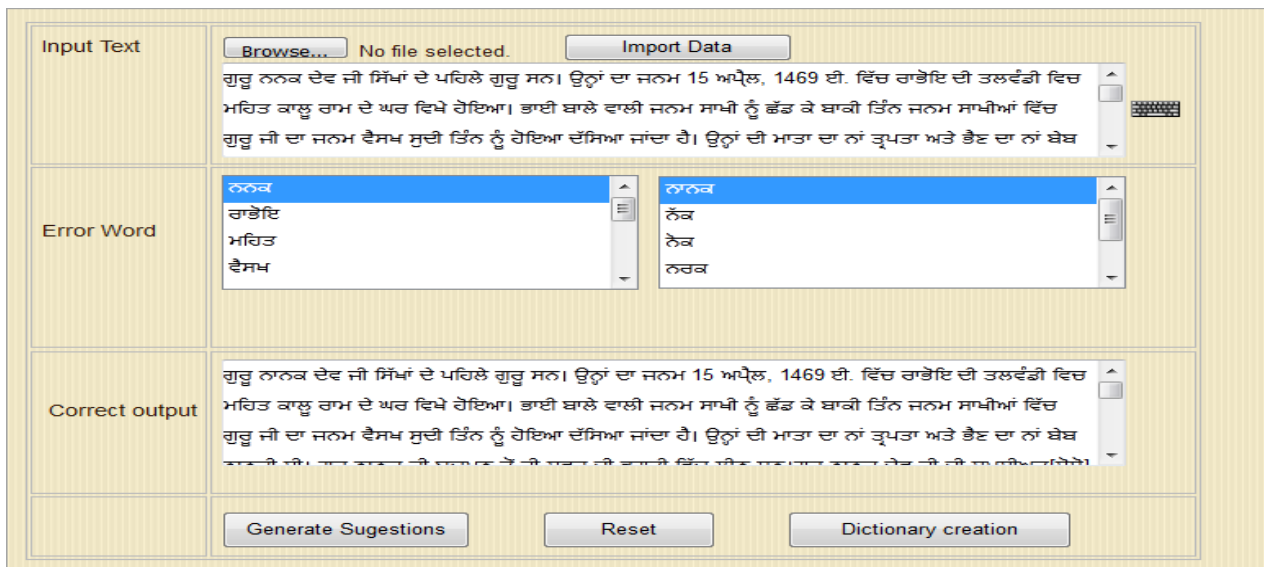
FLOW CHART FOR SMT APPROACH

IV. RESULT

Input:

ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਜੀ ਸਿੱਖਾਂ ਦੇ ਪਹਿਲੇ ਗੁਰੂ ਸਨ। ਉਨ੍ਹਾਂ ਦਾ ਜਨਮ 15 ਅਪ੍ਰੈਲ, 1469 ਈ. ਵਿੱਚ ਰਾਭੋਇ ਦੀ ਤਲਵੰਡੀ ਵਿੱਚ ਮਹਿਤ ਕਾਲੂ ਰਾਮ ਦੇ ਘਰ ਵਿਖੇ ਹੋਇਆ। ਭਾਈ ਬਾਲੇ ਵਾਲੀ ਜਨਮ ਸਾਖੀ ਨੂੰ ਛੱਡ ਕੇ ਬਾਕੀ ਤਿੰਨ ਜਨਮ ਸਾਖੀਆਂ ਵਿੱਚ ਗੁਰੂ ਜੀ ਦਾ ਜਨਮ ਵੈਸਖ ਸੁਦੀ ਤਿੰਨ ਨੂੰ ਹੋਇਆ ਦੱਸਿਆ ਜਾਂਦਾ ਹੈ। ਉਨ੍ਹਾਂ ਦੀ ਮਾਤਾ ਦਾ ਨਾਂ ਤ੍ਰਪਤਾ ਅਤੇ ਭੈਣ ਦਾ ਨਾਂ ਬੇਬ ਨਾਨਕੀ ਸੀ। ਗੁਰੂ ਨਾਨਕ ਜੀ ਬਚਪਨ ਤੋਂ ਹੀ ਸ਼ਵਰ ਦੀ ਭਗਤੀ ਵਿੱਚ ਲੀਨ ਸਨ। ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਜੀ ਦੀ ਸ਼ਖਸੀਅਤ [ਸੋਧੋ] ਗੁਰੂ ਨਾਨਕ ਦੇਵ ਸੰਸਾਰ ਦੇ ਉਨ੍ਹਾਂ ਮਹਾਨ ਵਿਅਕਤੀਆਂ ਵਿੱਚੋਂ ਹਨ, ਜਿਨ੍ਹਾਂ ਨੂੰ ਆਪਸ ਵਿੱਚ ਵਿਰੋਧ ਰੱਖਣ ਵਾਲੀਆਂ ਦੇਵੱਖ-ਵੱਖ ਕੌਮਾਂ ਨੇ ਪੂਰਾ-ਪੂਰਾ ਸਨਮਾਨ ਦਿੱਤਾ। ਉਹ ਸਿੱਖ ਧਰਮ ਦੇ ਬਾਨੀ, ਇੱਕ ਪਰਮਾਤਮਾ ਦੀ ਭਗਤੀ ਕਰਨ ਵਾਲੇ, ਸਾਰੇ ਸੰਸਾਰ ਨੂੰ ਇੱਕ ਸੂਤ ਵਿੱਚ ਪਿਰੇਇਆ। ਵੇਖਣ ਦੇ ਚਾਹਵਾਨ, ਦੀ ਨਦੁਖੀਆਂ ਦੇ ਸਮਰਥਕ ਅਤੇ ਮਹਾਨ ਸਮਾਜ ਸੁਧਾਰਕ ਸਨ। ਉਹ ਅਜਿਹੇ ਮਹਾਪੁਰਸ਼ ਸਨ, ਜਿਨ੍ਹਾਂ ਨੇ ਧਰਮ ਨਿਰਪੱਖਤਾ ਦਾ ਪ੍ਰਚਾਰ ਕੀਤਾ। ਗੁਰੂ ਜੀ ਇੱਕ ਸਮੁੱਚੇ ਸਾਹਿਤਕਾਰ, ਪਰਬੀਨ ਆਗੂ, ਵਿਸ਼ਵ ਧਰਮ ਦੇ ਨਿਰਮਾਤਾ, ਹਮਦਰਦ, ਨਿਰਭੈ, ਨਿਰਵੈਰ ਅਤੇ ਧਰਮ ਮਨੁੱਖ ਸਨ। ਉਨ੍ਹਾਂ ਦੀ ਚੁੰਬਕੀ ਸ਼ਖਸੀਅਤ ਦੀ ਪਾਰਸਫੂ ਹਨਾਲਕਈ ਹੋਰ ਵਿਅਕਤੀ ਵੀ ਵਿਲੱਖਣ ਸ਼ਖਸੀਅਤ ਦੇ ਮਾਲਕ ਬਣ ਗਏ।

Output:



The screenshot shows a software interface with the following components:

- Input Text:** A text area containing the input paragraph. Above it are buttons for "Browse..." (with "No file selected." below it) and "Import Data".
- Error Word:** A list of words identified as errors: ਨਨਕ, ਰਾਭੋਇ, ਮਹਿਤ, ਵੈਸਖ.
- Correct output:** A text area showing the corrected version of the input text, where the error words have been replaced by their correct forms: ਨਾਨਕ, ਨਾਨਕ, ਨੇਕ, ਨੇਕ, ਨਰਕ.
- Buttons:** "Generate Sugestions", "Reset", and "Dictionary creation" are located at the bottom of the interface.



ISSN: 2350-0328

International Journal of Advanced Research in Science, Engineering and Technology

Vol. 2, Issue 11 , November 2015

REFERENCES

- [1] 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
- [2] Neha Gupta, PratisthaMathur, *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 Engineering
- [3] 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.
- [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, Cybematics and Infomatics, 2007.
- [8] Amit Sharma &Pulkit Jain, "Hindi Spell Checker", Indian Institute of Technology Kanpur, April 17, 2013.
- [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 Symposium on Machine Translation, NLP and TSS, pp. 128-141, 2007.
- [17] Jesus Vilares& Manuel Vilares, "Managing Misspelled Queries in IR Application," Issue 8, October 2010.
- [18] Youssef Bassil& Mohammad Alwani, "Context-sensitive Spelling Correction using Google Web IT 5-Gram Information," Department of Computer and Information Science, Vol. 5,No.3, May 2012.