# Punjabi Spell Checker Using Dictionary Clustering

Harpreet Kaur[1], Gurpreet Kaur[2], Manpreet Kaur[3]

*M.tech. Student CSE Department, Assistant professor CSE Department, Assistant professor CSE Department*

*BBSBEC Fatehgarh Sahib,India , BBSBEC Fatehgarh Sahib,India, BBSBEC Fatehgarh Sahib,India*

***Abstract*— Spell checker is a fundamental need of every word processing document that examines the input text for misspelled words and provides possible suggestion for wrong word. In spell checking area, lot of work has been done in English language. Punjabi is the 10th most widely spoken language of the world. The design issues of English and Punjabi languages are different. The available Punjabi spell checkers use dictionary lookup detection technique and different correction techniques. The aim of this paper is to improve the processing speed and reduce the access time. It presents a different approach of spell checking using dictionary clustering. Dictionary is organized as clusters. Clustering is done on the basis of noun, pronoun, verb and adjective of Punjabi language. The category of the input text is identified at run time and that word is searched in the predicted cluster, which reduces the processing time. The execution time of the input text with and without clustering is compared. It is for the first time that Punjabi spell checker with dictionary clustering has been developed. Also the execution time of single word, a line and a paragraph is compared. The system detects approximately 83.3 percent of the wrong words and provides 98 percent of the correct suggestions for the misspelled words.**

***Keywords—Types of errors ; Punjabi spell checker ; error detection; error correction; dictionary clustering ;***

## I. Introduction

Spell checker is a software program that identifies the misspelled words in the input text by checking in the database and then provides possible suggestions for replacement of the identified wrong word. Now Spell checker becomes vital component of software like word processor, web browser and others. Spell checker has two main functions. The first one is to identify the misspelling performed by the user. Misspelling occurs due to poor writing skills of writer and spelling proficiency. The source data also give a challenge for good performance of the spell checker. The document produced by non-native speakers is more prone to errors than the native speakers. So vast range of spelling possibilities arises and spell checker should be instructed to identify.

The second function of the spell checker is to give possible suggestions for the identified wrong word and replace it in the input text. Main steps followed by the spell checker:-

- Take the input word from a file or database.
- Check the availability of the word in the database.
- If the word found in the database, then next word is searched.
- If the word has not found in the database, then nearest matching is given for the replacement of the identified wrong word.

Thus these steps are commonly followed by all spell checkers. The available Punjabi spell checker use dictionary lookup technique for error detection, in which every word of the database is searched to find the input words.

*Types of errors*

To classify the type of errors different studies have performed. The word error can be classified into two main categories:-

Real word and non-real word error

A *real* word error is a valid word that is acceptable in the lexicon but not correct according to the sentence.

mYN **Awpxw** dosq kol **ilAw**[

For

mYN **Awpxy** dosq kol **igAw**[

Awpxw and ilAw are acceptable words in the Punjabi lexicon but not correct according to the sentence.

A non real word error is not a valid word and cannot be found in the lexicon.

aus kol **bhq pYS**y hn[

For

aus kol **bhuq** pYsy hn[

bhq and pYSy are invalid words for the Punjabi lexicon. There are 7 types of errors are analyzed in Punjabi typed text:

- Insertion error (IE): When at least one extra character is inserted in the desired word. `iSMgwr -iSMngwr`
- Deletion error (DE): When at least one character is deleted in the desired word. `iSMgwr - iSgwr`
- Substitution error (SE): When at least one character is substituted by the other character. `iSMgwr - iSMGwr`
- Transposition error (TE): When two adjacent characters are transposed. `iSMgwr - iSMgrw`
- Run-on error (ROE): When there is space missing between two or more valid words. `iSMgwr krvwE- iSMgwrkrvwE`
- Split Word error (SWE): This is Opposite of Run-on error when some extra space is inserted between alphabets of a word. This error can be resolved by removing the extra space. `iSMgwr – iSMgw r`

*Error detection*

In every spell checker it is important task to detect the misspelled words. There are various non word error detection techniques. Spell checkers usually use dictionary lookup technique and N- gram analysis technique is used in OCR problems. Punjabi spell checkers use dictionary lookup technique.

