Design and Implementation of HINSPELL -Hindi Spell Checker using Hybrid approach

Baljeet kaur[1]***, Harsharndeep Singh***[2]

[1]Department of Computer Science & Engineering,
Baba Farid College of Engineering and Technology, Bathinda, India
baljeetmtech10@gmail.com

[2]Department of Information Technology
Baba Farid College of Engineering and Technology, Bathinda, India
harsharndeepsinghsivia@gmail.com

Abstract: A spell checker is an application program that flags words in a document that may not be spelled correctly. A spell checker is a basic need of a word processor of any language. Spell checker analyzes the written text in order to identify any misspellings and gives best correct suggestions for those misspellings. Most of work has been done in English and Punjabi language. Hindi is the third most spoken language in the world. In This paper the design, techniques and implementation of the Hindi spell checker is proposed. Error detection, Error correction by generating suggestions and replacement are the main features of this system. The system detects approximately 83.2% of the errors and provides 77.9% of the correct suggestions for the misspelled words.

Keywords: *Error detection, Error correction, HINSPELL, dictionary lookup, weight age algorithm, M.E.D, SMT.*

1. Introduction

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 (OCR). 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.

1.1 Types of Errors:

Techniques of error detection and correction were designed on the basis of type of spelling errors. According to various studies, spelling error can belong to two distinct categories: Non-word error and Real-word error [3].

Non-word errors are those error words that cannot be found in the dictionary. E.g. ग्यान for ज्ञान.

Typographic errors [14] categorized under non-word errors which occur when the correct spelling of the word is known but the word is mistyped by mistake. These errors are mostly related to the wrong key press. For example, typing आपमान for अपमान. Real-word errors are those error words that are acceptable words in the dictionary but not correct according to sentence. For example, मेरा घर उस और है (incorrect) for मेरा घर उस ओर है (correct) और is an acceptable word in the Hindi dictionary but it occurs as an error for ओर word. Possibility of spelling mistakes in Hindi language increases because Hindi is a highly confusing language. Hence Hindi spell checker is the solution for making input text correct.

2. Proposed Work

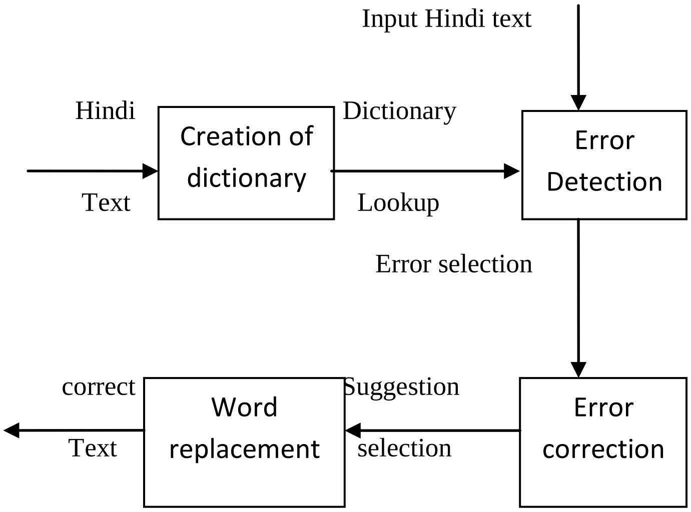
A few work is done in Hindi spell detection and correction field and it is not an easy task to identify errors in Hindi text. The spell checker systems are online available but as not standalone applications. Some paid Hindi spell checker software's are also online available.

HINSPELL is a web based spell checking and correcting application for Hindi language. HINSPELL only deals with non-word errors. The main features of HINSPELL are large correct database and user interactive.

2.1 implementation of HINSPELL

Two different applications are designed in HINSPELL. One is dictionary creation tool, executed once to create the own dictionary and second is a spell checker for Hindi language and it is implemented in c# language. At start, user gives the input Hindi text and the system detect the errors by looking up for that particular word into the created Hindi dictionary and provides the correct suggestions for that misspelled word in the suggestion list. After that user can select the suggestion from the suggestion list and replace errors accordingly. The final output is a corrected text without any spelling mistakes.

Figure 2.1: Architecture of HINSPELL[4]



2.1.1 Error Detection

The error detection process consists of detecting any spelling errors in the input text. In HINSPELL, dictionary lookup technique is applied for detecting errors in input text by checking each word of input text for its presence in to the created Hindi dictionary. If the word is found then it is a correct word otherwise it considers as an error word and that word will be added into Error word list.

