International Journal For Technological Research In Engineering
Volume 2, Issue 3, November-2014 *ISSN (Online): 2347 - 4718*

STUDY OF SPELL CHECKING TECHNIQUES AND AVAILABLE SPELL CHECKERS IN REGIONAL LANGUAGES: A SURVEY

Er. Sumreet Kaur Randhawa[1], Er.Charanjiv Singh Saroa[2]
[1]Student, [2]Assistant professor, Department of computer engineering
Punjabi university Patiala, Punjab, India

Abstract: Spellchecker is software that analyzes possible misspellings in the text. It is the process of detecting and sometimes providing some suggestions for incorrectly spelled words in a text. If dictionary of spell checker is larger than higher is the error detection and error correction rate. Though considerable work has been done in English language but not much work has been done in regional language of India including Punjabi. Punjabi world's 12th most widely spoken language. The only available spell checker for Punjabi is AKHAR and SUDHAR. NLP (Natural language processing) is a field of computer science concerned with interaction between computer and human language. In this paper we have discussed different techniques and about different spell checker available in different Indian regional languages and study about all spell checker whose websites are available and check their efficiency through their ranking. In future we will make a Punjabi spell checker. We will combine concept of AI (artificial intelligence) and NLP (natural language processing) to create Punjabi spell checker so that user can get appropriate word for misspelled word.

KEYWORDS: Punjabi spell checker, Gurmukhi spell checker, Error detection, Error correction, N-gram, Typing errors, NLP, misspelled words

I. INTRODUCTION

With the advent of the personal computer, it can be assumed that mistyping of words has increased. Thus, for many of us who do our own typing, the spell checkers in our word processors or other software have become indispensable. Spell Checker has following steps:-

1. Take a word as input from a file.
2. Preprocess it.
3. Check in the word whether that word is available.
4. If it is available then move to next one.
5.If word is not present then spell checker will check the closest match word with it and put it in the form of suggestions.

Thus two main issues related to spell checker are error detection and error correction. Spelling errors can be partitioned into different categories, that is, real-word error and non-word error. In this paper I am describing techniques of error detection, error correction, available spell checkers and efficiency of spell checkers in different languages whose websites are available. Error means a measure of the estimated difference between the observed and calculated value Spelling and typing errors are common errors made by humans. Errors may be of missing letters, extra letters, misspelled letters, or disordered letters

II. LITERATURE SURVEY

A we discuss there are two main issues related to spell checker i.e. error detection and error correction .Further there are two types of errors these are non-word errors and real – word errors or errors may be classified as Typographic errors and Cognitive errors. Many techniques are available for non-word errors. The error detection process usually consists of checking to see if an input string is a valid index or dictionary word. Efficient techniques have been devised for detecting such types of errors. The two most known techniques are n-gram analysis and dictionary lookup. Error correction means just to replace the incorrect with most likely corrected word. Techniques available for error correction are Edit distance, Similarity keys, Rule based technique n-gram based technique, neural technique, Probabilistic technique and neural network. Available websites for spell checkers for different languages are Hindikhoj.com, Star21.com, Shabdkosh.com, khandbahale.com, Shuddhoshabdo.com, Spellweb.com.

III. IMPORTANCE OF PUNJABI LANGUAGE

Language like Punjabi which has literary history older than English is 12th most largest spoken language of world has now ultimately came to an end. This is due to absence of local languages in educational system because schools and colleges play a necessary role in preserving languages and culture. Almost 110 million people possess Punjabi as their mother tongue but now English language is like punishing language in schools for students and we all know one could better understand things in mother tongue rather than in other language. Although Asian countries like china, japan, are teaching their students in their mother tongue. Suppose in a school science subject is there and rather than understanding concept of science or getting practical knowledge we pay attention to English words used in it. So Punjabi spell checker act as savior of Punjabi language.

IV. TECHNIQUES OF SPELL CHECKER

A. ERROR DETECTON

1. N GRAM ANAYLSIS

N-gram analysis is described as a method to find incorrectly spelled words in text and used for non-word errors. Instead of comparing each entire word in a text to a dictionary, just n-grams are controlled. 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. If n is set to one then the term used is a unigram, if n is two then the term is a Bigram, if n is three then the term is trigram. N-gram tables can take on a variety of forms but the simplest is bi-gram array which is 2-D array of size 41*41 whose element represents all possible 2-D letter combination of alphabet. The n-grams algorithms have the major advantage that they require no knowledge of the language that it is used with and so it is often called language independent or a neutral string matching algorithm. Using n-grams to calculate for example the similarity between two strings is achieved by discovering the number of unique n-grams that they share and then calculating a similarity coefficient, which is the number of the n-grams in common (intersection), divided by the total number of n-grams in the two words (union)

