Vol. (1) No. (1) March 2018

JK Research Journal in Mathematics and Computer Sciences

A study of spell checking techniques for Indian Languages

*Rakesh Kumar[1], Minu Bala[2], Kumar Sourabh[3]

[1]*Government Degree College Billawar, J&K-India*

[2]*Government Gandhi Memorial Science College Jammu, J&K- India*

[3]*Government Gandhi Memorial Science College Jammu, J&K- India*

Abstract: Spell checker is one of the most important tools for any language to be digitized. A spell checker is a software tool / plugin that identifies and corrects any spelling mistakes in a text in a particular language. Spell checkers can be combined with other research areas focusing on linguistics like Machine translation, Information retrieval, natural language processing etc. or they can be clubbed with other softwares like compilers, word processor softwares. In this paper authors have made a study on the developmental approaches as well as roles of spell checker with respect to various applications based on Indian languages.

Keywords: *Error Detection; Error Correction: Spell Checker; Machine translation; Information retrieval; N-gram.*

Introduction

There are many commercial as well as non commercial spelling error detection and correction tools available in the market for almost all popular languages. And every tool works on word level with the help of integral dictionary/Wordnet as the backend database for correction and detection. Every word from the text is looked up in the speller lexicon. When a word is not in the dictionary, it is detected as an error. In order to correct the error, a spell checker searches the dictionary/Wordnet for the word that is most resembled to the erroneous word. These words are then suggested to the user to choose the intended word. Spelling checking in used in various applications like machine translation, search, information retrieval etc. Spell checking technique comprises of two stages mainly error detection and error correction. In this paper we have studied various issues related spell checking techniques available so far as well as developmental approaches for error detection and correction for Indian languages.

Types of Spell Errors

In this section various techniques that were designed on the basis of spelling errors and trends also called error patterns have been studied. The most notable among these are the studies performed by Damerau [1]. According to these studies spelling errors are generally

rahulgoswami95@gmail.com
*Rakesh Kumar

divided into two types Typographic errors and Cognitive errors.

Typographic errors (Non Word Errors)

These errors occur when the correct spelling of the word is known but the word is mistyped by mistake. These errors are mostly related to the keyboard and therefore do not follow any linguistic criteria. A study by Damerau[1] shows that 80% of the typographic errors fall into one of the following four categories.

1.Single letter insertion; e.g. typing "apress" for presstyping "वालक" for बालक. 2.Single letter deletion, e.g. typing "pres" for press 3.Single letter substitution, e.g. typing "acress" for across typing "मोसम" for मौसम 4. Transposition of two adjacent letters, e.g. typing "acress" for caress. typing "मौमस" for मौसम These erroneous words are considered as misspelling of others words in the language. The errors produced by any one of the above editing operations are also called single-errors.

Cognitive errors (Real Word Errors)

These errors occur when the correct spellings of the word are not known. In the case of cognitive errors, the pronunciation of misspelled word is the same or similar to the pronunciation of the intended correct word.

"peace" for piece "कलम" for कमल "मकान उस **और** है" for मकान उस **ओर** है The words कलम and और are correct Hindi word but are misspelled instead of कमल and ओर respectively.

Error Detection & Correction

For error detection each word in a sentence or paragraph is tokenized by using a tokenizer and checked for its validity. The candidate word is a valid if it has a meaning else it is a non word. Two commonly used techniques for error detection is N-gram analysis and dictionary/Wordnet lookup.

Error correction consists of two steps: the generation of candidate corrections and the ranking of candidate corrections. The candidate generation process usually makes use of a precompiled table of legal n-grams to locate one or more potential correction terms. The ranking process usually invokes some lexical similarity measure between the misspelled string and the candidates or a probabilistic estimate of the likelihood of the correction to rank order the candidates. These two steps are most of the time treated as a separate process and executed in sequence. Some techniques can omit the second

process though, leaving the ranking and final selection to the user. The isolated-word methods that will be described here are the most studied spelling correction algorithms, they are: *edit distance [2], similarity keys [3], rule-based techniques [4], n-gram-based techniques[5], probabilistic techniques[6], neural networks[7] and noisy channel model [8,9]*. All of these methods can be thought of as calculating a distance between the misspelled word and each word in the dictionary or index. The shorter the distance the higher the dictionary word is ranked.

