Design and Implementation of Marathi Spell Checker Using Hybrid Approach

Anita A. Patil[1], Priti R. Sharma[2] , Sandip S. Patil[3]

Department of Computer Engineering, SSBT's COET Bambhori, Maharashtra, India[1]

Department of Computer Engineering, SSBT's COET Bambhori, Maharashtra, India[2]

Department of Computer Engineering, SSBT's COET Bambhori, Maharashtra, India[3]

Abstract- Spelling correction is a process of detecting and providing corrections for misspelled words in a text. Marathi is the official language of Maharashtra and Goa states, the fourth most widely spoken language in India. Marathi language is still in its early stage of research and development regarding natural language processing applications in comparison to other languages. Spelling detection and correction for Marathi language is an important task of natural language processing (NLP) which has not get sufficient attention till date. The main challenge while working with such regional languages is first needs to learn its char set and convert it into Unicode's. Thus the existing techniques that are being used to check the errors in English language are not used for Marathi Language. The proposed approach is combination of three techniques, which are dictionary lookup, string similarity function LCS and N- gram. Each approach has a particular purpose and task. Finding the misspelled words from given text is corrected by using Hybrid approach. The proposed mechanism gives accurate error detection and suggestion for correction of Marathi misspelled words.

Keywords- Longest Common Subsequence, Natural Language Processing.

I. Introduction

Spell checking applications are important part of several fundamental applications such as editors and search engines. Language processing is a very complex area of research. So it becomes more and more important to make a Marathi language user friendly as the government is focusing to increase its official use in letters and notifications. Now a day there are several word processing applications are in use, so the solutions for spelling error correction become more important to provide accurate and quality information through text. There are lots of applications and research available for English, Hindi spelling detection and correction but for Marathi very less work has been done. So we are trying to give implementation touch to current work of Marathi language non word and real word error correction by using Hybrid approach.

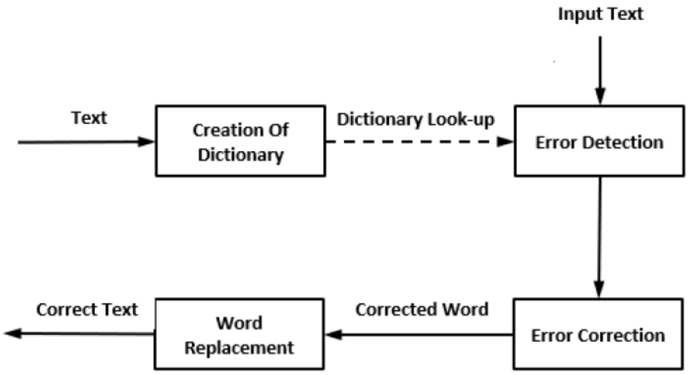

Figure 1.1 Conceptual Model for Spell Checker

Figure 1.1 shows Conceptual Model for Spell Checker which consist of dictionary creation application is a part of spell checker but is not needed to be executed every time because a dictionary that acts as database for spell checker is created using simple word adding approach. At run time dictionary look up, string similarity function and N-gram module goes through it for word traversing and gives corrected text based on dictionary. Error detection use dictionary look up for spelling error detection which checks each word of input text for its presence in dictionary. If that word is there in dictionary, then it is a correct word, otherwise it is put into the list of error words. Error correction in which erroneous words in the text are replaced by intended correct word and finally provide corrected text.

II. OBJECTIVE & SCOPE

To develop a system of Marathi spell checker for non word and real word spelling error detection and correction. Objective is to improve performance of Marathi spell checker system in terms of precision, recall and f-score.

Spell checking is the trending and most vital area to work for different regional languages in India. There are so many languages in India so it is wide scope to work for spell checker and suggestion work.

III. PROPOSED SYSTEM APPROACH

In proposed system, input Marathi text as UNICODE character UTF-8 format, after that text further tokenize into words. The hybrid approach is a combination of three techniques, each having a particular purpose and task.