2.1.2 Error correction

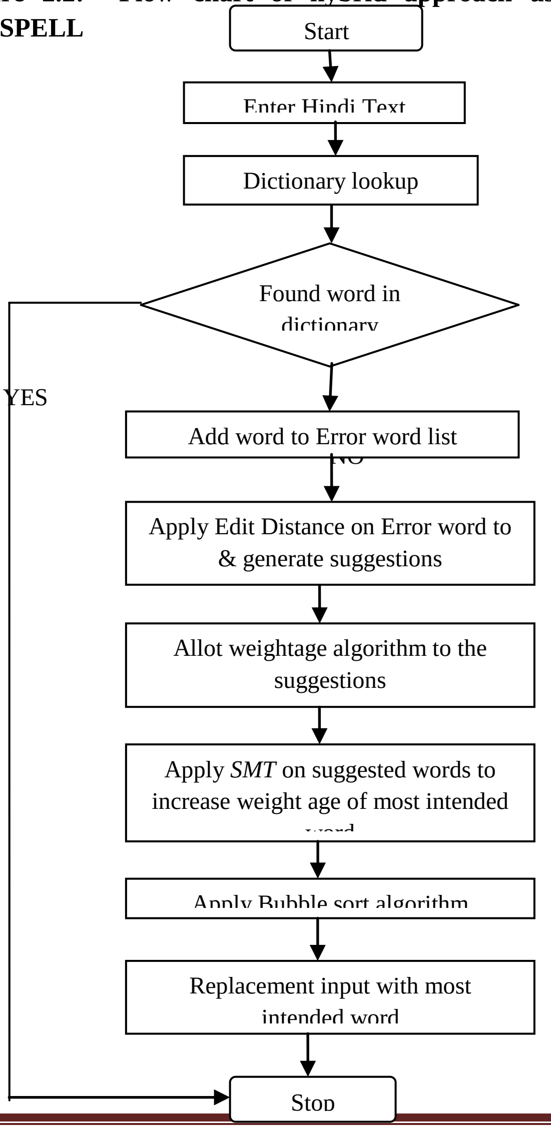
Error correction consists of two steps: the generation of correct possible suggestions for the error word and the ranking of suggestions [3]. Weightage algorithm, minimum edit distance and statistical machine translation techniques are used for error correction in HINSPELL. *Minimum edit distance (M.E.D)* applied on error word to generate possible suggestions for that word. In the process of basic editing operations i.e. Insertion, deletion and Substitution, *M.E.D* changes an error word into the possible correct word. Distance between error word and dictionary words are measured. The dictionary word having minimum distance with error word is ranked higher in suggestion list. Table 2.1 shows the possible suggestions and minimum edit distance (M.E.D) of some error words.

Table 2.1: Minimum Edit Distance (*M.E.D*)

Error word	**Possible suggestions**	**Operation & performed M.E.D**
आपमान	1.अपमान 2.आसमान	Deletion (ा) (*M.E.D*=1) substitution स (*M.E.D*=1)
अदमी	1.आदमी 2.अदली	Insertion (ा) (M.E.D=1) Substitution ल (M.E.D=1)
भुल	1.भूल 2.भील	Substitute (M.E.D=1) Deletion ु , insert ी (M.E.D=2)

Through weightage algorithm, weights are allocated to generated suggestions. Statistical machine translation (*SMT*) technique applied to give priority to suggestions with same minimum edit distance. SMT is applied on suggestions to find a most intended word from the list of suggestions. Minimum 3 words as an input are required for proper working of this technique. In HINSPELL, SMT compare the input text with paragraphs maintained into database for choosing the most intended suggestion. Priority is assigned by replacing suggestions with error word according to its previous and next word. If their exact combination is found in the database paragraphs then that suggestion is suggested as most suitable word. For example, महाराजा रणजीत पंजाव के राजा थे. In this sentence पंजाव word is an error word. According to user it may be possible that correct word will पंजाब or पंजा. SMT will give priority to suggestion by making word combinations like [रणजीत पंजाब के] and [रणजीत पंजा के]. The word combination which will be found into database that suggestion will be most intended suggestion. Most intended suggestion will be arranged on the top of the suggestion list by applying Bubble sort algorithm.

