A Survey on Creation of Hindi-Spell Checker to Improve the Processing of OCR

Shabana Pathan[1], Nikhil Khuje[2], Punam Kolhe[3]

[1]*Assistant Professor, Department of Information Technology, SVPCET, Nagpur, India*

[2,3]*Student, Department of Information Technology, SVPCET, Nagpur, India*

***Abstract*: This project investigates the principles of Optical Character Recognition used in the Tesseract OCR engine and techniques to improve its efficiency and processing. This mainly focuses on improving the Tesseract OCR efficiency for Hindi/Devanagari languages. Improving Hindi/Devanagari text extraction will increase Tesseract's performance and in turn will draw developers to contribute towards Hindi /Devanagari OCR. The project introduces the efficiency of new Tesseract OCR version 4.0 beta and heads us towards its capabilities which reflects in its output. 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. Thus the project mainly focuses on creation of Hindi-spell checker which in turn will provide clean final output and also will help to improve the processing of Tesseract OCR.**

***Keywords*: Hindi-Spell Checker, OCR**

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.

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. Objective

Our main objective is to convert image data into text data and obtain final noise free correct data.

We are about to design a model for user that uses the given tool for checking hindi spelling. To help the users to know the exact spelling of the words by designing the tool that gives it. This will help in another project related to NLP.

3. Literature survey

This documentation describes the OCR platform, one of the Open Source OCR tool Tesseract systems which originated within the project “Creation of Hindi-Spellchecker to improve the processing of Tesseract OCR” which includes conversion of printed text into editable text. It contains the detail specification of Accuracy of OCR, its methods, History of Open Source OCR tool Tesseract, which include the details as, it was modified and improved in 1995 with greater accuracy. In late 2005, HP released Tesseract for open source. The newer Tesseract versions are released and available, which is developed and maintained by Google, providing support for various languages. This documentation is helpful for understanding Accuracy of OCR, that is dependent on text preprocessing and segmentation algorithms. This documentation helps in understanding architecture of tesseract OCR and its working. It contains detailed information about the phases in the architecture of Tessseract OCR and experiment result of OCR performed by Tesseract on different kinds images in description [1].

The automatic spell checker system of a text is a tool of words verification and correction included in the text. This correction concerns many verification and correction axes namely:

- Spelling verification and correction: consists of verifying the spelling of a word. If the word is misspelled, the system suggests a set of words that are the nearest lexically to the erroneous one.
- Syntactic verification and correction: consists of verifying the syntactic accordance. A word with a syntactic error is spelled correctly but has a wrong structuring.
- Semantic verification and correction: consists of verifying the meanings of the text's words. A words semantically wrong is correct syntactically and correctly spelled but incorrect semantically in its context. [2].

The task of spell checking is primarily divided into two parts:

1) Error Detection
2) Error Correction

The first part consists of identifying the errors in the typed text. This part uses a language model which accounts for the words allowed in the language. Language models may vary from a simple list of permitted words to finite state graphs that accept words with valid spellings in the language. The second part consist of rectifying the spelling mistakes made by the user. This requires an error model which tries to find out the candidate replacements of a mis-spelled word. This part also includes ranking of the candidate replacements. Ranking may be done on the basis of edit distances, string similarity measures, phonetic measures or word frequency [3].

A framework for assisting human while correcting the OCR errors in documents, mostly dedicated to Indian Languages. Tested on Sanskrit, Hindi, Marathi and English.

The interactive features as of now are:

1. Error detection.
2. Generating Suggestions (will replace this to auto-completion in Future Work).
3. Following information is updated on the fly, after correction of each page:
   i. A domain specific dictionary which is uploaded on the fly.
   ii. The domain specific dictionary is also uploaded with OCR words with high confidence as the user starts working on the document.
   iii. OCR confusions specific to the documents which are uploaded on the fly [4].

GNU Aspell is a Free and Open Source spell checker designed to eventually replace Ispell. It can either be used as a library or as an independent spell checker. Its main feature is that it does a superior job of suggesting possible replacements for a misspelled word than just about any other spell checker out there for the English language. Unlike Ispell, Aspell can also easily check documents in UTF-8 without having to use a special dictionary. Aspell will also do its best to respect the current locale setting. Other advantages over Ispell include support for using multiple dictionaries at once and intelligently handling personal dictionaries when more than one Aspell process is open at once [5].

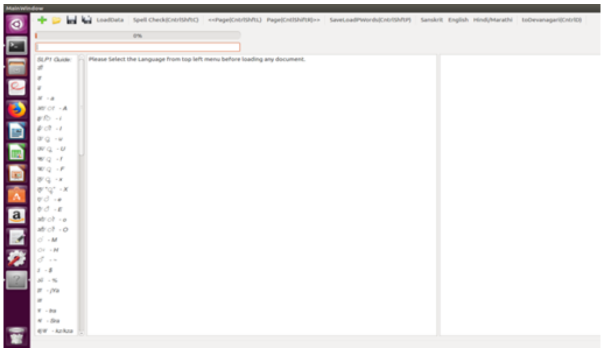
4. Proposed system

A few work has been done in Hindi spell detection and correction field by hunspell spell checker. Hunspell spell checker only checks the substring of words so we used IIT-bombay tool which is mainly designed for sanskrit language.

The proposed system displays the text associated with it's image at the same time on the screen. After the text is displayed the tool searches for error and provide suggestions for wrong words. For this we build dictionary of hindi words. It checks OCR output with Google doc output for finding wrong words and correct them by giving suggestions.