The task of the first technique dictionary look up is to detect non-word errors using Marathi word dictionary. The task of the second technique string similarity function longest common subsequence is to generate a list of candidate spellings for every detected error in the text using word dictionary. The task of the third technique N-gram is to perform context-sensitive error correction and select the best appropriate spelling candidate word using context dictionary provided by corrected text.

Figure 1.2 shows the Block Diagram for proposed Hybrid approach, input Marathi text is in Devanagari script convert into as UNICODE character UTF-8 format. The input Marathi text as UNICODE character and that containing misspelled word.

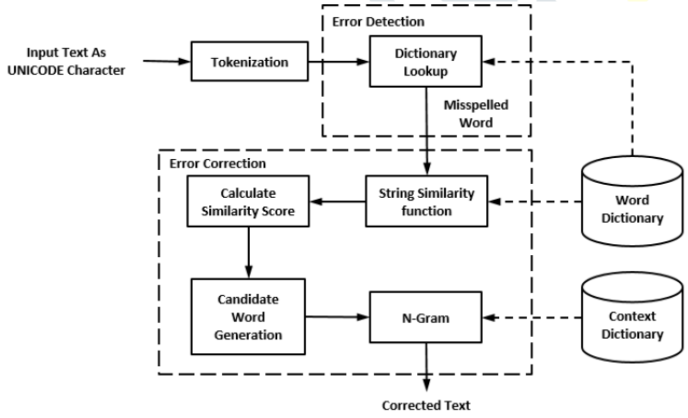

Fig.1.2 Block Diagram of Proposed System

The aim of this phase is understand Devanagari script to programming language. A spell checker is composed of mainly three components, tokenization, error detection and candidate word generation for error correction. Bunch of techniques that uses the database for spelling error detecting and providing spelling suggestions for misspelled words correction. These techniques generally provide three types of functionalities. At start user gives the input and the system detect the misspelled word by looking up for that particular text into the Marathi word dictionary and provide the suggestion to correct misspelled words. The final output is a corrected text without any spelling mistakes.

Tokenization: Tokenization is the first component of proposed approach. It is splitting the text into sentences because two misspelled words in different sentences have no relationship with each other. The tokenization is required to separating the words from the sentences and produces the tokens. The tokens act as a particular word which is extracted from the text.

Error Detection: Error detection is the second component of proposed approach. It is responsible for checking whether the input word is misspelled or not. The error detection component works first is error detection by using dictionary lookup finding approximate matches in word dictionary. If the input word exists in the word dictionary, the spell checker does nothing. It detects the errors in the input text when input word not exist in dictionary.

Algorithm 1: Dictionary Lookup Error Detection

```
Procedure Check Misspelled Word
Input: Marathi text
Output: List of misspelled word
Function ERROR DETECTION(word, dictionary) returns misspelled word
Enter text as UNICODE character ( S€si )
Split the text on every space and store the word into an array W
Searches for every w[i] in Marathi dictionary
Flag ← binary search(Marathi dictionary, w[i])
If flag = true then
    i ← i+1 //move to the next word w[i]
    w[i] ← found in dictionary
Else
    w[i] ← not found in dictionary //word add into misspelled word list
End if
End procedure
```

Algorithm 1, the proposed error detection algorithm detects misspelled word w is a word in the original text and n is the total number of words in the text. The process starts by validating every word w[i] in Marathi word dictionary. If w[i] is found, then w[i] is the correct. Otherwise if the word w[i] is not found, then w[i] is the misspelled hence a correction is required. Binary search be employed to speed up the execution time of error detection. Ultimately a list of misspelled word is generated and is denoted by E = w1, w2,w3, wm where m is the total number of word errors detected in the original text.

Error Correction: Error correction is the third component of proposed approach. When an input word is detected as a misspelled in written text then spelling correction techniques are applied on misspelled word to correct the word or providing correct suggestions for that word this process is called spell correction. Error correction which consist of main two phase's first candidate word generation and second choose the most likely candidate to fix then provide corrected text.