*Error correction*

Correction of detected misspelled words is consists of two steps: generation of the suggestion list and ranking of the list. Some error correction techniques can omit the second step leaving the ranking and final selection to the user. The commonly used spelling correction algorithms are minimum edit distance, similarity keys, rule-based, n-gram-based, probabilistic and neural networks. All these algorithms based on the concept of calculating distance between misspelled word and the word in the dictionary. Raking of the word is based on the distance, if the distance is small high the rank of the word.

## II. REALTED WORK

Lehal [1] specifies that designing a spell checker for Punjabi language poses problems that are totally different from the foreign language because Punjabi language has different phonetic properties and grammatical rules, which complicates the design of spell checker. Also there is no proper standardization for Punjabi keyboard layouts and fonts. Lehal and Bhagat [2] have done analysis of Punjabi typing errors and defined a pattern for typing errors. It is based on types of errors, positional effects and first position error, phonetic effects, word length effects. They have utilized the information collected from error analysis to design a single error based Suggestion list for a Punjabi spell checker. Kaur and Bhatia [3] proposeed a spell checker for Gurmukhi script that is a standalone application. Dictionary creation tool is used to create dictionary which acts as a database for the spell checker. This tool is executed once. The system checks the input text for misspelled words and provides suggestions for the wrong words. It gives an option to add right words to the database that increases the size of database for future use. Mishra and kaur [4] also developed an online Punjabi spell checker Raftaar and also proposed a new algorithm for correction of wrong words. It gives 80% accuracy of words in this research work. Kaur and Garg [5] proposed a system which is hybrid combination for spell checker and grammar checker for Punjabi language. They developed a new algorithm. It is not easy to make Punjabi spell checker and grammar checker because there is no approved format of Punjabi spellings and lack of basic layout of Punjabi keyboard. To write a grammar checker for natural language dictionary of all words and parts of speech of every word are used. A grammar checker verifies each word, preprocess the word and assign parts of speech tags and make phrases.

Telugu is an agglutinating language and complex morphology with internal as well as external Sandhi. Uma Maheshwar Rao, G., Amba P. Kulkarni, Christopher Mala and Parameshwari [6] describe the external Sandhi in NLP and developed a spell checker. Rule based techniques are used because Telugu is a morphological rich language. They define that morphological validation by a Morphological Analyzer is the core component of the Telugu Spell-Checking and Sandhi splitter rules are defined. A spell checker includes two functions: set of routines and an algorithm for comparison of unrecognized words. Alkhafaji [7] according to his survey, there is no technique that can give 100% success rate in the situations where the misspelled word includes more than one type of errors. In his method, the situation with the more than one type of errors like one or more letter transposition, deletion and insertion has been manipulated.

Amorim and Zampieri [8] proposed a method that minimized distance calculations while finding a nearest possible match for the misspelled word. This method didn't remove a single word from the database. This method implements the dictionary clustering to improve access time and performance which makes the algorithm faster than other methods. Anomalous pattern initialization and partition around medoids (PAM) techniques are combined first time in developing spell checker. Youssef Bassil [9] described a shared-memory parallel spell-checking algorithm for detection and correction of misspelled words in typed text. Dictionary based approach has limited vocabulary and scattered data. So it is difficult to cover essential words and fails to detect all wrong words in the input text. The presented algorithm is based on Yahoo! N-Grams Dataset which contains trillions of words and n-grams, basically obtained from the internet. This method corrected 94% of the total errors, scattered as 99% non-word errors and 65% real- word errors.

All Rights Reserved © 2012 IJSETR

## III. PUNJABI SPELL CHECKERS

### *AKHAR*

AKHAR is the first Punjabi spell checker. To develop Punjabi spell checker detailed analysis of Punjabi errors is done. This spell checker is the part of the commercial Punjabi word processor AKHAR. It detects and corrects only non real word errors. It uses dictionary lookup error detection technique and reverse minimum edit distance error correction technique. The steps followed by the spell checker are:

- Upload the file and take content as input.
- Pre-process the word
- Look for the word in dictionary
- If the word is present in the dictionary, pass onto the next one.
- If the word is not found, then look for the closest matching patterns and present them in the form of suggestions [5].