Aspell will suggest the possible replacement for your misspelled words in a word/document. Unlike other spell checkers like Ispell, Aspell can also easily check documents in UTF-8 without having to use a special dictionary.

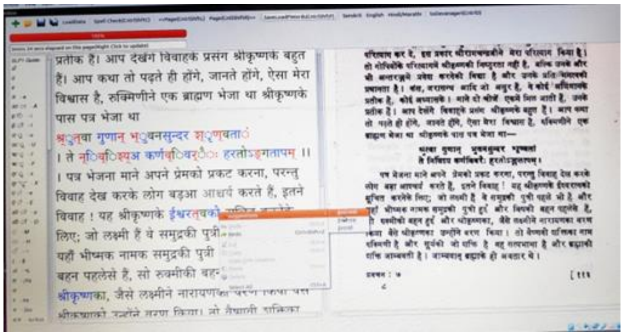
5. Implementation



Please Select the Language from top left menu before loading any document.

Fig. 1. System Design

We use IIT-Bombay tool for checking hindi spelling. Basically IIT-bombay tool is designed for sanskrit language so we modified it for hindi language. This tool requires two different OCR's converted data, from which you can make use of google OCR (i.e GEROCR) as it gives more better result than others, and second one can be tesseract OCR(i.e IEROCR). The proposed system displays the text associated with it's image at the same time on the screen which is stored in Inds folder. when text is displayed on the screen, it is appeares in red color. After clicking on load data button, it loads dictionary words, OCR words, Indsenz words, google doc words, PWords, domain dictionary. then click on spell checker, it shows wrong words in different color. conjoined word is formed by combination of a minimum number of words from the word dictionary. The problem in reading the conjoined words, due to their large length, is also an important factor while curating the OCR errors. To overcome this, we provide a user-friendly color coding scheme in our framework for the partial dictionary string matches, for each combined word. The words verified as correct are marked as black by our framework. The gray words are the words that have been marked as correct by the user (previously at a different location in the document) and the purple words are ones that have been auto-corrected by the system. The user is required to right click to generate suggestions. Each multi-colored (green and blue) word is a conjoined word consisting of substrings which are valid words in either the word dictionary or the domain vocabulary (which is updated on the fly with corrections). The colors (green and blue) differentiate the adjacent valid substrings of the conjoined word. After the text is displayed the tool searches for error and provide suggestions for wrong words. after correcting wrong word click on saveLoadPWord which automatically store the correct word document to the correct folder.



प्रतीक है। आप देखंग विवाहक प्रसंग श्रीकृष्णक बहुत
हैं। आप कथा तो पढ़ते ही होंगे, जानते होंगे, ऐसा मेरा
विश्वास है, रुक्मिणीने एक ब्राह्मण भेजा था श्रीकृष्णके
पास पत्र भेजा था
श्रुतवा गुणान् भुवनसुन्दर शृणवतां
। ते नृविश्यअ कर्णविवरः हरतोऽङ्गतापम् ॥
। पत्र भेजना माने अपने प्रेमको प्रकट करना, परन्तु
विवाह देख करके लोग बड़आ आश्चर्य करते हैं, इतने
विवाह ! यह श्रीकृष्णके ईश्वरतवको
लिए; जो लक्ष्मी हैं ये समुद्रकी पुत्री
यहाँ भीष्मक नामक समुद्रकी पुत्री
बहन पहलेसे है, सो रुक्मीकी बहन
श्रीकृष्णका, जैसे लक्ष्मीने नारायणका

Fig. 2. Examples of incorrect OCR words with improved readability

After all incorrect words are replaced by correct words, click on SaveLoadPwords which automatically save the file in correct folder.

Aspell is a utility program that connects to the Aspell library so that it can function as an ispell -a replacement, as an independent spell checker, as a test utility to test out Aspell library features, and as a utility for managing dictionaries used by the library. The Aspell library contains an interface allowing other programs direct access to it's functions and therefore reducing the complex task of spell checking to simple library calls. The default library does not contain dictionary word lists. To add language dictionaries, please check your distro first for modified dictionaries, otherwise look here for base language dictionaries.

6. Conclusion

In this paper we demonstrate different approaches for hindi OCR. Error confusions and domain specific vocabularies grow on-the-fly with user corrections. Our framework leverages generic word dictionaries and a domain-specific vocabulary grown incrementally based on user corrections from the current on the OCR document. It also learns OCR specific confusions on-the-fly. We have incorporated word conjoining rules to parse OCR words and discover their potentially correct sub-strings. Furthermore, we have presented a dual engine environment to cross-verify potential errors and corrections.

References

[1] Chirag Indravadan bhai Patel, "Optical Character Recognition by Open source OCR Tool Tesseract: A Case Study" in International Journal of Computer Applications 55(10):50-56 October 2012.

[2] Mohammed V, Rabat- Morocco B Eradiass Team, “The International Conference on Advanced Wireless, Information, and Communication Technologies (AWICT 2015)” The context in automatic spell correction, ENSIAS, University, FSJES, University Mohammed V, Rabat- Morocco.

[3] Tommi A Pirinen and Krister Lindén, “Creating and Weighting Hunspell Dictionaries as Finite-State Automata”, University of Helsinki, Department of Modern Languages Unionkatu 40, FI-00014 University of Helsinki, Finland.

[4] Rohit Saluja, Devraj Adiga, Ganesh Ramkrishnam, Parag Chaudhari, “A Framework for Document Specific Error Detection and Corrections,” in Indic OCR" The context in spell correction, IIT-Bombay.

[5] Scott Nesbitt; “Check spelling at the linux command line with Aspell,” opensource.com.