Algorithm 2: Context Sensitive Error Correction

```
Procedure Corrected Text
Input: Misspelled words
Output: Corrected text
Get the misspelled words
Apply Longest common subsequence edit distance
String similarity score between two word
Score ← Substring Match (dictionary word, misspelled word)
Candidates ← MAX (Get common words (results))
If candidate word C = w1,w2,......etc then
    Apply N-gram for choose corrected candidate word
    Search with context of word for misspelled word
```

```
Current misspelled word wi and its contexts are $w^{i-1}$, $w^{i-2}$, $w^{i+1}$ and $wi^{+2}$
Count =P1 $(w^{i} \mid w^{i-1})$ + P2 $(w^{i} \mid w^{i+1})$ + P3 $(w^{i} \mid w^{i+1},w^{i-1})$
S = $w^{i-2}$ $w^{i-1}$ $w^{i}$ $w^{i+1}$ $w^{i+2}$ .etc.
Searches for S context words and returns its frequency
Count[i] ←Binary search (context dictionary, S)
Index ← MAX(count)
RETURN candidates[index]
Else
State Return corrected text
End If
End Procedure
```

Algorithm 2, the proposed error correction algorithm have two phases first generation of candidate word corrections are denoted by C=w1, w2,....,etc. where w1, w2 denotes a particular candidate spelling. Second phase is the proposed context-sensitive spelling error correction algorithm takes each generated candidate c^{ik} with context of word left and right side words of misspelled word in the original text, leading to S = w^{i-2} w^{i-1} wi w^{i+1} w^{i+2} ...,etc. where S denotes sentence, w(i) denotes the original error word. The candidate word that belongs to the sentence S with the highest count is selected as a replacement for originally detected error word.

Candidate word generation: Candidate words generation get misspelled word from error detection component send to string similarity function. Edit distance length of the longest common subsequence are special cases of character distance and similarity respectively. Use the longest common subsequence (LCS) measure with some normalization and small modifications for the string similarity measure.

In classical LCS, the common subsequence needs not be consecutive in spelling correction. A consecutive common subsequence is important for a high degree of matching. Used maximal consecutive longest common subsequence MCLCS matching at first MCLCSf middle character MCLCSm and last character MCLCSl takes two strings as input and returns the shorter string or maximal consecutive portions of the shorter string that consecutively match with the longer string, where matching must be from first, middle and last character for both strings.

N-gram Model: Context-sensitive error correction algorithm, present the context-based method to check spelling with large scale of N-gram model on Marathi text. N-gram model work, the context-sensitive spelling corrections almost take the context in left side but in this proposed approach take the context at both sides of the misspelled word to improve the systems performance. Using the context in both sides get more clues to choose the best candidate in confusion set.

The context used in proposed system is the both side context words of misspelled word surrounding. The probabilistic classifier which is used to build the language model for spelling correction, in order to find best appropriate suggestions to correct spelling errors. To find the best appropriate candidate string candidate word C for the current misspelled word w^{i} and its context are w^{i+1}, w^{i+2}, w^{i-1} and w^{i-2} etc is a context dictionary words. To decide which best appropriate candidate of a misspelled word in given context dictionary. Provide text by replacing all incorrect sentences with correct sentences and reconstruct final output corrected text.

IV. RESULTS AND DISCUSSION

F-Score: F score is used to measure of a test's accuracy. It considers both the precision and recall of the test to compute the score. The F-score is weighted average of the precision and recall, where F-score reaches its worst value at 0 and best value 1. The F-measure or balanced F-score is the harmonic mean of precision and recall.

$$\textbf{F-Score} = \frac{2}{((1/Recall) + (1/Precision))}$$

Table 1.1 shows the F-score values of Hybrid approach and Edit Distance Approach. The F-Measure calculation for Marathi spell corrections considered the values of precision and their respective recall values. By applying the formula given in above Equation, calculates the F-Measure values. Figure 1.3 is plotted by considering the F-score.