### *SUDHAAR*

SUDHAAR is developed after AKHAR. SUDHAAR is a Punjabi word which means improvement or correction. It is a standalone spell checker for Gurmukhi script. It is an open source and uses dictionary lookup error detection technique and minimum edit distance error correction technique. It detects and corrects only non word errors and single word errors. The system detects approximately 87% of the errors and provides 90% of the correct suggestions [3].
Three modules are:

- Creation of Punjabi dictionary
- Error detection and correction
- Replacement

### *RAFTAAR*

Raftaar is a recently developed spell checker for Punjabi language. It is an online spell checker developed in an ASP.NET language with Sql Server 2005. The key features of Raftaar Punjabi spell checker are rich database, online feature, user friendly, email and printing option [4].

### *Spell checker using hybrid approach*

This system is a hybrid combination of spell checker and grammar checker for Punjab language. Primarily the system checks for spelling errors, then checks for grammatical errors in the text. When some input text is given to system, it is passed through spelling checker and grammar checker to give the input text free from both grammatical and spelling errors. A comparative good and efficient algorithm has been proposed which is hybrid combination of spell checker and grammar checker for Punjabi which saves time and cost [5].

## IV. MOTIVATION FOR WORK

Spell checker is very helpful when user works with word processing documents, e-mail and letters etc., user does not need to worry about the spelling errors. For the convenience of the writers spell checkers are designed which can be inbuilt program in the word processors or standalone application. There are already some spell checkers available in Punjabi language like AKHAR, SUDHAAR, Raftaar, Hybrid combination of spell checker and grammar checker, spell checker using unification method. All these spell checker use dictionary lookup technique for error detection and minimum edit distance technique for error correction. In dictionary lookup, it is feasible to store large-sized dictionaries in memory for instant access without storing the whole data on the disk. But, in dictionary-based techniques, similarity keys are used to compute the distance between the target wrong word and possible candidates correct words in the dictionary.

So Bigger the dictionary, accuracy may increases but number of calculations is also rises, which in turn make the algorithm's performance slower. We have one alternative to use dictionary arranged as clusters, which will improve the performance and access time.

We will develop an algorithm based on this clustering technique. The scope of our study is to develop a system which finds the correct word in clusters, made on the basis of noun, pronoun, verb and adjectives of Punjabi language.

## V. PROPOSED SCHEME

This paper proposes the clustering technique used with dictionary lookup, which makes the correction algorithm faster. To develop this spell checker following steps are followed:

1. Data collected from different sources and clustered database created.
2. Suffix and prefix rules made for pronoun, verb and adjectives.
3. Algorithm designed for detection and correction of wrong words using the rules.

### *Dictionary clustering*

Database has been created by collecting words from newspapers, dictionary and writers. All collected words have been grouped into four categories or clusters. Clusters are noun, pronoun, verb and adjective of the Punjabi language. To identify the category of the input word at run time different rules are made for the pronoun, verb and adjectives of language. The algorithm uses these rules in detection and correction process of spell checking. Rules have been generated according to suffix, prefix and length of the words [10]. When the rule matches for the input word its category has been identified. And that particular word will be searched in the identified cluster. So access time and execution time decreases and performance improves.

The system also allows the user to add word to dictionary which are correct but not present in the database. It will increase the accuracy of the system, database becomes rich in words. This system compares the execution time of single words, lines and paragraphs with and without clustering.

The category of the wrong word is identified at the run time by two methods. First using the suffix, prefix and length based rules and second is with location of the wrong word in the sentence. The word at the first location can be a noun or pronoun and word after the middle of the sentence can be verb or adjective.

The steps followed by the spell checker with dictionary clustering are shown below in Fig 1. Input file is uploaded and input text is given in the form of a word, a line or a paragraph. This word line or paragraph is scanned by the spell checker for the wrong by searching the database.