Error Detection Approaches

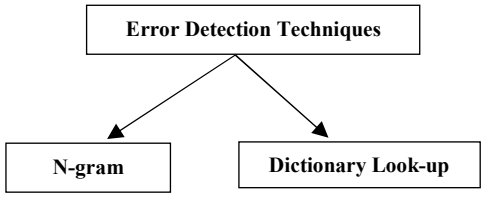

Fig. 1

N-gram Analysis

R.K Sharma [10] suggested in his studies that N-gram analysis is a method to find incorrectly spelled words in a mass of text. Instead of comparing each entire word in a text to a dictionary, just n-grams are checked. A check is done by using an n-dimensional matrix where real n-gram frequencies are stored. If a non-existent or rare n-gram is found the word is flagged as a misspelling, otherwise not. An n-gram is a set of consecutive characters taken from a string with a length of whatever n is set to. This method is language independent as it requires no knowledge of the language for which it is used. In this algorithm, each string that is involved in the comparison process is split up into sets of adjacent n-grams. The similarity between two strings is achieved by discovering the number of unique n-grams that they share and then calculating a similarity coefficient, i.e. the number of the n-grams in common (intersection), divided by the total number of n-grams in the two words (union)

Dictionary Lookup

A dictionary/Wordnet is a lexical source that contains list of correct words a particular language. The non-word errors can be easily detected by checking each word against a dictionary. The drawbacks of this method are difficulties in keeping such a dictionary up to date, and sufficiently extensive to cover all the words in a text. These resources are used for preparing, processing and

managing linguistic information and knowledge needed for the computational processing of natural language [11]. An example of such large scale lexical resources is given by linguistic ontology that covers many words of a language and has a hierarchical structure based on the relationship between concepts. We suggested to use these dictionaries, and especially WordNet [12], the most important lexical resource available. It covers nouns, verbs, adjectives and adverbs. Therefore, we extracted all words contained in it with all its linguistic relationships.

Error Correction Approaches

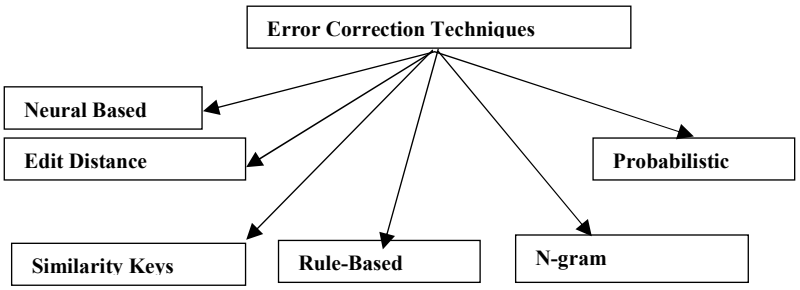

Fig.2

Edit Distance

It is one of the simplest methods based on the assumption that the person usually makes few errors i.e. only one erroneous character operation (insertion, deletions, and substitutions) necessary to covert a dictionary word in to the non word. Edit distance is useful for correcting errors resulting from keyboard input, since these are often of the same kind as the allowed edit operations. It is not quite as good for correcting phonetic spelling errors, especially if the difference between spelling and pronunciation is big as in English or French.

Similarity Keys

A index/key is assigned to each dictionary word for comparing with the key computed for the non word .The word for which the keys computed for the non word .The word for which the keys are most similar are selected as suggestion .such an approach is speed effective as only the words with similar keys have to be processed With a good transformation algorithm this method can handle keyboard errors.

Rule-based Techniques

Rule-based methods are interesting. They work by having a set of rules that capture common spelling and typographic errors and applying these rules to the misspelled word. Intuitively these rules are the "inverses" of common errors. Each correct word generated by this process is taken as a correction suggestion. The rules also have probabilities, making it possible to rank the suggestions by accumulating the probabilities for the applied rules. Edit distance can be viewed as a special case of a rule-based method with limitation on the possible rules.

N-gram-Based Techniques

N-grams can be used in two ways, either without a dictionary or together with a dictionary. Used without a dictionary, n-grams are employed to find in which position in the misspelled word the error occurs. If there is a unique way to change the misspelled word so that it contains only valid n-grams, this is taken as the correction. The performance of this method is limited. Its main virtue is that it is simple and does not require any dictionary. Together with a dictionary, n-grams are used to define the distance between words, but the words are always checked against the dictionary. This can b done in several ways, for example check how many n-grams the misspelled word and a dictionary word have in common, weighted by the length of the words.