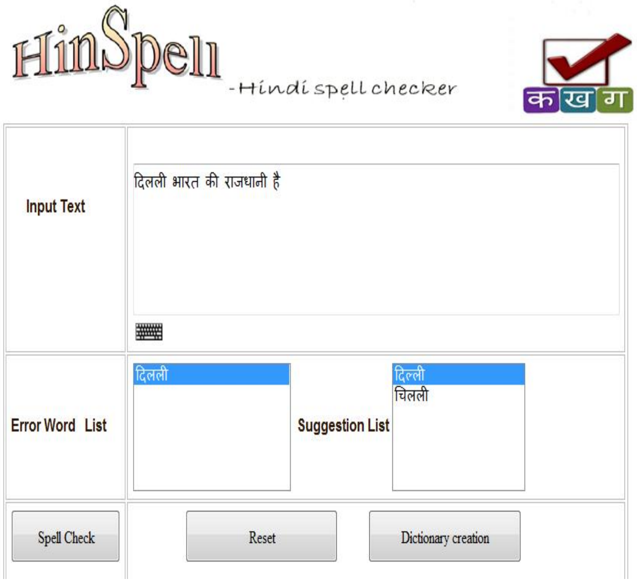
Figure 2.2: Flow chart of hybrid approach used in HINSPELL


Start
Enter Hindi Text
Dictionary lookup
Found word in dictionary
YES
Add word to Error word list
NO
Apply Edit Distance on Error word to & generate suggestions
Allot weightage algorithm to the suggestions
Apply SMT on suggested words to increase weight age of most intended word
Apply Bubble sort algorithm
Replacement input with most intended word
Stop

2.2 Outlook of HINSPELL

Figure 2.3: shows the user interface of HINSPELL. User gives the input and click on spell check button, error words will be shown into error word list. When the User will select the error word from the error word list; possible suggestions will be shown into suggestion list.

If the most intended suggestion found in the suggestion list then the user will replace the error word by selecting most intended word from the suggestion list. Similarly users can perform tasks like Reset; dictionary creation etc. virtual Hindi keyboard can be used for typing the text. Hindi text can be typed in any format like Devanagari, Dogri script etc.



दिल्ली भारत की राजधानी है

दिल्ली

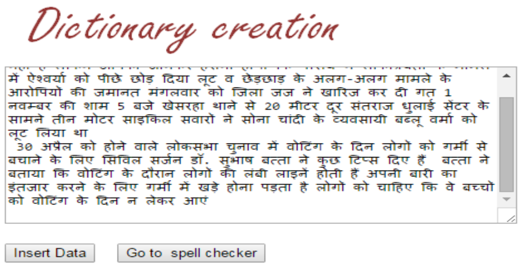
दिल्ली
चिल्ली

Figure 2.3 User Interface of HINSPELL

2.2 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 HINSPELL. As Shown in figure 2.4 by clicking on insert data button, words will be added into database of the spell checker.

Figure 2.4: Dictionary creation tool



में ऐश्वर्या को पीछे छोड़ दिया लूट व छेड़छाड़ के अलग-अलग मामले के आरोपियों की जमानत मंगलवार को जिला जज ने खारिज कर दी गत 1 नवम्बर की शाम 5 बजे खेसरहा थाने से 20 मीटर दूर संतराज धुलाई सेंटर के सामने तीन मोटर साइकिल सवारों ने सोना चांदी के व्यवसायी बब्लू वर्मा को लूट लिया था

30 अप्रैल को होने वाले लोकसभा चुनाव में वोटिंग के दिन लोगों को गर्मी से बचाने के लिए सिविल सर्जन डॉ. सुभाष बत्ता ने कुछ टिप्स दिए हैं बत्ता ने बताया कि वोटिंग के दौरान लोगों की लंबी लाइनें होती हैं अपनी बारी का इंतजार करने के लिए गर्मी में खड़े होना पड़ता है लोगों को चाहिए कि वे बच्चों को वोटिंग के दिन न लेकर आएं

.

.

3. Results Analysis

In this research, 870 misspelled words randomly collected from books, newspapers and peoples etc as input to test the system. In the result analysis there are 724 words detected as error words and system generates correct suggestions for 678 words. Hence detection rate of the system reaches 83.2% approximately and correction rate of the system reaches 77.9% approximately. Accuracy depends on the length of the characters and no. of editing operations required to change an error word into correct word.

Table 3.1. Results of HINSPELL

Dataset &character length	Total misspelled words	Detected as an error word	Intended word in suggestion list
D1(L=2)	227	181	150
D2(L=3)	260	212	206
D3(L=4)	223	189	183
D4(L=5)	160	142	139

Here D and L denote dataset and length of character simultaneously.

4. Conclusion and Future Scope

This paper presents the HINSPELL-Hindi spell checker system which is not a part of any word processor or website. This system only deals with non word errors. Real word errors are subject of future research. The system gives the approximately 83.2% detection rate and 77.9% Correction rate. After applying SMT Technique, the accuracy of the system increases but response time of the system also increases so there is a scope of improvement in implementation of SMT with less response time. HINSPELL can also be used for other languages with modification of dictionary and keyboard.

References

[1] Shikha kabra, Ritika Agarwal, February (2014) "Auto spell suggestion for high quality speech synthesis in Hindi", international journal of computer applications, volume87-no17.