If wrong word found in the database then clustering algorithm applied on the wrong word. This algorithm identifies the cluster of the word on the basis of prefix, suffix and length rules. After identification of cluster, wrong word's intended correct word is found in the respective cluster. When a match found, the wrong word in the input text is replaced with correct word. If the word is not found in the database then it has the option to add that word in the database.

Fig. 2 describes the identification of the cluster on the basis of location of the wrong word in the sentence. If the location of the wrong word is first then in the most cases it can be a noun or pronoun according to Punjabi grammar. And if the wrong word found after the middle of the sentence then in the most cases it can be verb or adjective.

## VI. RESULTS AND DISCUSSION

Punjabi Spell checker with dictionary clustering has been implemented in visual basic dot net and MS Access at the back end. This system is offline spell checker.

Fig. 3 shows the main window of the spell checker. In this figure working of spell checker without clustering is shown. The time taken for execution is 1700.1488 milliseconds. Input can be a word, a line or a paragraph. Input can be given in two forms:-

- User can copy the input text from the destination and paste it in the text area of spell checker screen.

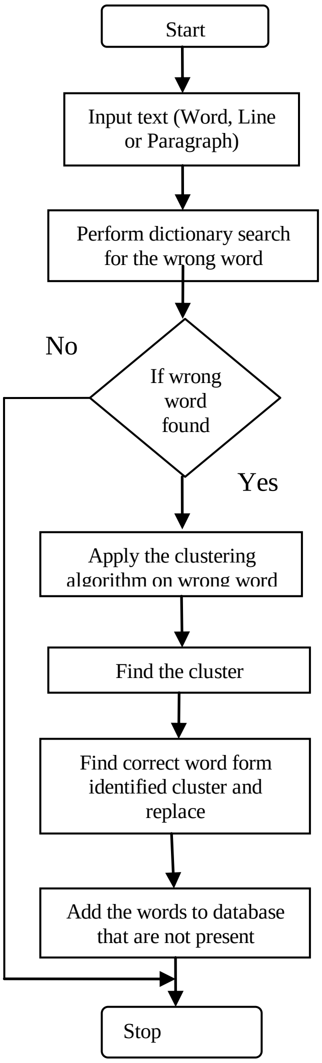

Start
Input text (Word, Line or Paragraph)
Perform dictionary search for the wrong word
No
If wrong word found
Yes
Apply the clustering algorithm on wrong word
Find the cluster
Find correct word form identified cluster and replace
Add the words to database that are not present
Stop

Fig. 1. Spell checker flowchart with clustering

- The user can upload the text file form the system by Upload File button of spell checker.

Spell check button detects the wrong words from the given input text. The wrong words are displayed in the grid view. Correct button replace the wrong words with the correct ones and display the error free file in the second rich text box.

Fig. 4 describes the spell checker with clustering and execution time taken by the spell checking process also shown. The execution time taken by spell checker is 486.2922 milliseconds. Also the words which are not present in the database and are valid words, added to the database in both the cases with and without clustering.

All Rights Reserved © 2012 IJSETR

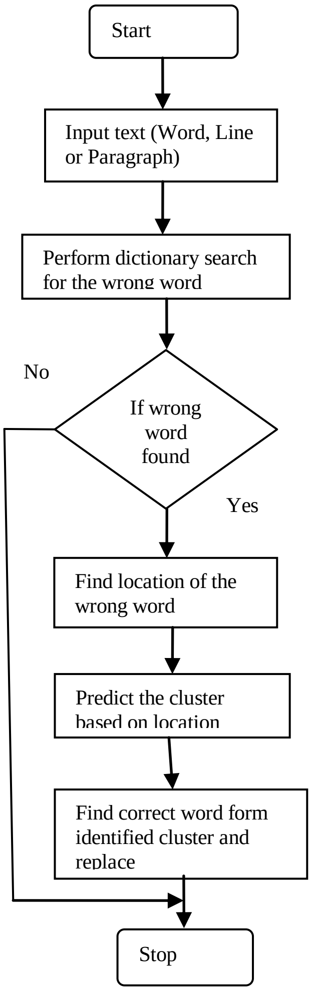

Start
Input text (Word, Line or Paragraph)
Perform dictionary search for the wrong word
No
If wrong word found
Yes
Find location of the wrong word
Predict the cluster based on location
Find correct word form identified cluster and replace
Stop