No of words	No of sentences	F-score of Hybrid approach	F-score of Edit Distance
25	5	0.72	0.63
32	7	0.66	0.59
50	11	0.70	0.56
64	15	0.68	0.55

Table 1.1 F-Score values of Hybrid approach and Edit distance

The proposed experiment of Hybrid approach for Marathi spell checker is successfully satisfied problem statement. According to models of Marathi spell checker is successfully implemented dictionary creation, dictionary look-up, longest common subsequence and N-gram. The performance of given system is measured by factors such as precision, recall and F-score. The experimental result performance of the system is higher compared based on precision, recall and f-score of the proposed system.

F-measure values are changes according to the precision and recall values change. Average precision values for Hybrid approach is 0.78 and Edit Distance is 0.64 respectively. Average Recall values for Hybrid approach is 0.75 and Edit Distance approach is 0.61.

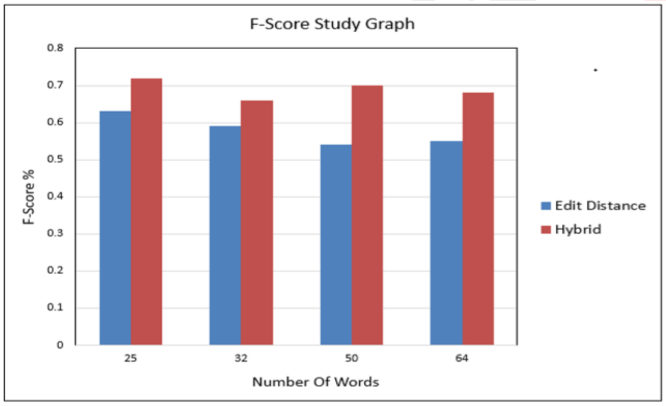

Fig. 1.3 F-Score study graph

The F-score values are calculated from the precision and recall measures. The F-score values are calculated from the precision and recall measures. Average F-score value for of the Hybrid approach is 0.70 and Edit Distance is 0.58.

Results of experiment states that the the precision values of hybrid approach are higher and get better accuracy in terms of the correction of text. The proposed system gives exactness in Marathi spell checker system using hybrid approach than edit distance. Because in edit distance approach longest common subsequence string similarity function use same character matching between misspelled word and dictionary word provide sometime wrong candidate word suggestion for correction misspelled word.

CONCLUTION

The work done for Marathi spell checker using hybrid approach achieved high accuracy as compare to edit distance technique. Edit Distance technique failed to achieve high accuracy while working with misspelled word which gives list of suggestions for correction. To overcome all existing drawbacks, proposed hybrid approach with a combination of dictionary lookup, string similarity LCS and N-gram into a single algorithm. The proposed work has been implemented strong framework of Marathi text provide accurate suggestion for correction using hybrid approach. The proposed mechanism achieves high accuracy towards error detection, suggestion for correction of misspelled words.

FUTURE WORK

In future, the scalability of the Marathi spell checker get extends for the corrections of syntax and semantics in the Marathi sentence and it can be used in Marathi translation system.

REFERENCES

[1] Islam, Aminul and Diana Inkpen, **`Real-word spelling correction using Google Web 1T n-gram with backoff_'** , In 2009 International Conference on Natural Language Processing and Knowledge Engineering, pp. 1-8, IEEE, 2009.

[2] Etoori, Pravallika, Manoj Chinnakotla and Radhika Mamidi, **`Automatic Spelling Correction for Resource-Scarce Languages using Deep Learning**' , In Proceedings of ACL 2018, Student Research Workshop, pp. 146-152, 2018.

[3] Kumar, Rakesh, Minu Bala and Kumar Sourabh, ‘**A study of spell checking techniques for Indian Languages’** , JK Research Journal in Mathematics and Computer Sciences 1, no. 1, 2018.

[4] Bhaire, Vibhakti V., Ashiki A. Jadhav, Pradnya A. Pashte and P. G. Magdum, ‘**Spell checker’** , International Journal of Scientific and Analisys Publication 5, Issue 4, pp. 1-5, 2015.