Probabilistic techniques

They are, simply put, based on some statistical features of the language. Two common methods are transition probabilities and confusion probabilities. Transition probabilities are similar to n-grams. They give us the probability that a given letter or sequence of letters is followed by an16 other given letter. Transition probabilities are not very useful when we have access to a dictionary or index. Given a sentence to be corrected, the system decomposes Each string in the sentence into letter n-grams and retrieves word candidates from the lexicon by comparing string n-grams with lexicon-entry n-grams. The retrieved candidates are ranked by the conditional probability of matches with the string, given character confusion probabilities. Finally, a word-bigram model and a certain algorithm are used to determine the best scoring word sequence for the sentence. They claim that the system can correct non-word errors as well as real word errors and achieves a 60.2 % error reduction rate for real OCR text.

Neural Networks

Neural networks are also an interesting and promising technique, but it seems like it has to mature a bit more before it can be used generally. The current methods are based on back-propagation networks, using one output node for each word in the dictionary and an input node for every possible n-gram in every position of the word, where n usually is one or two. Normally only one of the outputs should be active, indicating which dictionary words the network suggests as a correction. This method works for small (< 1000 words) dictionaries, but it does not scale well. The time requirements are too big on traditional hardware, especially in the learning phase.

Survey of ExitingSpell Checkers

There are many spell checkers for Indian languages are developed by using above techniques. This section provides brief discussion and evaluation of some available spell checkers.

A. *Bangla Spell Checker [13*] P.Kundra et.al in their study suggested Bangla Spell Checker which can act in both offline and online notes. It has a graphics user interface via an editor. Whenever the user types bangle text, it checks for wrong spelling and gives suitable suggestion. On the other hand, one can run this software on previously typed material such as – An OCR Document. For a single error word, word is found within top 4 words in the suggestion list. Words having more than 1 error are so captured and for most of them, words are in the upper half in the suggestion list. However, Suggestions cannot be given on some inflected words. It has facility to add new words in the dictionary against which spellings are checked.

B. *Oriya Spell Checker [14*] Manisha Das et.al in their research have come out with error detection and automatic or manual correction for miss spelled words successfully by Oriya Spell Checker tool developed by them. Their study suggest that there has been developed some algorithm to perform OSC in order to find out more accurate suggestion for a miss spelled word. The words are indexed according to their word length in word's data base in order for effective searching. On the basis of the miss spelled word, it takes into account the number of:

i) Matching Characters,

ii) Matching Characters in the forward direction,

iii) Corresponding Matching Characters in backward direction to give more accurate, suggestive words for the miss spelled word. The OSC is running successfully in word processor using this technique, Hindi spell checker and English spell checker has also been developed for word processor.

The word files are stored in ASCII Format and supports Oriya, Hindi and any font in English. This software is designed using Java and Java Swing for both the Windows 98 / 2000 / NT and the Linux – OS.

C. *Marathi Spell Checker [15]* Prof.Puspak Bhattcharya a renowned Indian scientist came up with a standalone spellchecker is being built for Marathi. The spellchecker will be available to spell Check document in a given Encoding. From the CIIL (Central Institute of Indian Language) Corpus, 12886 distinct words have been listed. A morphological analysis is being carried out on the collection of words. For e.g., An automatic grouping algorithm identified 3975 groups out of 12886 distinct word. 1st word is usually the route word. Thus there are approximately 4000 route words from Marathi Corpus. A manual proof reading will be done on these results. The morphology will be enriched.

D. Annam (Tamil Spell Checker) [16] Tamil spell checker is used as a tool to check the spelling of Tamil words. It provides possible suggestion for erroneous words. User has the provision to select the suggestion among the list, ignore the suggestion or add the particular word to the dictionary. This module extract the route word from the given word (Noun / Verb) with the help of morphological analyzers and the route word is checked in dictionary and is found, the word is termed as correct word. Otherwise, the correction process is activated. The correction process includes error handling and suggestion generation module. After finding the types of error, the right form of suffix Noun or Verbs are given as input to the suggestion generation module. With the help of morphological generator the correct word is generated, And this module also handles the operation like – Select, change or Ignore the suggested word and adding the word to the dictionary.