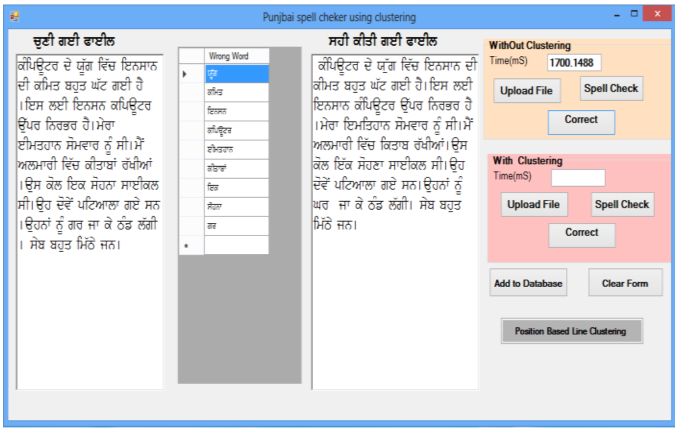
Fig. 2. Flowchart with position based cluster prediction

The TABLE I show the details of input text taken in this study. It describes the number of files that contains single words, number of files that contains lines and paragraphs.

The average execution time of spell checking process with and without clustering is shown in TABLE II. The average execution time taken by a word with clustering is four times, for a line five times and for a paragraph six times lesser than without clustering.

The input text with 250 words has been uploaded in the spell checker. And that text contains 60 wrong words. The spell checker detects 50 wrong words and added 30 words to the database. Correction module corrects 49 words out of 50 detected wrong words. TABLE III shows the accuracy of the system. The accuracy for detection is the ratio of detected wrong words to the total number of wrong words. And for correction it is the ratio of corrected words to the detected wrong words.

The spell checker detects 83.3% misspelled words and corrects 98% detected wrong words.


Punjbai spell cheker using clustering

ਚੁਣੀ ਗਈ ਫਾਈਲ

ਕੰਪਿਊਟਰ ਦੇ ਯੂੱਗ ਵਿੱਚ ਇਨਸਾਨ
ਦੀ ਕਮਿਤ ਬਹੁਤ ਘੱਟ ਗਈ ਹੈ
।ਇਸ ਲਈ ਇਨਸਨ ਕਪਿਊਟਰ
ਉੱਪਰ ਨਿਰਭਰ ਹੈ।ਮੇਰਾ
ਈਮਤਹਾਨ ਸੋਮਵਾਰ ਨੂੰ ਸੀ।ਮੈਂ
ਅਲਮਾਰੀ ਵਿੱਚ ਕੀਤਾਬਾਂ ਰੱਖੀਆਂ
।ਉਸ ਕੋਲ ਇਕ ਸੋਹਨਾ ਸਾਈਕਲ
ਸੀ।ਉਹ ਦੋਵੇਂ ਪਟਿਆਲਾ ਗਏ ਸਨ
।ਉਹਨਾਂ ਨੂੰ ਗਰ ਜਾ ਕੇ ਠੰਡ ਲੱਗੀ
। ਸੇਬ ਬਹੁਤ ਮਿੱਠੇ ਜਨ।

| | Wrong Word |
|---|---|
| ▶ | ਯੂੱਗ |
| | ਕਮਿਤ |
| | ਇਨਸਨ |
| | ਕਪਿਊਟਰ |
| | ਈਮਤਹਾਨ |
| | ਕੀਤਾਬਾਂ |
| | ਇਕ |
| | ਸੋਹਨਾ |
| | ਗਰ |
| * | |