[2] 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.

[3] Neha Gupta, Pratistha Mathur, December (2012), "Spell Checking Techniques in NLP": A Survey, International Journal of Advanced Research in Computer Science and Software Engineering Volume 2, Issue 12.

[4] 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.

[5] Mrs. Namrata Tapaswi, April (2012), Dr. Suresh Jain, Mrs. Vaishali chourey, "morphological-based spell checker for Sanskrit sentences", international journal of scientific & technology research volume1, issue 3,

[6] Amit Sharma & Pulkit Jain, April (2013) "Hindi Spell Checker", Indian Institute of Technology Kanpur".

[7] Li Zhao, (2009) "Based on the Phonetic Spelling Correction System Research and Implementation", Xi'an Technological University, Xi'an, china, IEEE.

[8] S. Dasgupta, C.H. Papadimitriou and U.V. Vazirani, "Algorithms", p.173, available at http:/ / www.cs.berkeley.edu/ vazirani/ algorithms.html

[9] Gurpreet Singh Lehal, (2007) "Design and Implementation of Punjabi Spell Checker", International Journal of Systemic, Cybemetics and Informatics, pp.70-75.

[10]G S Lehal & Meenu Bhagat, "Spelling Error Pattern Analysis of Punjabi Typed Text", In Proceedings of the 2007 International Symposium on Machine Translation, NLP and TSS, pp. 128-141.

[11]Francesco Bonchi, Ophir Frieder, Franco Maria Nardini, Fabrizio Silvestri and Hossein Vahabi, (2012) "Interactive and Context-Aware Tag Spell Check and Correction"

[12]Suzan Verberne, (2002) "Context-sensitive spell checking based on word trigram probabilities".

[13]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.

[14]F.J. Damerau, (1964), "A Technique for Error Detection and Correction of Spelling Errors", Communication ACM, pp. 171-176.

[15]Monisha Das, S. Borgohain, JuliGogoi, S. B. Nair, "Design and Implementation of a Spell Checker for Assamese", in proceedings of the (2002) Language Engineering Conference (LEC'02), pp. 156.

[16]R.E. Gorin, (1971) "SPELL: A spelling checking and correction program", Online documentation for the DEC-10 computer.

[17]Ritu aggrawal, September (2007), "Hindi editor with spell checker", Vinayaka Mission University, Salem.

[18]Peterson James (1980) "Computer Programs for Detecting and Correcting Spelling Errors", Computing Practices Communications of the ACM.

[19]Tanveer Siddiqui, U.S.Tiwary (2008), "Natural Language Processing and Information Retrieval" Oxford university press.

[20]Prof.Puspak Bhattcharya and Prof. Rushikersh Josh, "Design and implementation of morphology based spellcheckers for Marathi", TDIL Newsletter.

[21]Mukand Roy, Gaur Mohan, Karunesh K Arora, "Comparative study of spell checker algorithm for building a generic spell checkers For Indian language C-DAC NODIA ,India.

[22]Veena Dixit, Satish Dethe, Rushikesh K. Joshi, "Design and Implementation of a Morphology-based Spellchecker for Marathi, an Indian Language", Indian Institute of Technology Bombay, India.

[23]K. Kukich (1992) "Techniques for automatically correcting words in text", ACM Computing Surveys. 24(4): 377-439.

[24]Robert & Cherry, Lorinda L, March (1975), "Computer Detection of typographic errors", IEEE Trans Professional Communications, vol. PC-18, no.1 pp 54-64.

[25]Ajit Kumar, vishal Goyal, "Tdil programme: a government initiative", Department of Computer Science, Punjabi University, Patiala.

[26]Deepak Seth, Mieczyslaw M. Kokar2, "SSCS: A Smart Spell Checker System Implementation Using Adaptive Software Architecture", Northeastern University, Boston, MA 02115, USA.

[27]Hindi spell checker available at https://addons.mozilla.org/en-US/firefox/addon/hindi-spell-checker/

[28]Hinkhoj spell checker available at http://dict.hinkhoj.com/spell-checker/check-spelling.php

[29] Spell guru available at http://bhashagiri.com/

Author Profile

Baljeet Kaur is a Student of M.Tech (computer science Engg.) at Baba Farid College of engineering and technology, Bathinda. She has received her B.Tech in Computer Sciences from Baba Farid College of engineering and technology, Bathinda in 2012. She is persuing her M.Tech Thesis in the area of Natural Language Processing.

Harsharndeep Singh received M.Tech degrees in Information Technology from Maharishi Markandeshwar University, Mullana, Ambala in 2012. He is working as Assistant Professor in Department of Information Technology at Baba Farid College of Engineering and Technology, Bathinda, India.