E. Malayalam Spell Checker [17] It is a software sub system that can be executed with Microsoft word as a macro or the Malayalam editor, developed by CDAC, Tiruvananthpuram, to check the spelling of words in a Malayalam text file. While running as a macro in a word, it functions as an offline spell checker in the sense that one can use this software with the previously typed text file only. Both offline and online checking are possible when it is integrated with the text editor. It generates suggestion for wrongly spelled words. The system based on dictionary look up approach. This module split the input words into route word, suffixes, post positions etc. Checks the validity of each using the rule database finally it will check the dictionary to find the route word is present in the dictionary. If anything goes wrong in the checking. It is detected as an error and the error word is reprocessed to get three 3 – 4 valid words which are displayed as suggestion. The user can add new

words into a personalized database file, which can be added to the dictionary, if required. is integrated with the text editor. It generates suggestion for wrongly spelled words. The system based on dictionary look up approach. This module split the input words into route word, suffixes, post positions etc. Checks the validity of each using the rule database finally it will check the dictionary to find the route word is present in the dictionary. If anything goes wrong in the checking. It is detected as an error and the error word is reprocessed to get three 3 – 4 valid words which are displayed as suggestion. The user can add new words into a personalized database file, which can be added to the dictionary, if required.

F. Akhar (Punjabi Spell Checker)

A language sensitive Punjabi / English spell checker has been provided in Akhar. Akhar can automatically detect the language and invokes the respective spell checker. The Unicode complaint Punjabi Spell Checker is font independent and can work on any types of the popular Punjabi fonts such as, Anantpur Sahib, Amritlipi, Jasmine, Punjabi, Satluj etc. This removes the contrast on the user to type the text in pre defined font only**.**

Conclusion

In this paper we have studied the area of Spell checking techniques as well as various detection and correction techniques that are useful in finding the text with error. A study has been done in context with various spell checkers available in Indian languages. Objective of our study is to design spell checker for Dogri as well as Urdu language with approach based on dictionary lookup techniques for detection and minimum edit distance techniques for correction of result. As we have come to a conclusion in our study that design based on above mentioned approaches by various researchers is a suited design for Dogri language also, because of similar morphological structure of the language. For implementing the design of spellchecker in Urdu a separate study shall be required to understand structure of Urdu language.

References

[1] F.J Damerau(1964), “A technique for computer detection and correction of spelling error”, Communication ACM.

[2] R. A. Wagner and M. J. Fisher, “The string to string correction problem,” Journal of Assoc. Comp. Mach., 21(1):168-173, 1974.

[3] J. J. Pollock and A. Zamora, "Collection and characterization of spelling errors in scientific and scholarly text," Journal Amer. Soc. Inf. Sci., Vol. 34, No. 1, pp. 51–58, 1983.

[4] E. J. Yannakoudakis and D. Fawthrop, "An intelligent spelling error corrector," Information Processing and Management, 19:1, 101-108, 1983.

[5] Jin-ming Zhan, Xiaolong Mou, Shuqing Li, Ditang Fang, "A Language Model in a Large-Vocabulary Speech Recognition System," in Proc. of Int. Conf. ICSLP98, Sydney, Australia, 1998.

[6] K. Church and W. A. Gale, "Probability scoring for spelling correction," Statistics and Computing, Vol. 1, No. 1, pp. 93–103, 1991.

[7] V. J. Hodge and J. Austin, "A comparison of standard spell checking algorithms and novel binary neural approach," IEEE Trans. Know. Dat. Eng., Vol. 15:5, pp. 1073-1081, 2003.

[8] E. Brill and R. C. Moore, "An improved error model for noisy channel spelling correction," in Proc. 38th Annual Meet. of the Assoc. for Comp. Ling., Hong Kong, 2000, pp. 286–293.

[9] K. Toutanova and R. C. Moore, "Pronunciation modeling for improved spelling correction," in Proc. 40th Annual Meeting of the Assoc. for Comp. Ling., Hong Kong, 2002, pp. 144–151.

[10] R.K Sharma, "The Bilingual Punjabi English spell checker", Resource centre for Indain language Technology Solution, TDIL newsletter.

[11] W. Peters, "Lexical Resources," NLP group, Dept. of Comp. Sc., Uni. of Sheffield, 2001.

[12] C. Fellbaum, "WordNet, an electronical lexical database," Cambridge, MIT Press, 1998.

[13] P.Kundra and B.B Charudhari (1999),"Error pattern in Bangla text", international Jounral of Dravidian Linguistics.

[14] Manisha Das,S.Borgohain,Juli Gogai,S.B Nair (2002),"Desingn and implemention of a spell checkers for Assamese",Language Engineering Conference.

[15] Prof.Puspak Bhattcharya and Prof. Rushikersh Josh, "Design and implantation of morophology based spellcheckers for Marathi",TDIL Newsletter.

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

[17] R.Ravindra Kumar ,K.G S ulochana,"Malayalam spell checker ",Resource centre for Indain language Technology Solution ,TDIL newsletter.