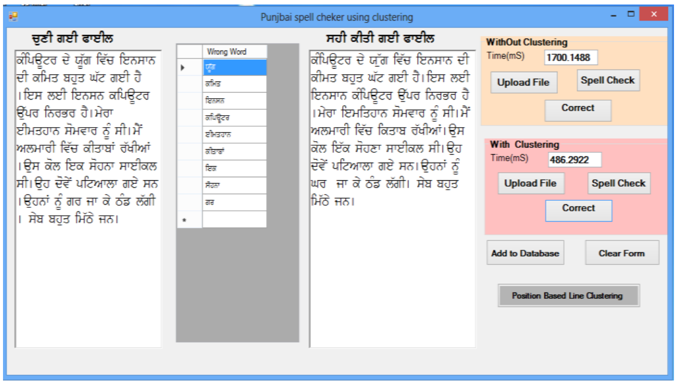
ਸਹੀ ਕੀਤੀ ਗਈ ਫਾਈਲ

ਕੰਪਿਊਟਰ ਦੇ ਯੁੱਗ ਵਿੱਚ ਇਨਸਾਨ ਦੀ
ਕੀਮਤ ਬਹੁਤ ਘੱਟ ਗਈ ਹੈ।ਇਸ ਲਈ
ਇਨਸਾਨ ਕੰਪਿਊਟਰ ਉੱਪਰ ਨਿਰਭਰ ਹੈ
।ਮੇਰਾ ਇਮਤਿਹਾਨ ਸੋਮਵਾਰ ਨੂੰ ਸੀ।ਮੈਂ
ਅਲਮਾਰੀ ਵਿੱਚ ਕਿਤਾਬ ਰੱਖੀਆਂ।ਉਸ
ਕੋਲ ਇੱਕ ਸੋਹਣਾ ਸਾਈਕਲ ਸੀ।ਉਹ
ਦੋਵੇਂ ਪਟਿਆਲਾ ਗਏ ਸਨ।ਉਹਨਾਂ ਨੂੰ
ਘਰ  ਜਾ ਕੇ ਠੰਡ ਲੱਗੀ। ਸੇਬ ਬਹੁਤ
ਮਿੱਠੇ ਜਨ।

WithOut Clustering
Time(mS) 1700.1488

With Clustering
Time(mS)

Fig. 3. User interface of spell checker (Without clustering)


Punjbai spell cheker using clustering

ਚੁਣੀ ਗਈ ਫਾਈਲ

ਕੰਪਿਊਟਰ ਦੇ ਯੂੱਗ ਵਿੱਚ ਇਨਸਾਨ
ਦੀ ਕਮਿਤ ਬਹੁਤ ਘੱਟ ਗਈ ਹੈ
।ਇਸ ਲਈ ਇਨਸਨ ਕਪਿਊਟਰ
ਉੱਪਰ ਨਿਰਭਰ ਹੈ।ਮੇਰਾ
ਈਮਤਹਾਨ ਸੋਮਵਾਰ ਨੂੰ ਸੀ।ਮੈਂ
ਅਲਮਾਰੀ ਵਿੱਚ ਕੀਤਾਬਾਂ ਰੱਖੀਆਂ
।ਉਸ ਕੋਲ ਇਕ ਸੋਹਨਾ ਸਾਈਕਲ
ਸੀ।ਉਹ ਦੋਵੇਂ ਪਟਿਆਲਾ ਗਏ ਸਨ
।ਉਹਨਾਂ ਨੂੰ ਗਰ ਜਾ ਕੇ ਠੰਡ ਲੱਗੀ
। ਸੇਬ ਬਹੁਤ ਮਿੱਠੇ ਜਨ।

| | Wrong Word |
|---|---|
| ▶ | ਯੂੱਗ |
| | ਕਮਿਤ |
| | ਇਨਸਨ |
| | ਕਪਿਊਟਰ |
| | ਈਮਤਹਾਨ |
| | ਕੀਤਾਬਾਂ |
| | ਇਕ |
| | ਸੋਹਨਾ |
| | ਗਰ |
| * | |