[5] Bassil, Youssef and Mohammad Alwani, ‘**Context-sensitive spelling correction using google web 1t 5-gram information’** , Canadian Center of Science and Education 5, No. 3, pp. 37-48, 2012.

[6] Huong, Nguyen Thi Xuan, Tran-Thai Dang and Anh-Cuong Le, ‘**Using large n-gram for Vietnamese spell checking**’ , In Knowledge and Systems Engineering, pp. 617-627, Springer, Cham, 2015.

[7] Faili and Heshaam, ‘**Detection and correction of real-word spelling errors in Persian language’** . In Proceedings of the 6th International Conference on Natural Language Processing and Knowledge Engineering (NLPKE-2010), pp. 1-4, IEEE, 2010.

[8] Wasala, Asanka, Ruvan Weerasinghe, Randil Pushpananda, Chamila Liyanage and Eranga Jayalatharachchi, ‘**A data-driven approach to checking and correcting spelling errors in sinhala’** , The International Journal on Advances in ICT for Emerging Regions 3, no. 1, pp. 11-24, 2010.

[9] Shailza Kanwar, Manoj Sachan and Gurpreet Singh,August, **‘N-gram Solution for Error Detection and Correction in Hindi Language’** , Journal of Advanced Research in Computer Science 8, No. 7, pp. 667-670, 2017.

[10] Dixit, Veena, Satish Dethe and Rushikesh K. Joshi, **‘Design and implementation of a morphology-based spellchecker for Marathi and Indian language’** , ARCHIVES OF CONTROL SCIENCE 15, pp. 301, no. 3, pp. 251258, 2005.

[11] Das, Monisha, Samir Borgohain, Juli Gogoi and Shivashankar B. Nair, **'Design and implementation of a spell checker for Assamese'** , In Language Engineering Conference, pp. 156-162. IEEE, 2002.

[12] Fahda, Asanilta and Ayu Purwarianti, **'A statistical and rule-based spelling and grammar checker for Indonesian text'** , International Conference on Data and Software Engineering (ICoDSE), pp. 1-6. IEEE, 2017.

[13] Gaddisa Olani Ganfure and Dr. Dida Midekso, **'Design and implementation of morphology based spell checker'** , International Journal Scietific and Technology Research 3, no. 12, pp. 118-125, December 2014.

[14] Hodge, Victoria J. and Jim Austin**, 'A comparison of standard spell checking algorithms and a novel binary neural approach'** , IEEE transactions on knowledge and data engineering 15, no. 5, pp. 1073-1081, 2003.

[15] UzZaman, Naushad and Mumit Khan. **'A double metaphone encoding for Bangla and its application in spelling checker**' , In 2005 International Conference on Natural Language Processing and Knowledge Engineering, pp. 705-710, IEEE, 2005.

[16] Chitra L. Mahajan and Sandip S. Patil, **'Word Sense Disambiguation For Devnagri Script'** ,is published in International Journal of Creative Research Thoughts (IJCRT), Volume 5, Issue 12,Dec. 2017.

[17] Darshana Bhole and Sandip S. Patil, **'The Study and Review of Paraphrase Detection Techniques in Machine Learning'**, is published in International Journal of Innovative Research and Science, Engineering and Technology(IJIRSET), vol:6, Special Issue:1, Jan. 2017

AUTHORS

First Author – Miss. Anita A. Patil, ME-Comp, Department of Computer Engineering, SSBT's COET Bambhori, Maharashtra, India. patilanita2512@gmail.com

Correspondence Author Prof. Miss. Priti R. Sharma, ME-Comp, Department of Computer Engineering, SSBT's COET Bambhori, Maharashtra, India. pritirsharma@gmail.com

Correspondence Author Prof. Mr. Sandip S. Patil, ME-Comp, Department of Computer Engineering, SSBT's COET Bambhori, Maharashtra, India. sspatiljalgaon@gmail.com