ਸਹੀ ਕੀਤੀ ਗਈ ਫਾਈਲ

ਕੰਪਿਊਟਰ ਦੇ ਯੁੱਗ ਵਿੱਚ ਇਨਸਾਨ ਦੀ
ਕੀਮਤ ਬਹੁਤ ਘੱਟ ਗਈ ਹੈ।ਇਸ ਲਈ
ਇਨਸਾਨ ਕੰਪਿਊਟਰ ਉੱਪਰ ਨਿਰਭਰ ਹੈ
।ਮੇਰਾ ਇਮਤਿਹਾਨ ਸੋਮਵਾਰ ਨੂੰ ਸੀ।ਮੈਂ
ਅਲਮਾਰੀ ਵਿੱਚ ਕਿਤਾਬ ਰੱਖੀਆਂ।ਉਸ
ਕੋਲ ਇੱਕ ਸੋਹਣਾ ਸਾਈਕਲ ਸੀ।ਉਹ
ਦੋਵੇਂ ਪਟਿਆਲਾ ਗਏ ਸਨ।ਉਹਨਾਂ ਨੂੰ
ਘਰ  ਜਾ ਕੇ ਠੰਡ ਲੱਗੀ। ਸੇਬ ਬਹੁਤ
ਮਿੱਠੇ ਜਨ।

WithOut Clustering
Time(mS) 1700.1488

With Clustering
Time(mS) 486.2922

Fig. 4. User interface of spell checker (With clustering)

TABLE I. DETAIL OF INPUT TEXT

| *Input Text* | *Number* |
|---|---|
| Single word | 35 |
| Line | 15 |
| Paragralph | 6 |

TABLE II. AVERAGE EXECUTION TIME

| Average execution time(in milliseconds) | | |
|---|---|---|
| *Input Text* | *Without clustering* | *With clustering* |
| Single word | 40.9527 | 10.9144 |
| Line | 456.6246 | 86.0015 |
| Paragralph | 3491.87 | 527.895 |

TABLE III. ACCURACY TABLE

| Accuracy | | |
|---|---|---|
| ***Input Text*** | ***Detection (%)*** | ***Correction (%)*** |
| Paragraph | 83.3 | 98 |

## VII. CONCLUSION AND FUTURE SCOPE

In this paper we have developed a Punjabi spell checker with dictionary categorized as clusters. The execution time of spell checking process has been analyzed which is approximately 5 times lesser than spell checker without clustering. The system has given the 83.3% accuracy of detection process and 98% of correction process. Further research can be elaborated to refine the position based prediction of clusters and this technique can be implemented in other languages to develop a spell checker.

## *References*

[1] G.S. Lehal(2007), "design and implementation of Punjabi spell checker", International journal of systemic cybernetics and informatics, pp.70-75.

[2] G. Singh Lehal and Meenu Bhagat, "Spelling Error Pattern Analysis of Punjabi Typed Text", In Proceedings of International Symposum on Machine Translation, NLP and TSS, pp. 128-141, 2007.

[3] R. Kaur and P. Bhatia, "Design and Implementation of SUDHAAR-Punjabi Spell Checker", International Journal of Information and Telecommunication Technology, Vol. 1, Issue 15 May, 2010.

[4] R. Mishra and N. Kaur, "Design and Implementation of Online Punjabi Spell CheckeBased on Dynamic Programming", International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 3, Issue 8, August 2013.

[5] J. Kaur and K. 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.

[6] U. Maheshwar Rao, G., Amba P. Kulkarni, C. Mala and

Parameshwari and K, "TELUGU SPELL-CHECKER", Center for Applied Linguistics and Translation Studies University of Hyderabad Hyderabad, India.

[7] K. Hussein Alkhafaji, "A New Algorithm to Design and Implementation of Multilingual Spellchecker and Corrector", Journal of Al Rafidain University College 32 ISSN (1681 – 6870) Issue No. 32/2013.

[8] R. Cordeiro de Amorim and M. Zampieri, "Effective Spell Checking Methods Using Clustering Algorithms" , Proceedings of Recent Advances in Natural Language Processing, pages 172–178, Hissar, Bulgaria, 7-13 September 2013.

[9] Y. 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.

[10] Sabu M. Thampi,A. Gelbukh and J. Mukhopadhyay," Advances in Signal Processing and Intelligent Recognition Systems", Advances in Intelligent Systems and Computing Volume 264.

All Rights Reserved © 2012 IJSETR