2. DICTIONARY LOOKUP

This technique simply lookup every word in the dictionary if the word is not there then it is said to be an error. Dictionaries have their own characteristics and storage requirements. It is a straightforward task. Large dictionary might be a dictionary with most common word combined with a set of additional dictionaries for specific topics such as computer science or economy .Big dictionary also uses more space and may take longer time to search. The non-word errors can be detected as mentioned above by checking each word against a dictionary. The drawbacks of this method are difficulties in keeping such a dictionary up to date. At the same time one should keep down system response time. Too small a dictionary can give the user too many false rejections of valid words. The most common technique used for gaining fast access in dictionary is Hash tables. In order to lookup a string, one has to compute its hash address and retrieve the word stored at that address in the pre constructed hash table. If the word stored at the hash address is different from the Input string, a misspelling is flagged. Hash tables main advantage is their random-access nature that eliminated the large number of comparisons needed to search the dictionary. The main disadvantage is the need to devise a clever hash function that avoids collisions. To store a word in the dictionary we calculate each hash function for the word and set the vector entries corresponding to the calculated values to true. To find out if a word belongs to the dictionary, you calculate the hash values for that word and look in the vector. If all entries corresponding to the values are true, then the word belongs to the dictionary, otherwise it does not.

B. ERROR CORRECTION TECHNIQUES

1. EDIT DISTANCE

Edit Distance Edit distance is a simple technique. first edit distance spelling error correction algorithm was implemented by Damerau Simplest method is based on the assumption that the person usually makes few errors if ones,i.e. errors from keyboard input therefore for each dictionary word .The minimal number of the basic editing operations (insertion, deletions, substitutions) necessary to covert a dictionary word in to the non-word As 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.

2. N-GRAM TECHNIQUE

N-grams can be used in two ways, either without a dictionary or together with a dictionary. Letter N-gram including tri-gram, bi-gram and uni- gram have been used in variety of ways in text recognition and spelling correction techniques. 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. It 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.

3. SIMILARITY KEYS

In this technique we map every string into a key such that the similarly spelled strings will have similar key. It is known as SOUNDEX system. In this it is not necessary to directly compare misspelled string to each word in dictionary. For example suppose a customer comes in a bank and said his name is zayheijendn. So in this case you cannot ask him to speak his name as his English is poor and others customers are waiting. So we want a key which sounds like his name and find a name resembles with it.

4. RULE BASED TECHNIQUE

It attempts to represents knowledge commonly spelled errors i.e. mistyping by mistake in the form of rules for converting it into a valid words. Each word which is correct can be taken as a suggestion. It consist the process consist of applying all applicable rules to a misspelled string.

5. NEURAL NETWORK

This method works on small dictionaries. Back propagation algorithm is used in neural network.it consist of 3 layers input, hidden and output layer. It has potential to adapt specific error pattern. In this input information is represented by on- off pattern. A=1 indicates that node is turned on and A=0 means node is turned off. For e.g. in spell checking applications a misspellings represented as binary n-gram vector may be taken as input and output pattern might be vector of m elements means number of words in lexicon.

6. PROBALISTICS TECHNIQUES

Based on some statistical features of the language-gram technique led to probalistic technique in spell correction and text recognition. Two methods are used in this. Transition probabilities which is similar to n-grams .It is language indepdent.C onfusion probabilities estimates of how often a given letter is mistaken. It is source indepdent.

V. AVAILABLE SPELL CHECKERS

There are many spell checkers for Indian languages are developed by using above techniques. This section provides brief discussion of some available spell checkers and websites available for those spell checkers. The number of spell checker available for different regional languages are

Hindi, Marathi, Bangla, Tamil, Malayalam, and Punjabi spell checkers.

A. Hindi spell checker

Name of spell checker	Hindi
website	Shabdkosh.com
Global rank	3,497
India rank	296
Daily page viewer per visitor	4.62
Total site linking in	462

B. Marathi Spell Checker

Under this ongoing project, 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.

Name of spell checker	Marathi
Website	Khandbahale.com
Global rank	320,723
India rank	30,873
Daily page viewer per visitor	2.50

C. Bangla Spell Checker

Bangla Spell Checker can act in both offline and online notes. It has a graphics user interface via an editor. When we typed any text it checks for wrong spelling and gives suitable suggestion. 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. It has facility to add new words in the dictionary against which spellings are checked.

Name of spell checker	Shuddhoshabdo.com
Global rank	19,779,445
India rank	N/A
Daily page viewer per visitor	1

D. Oriya Spell Checker

It is available online Named NAAVI. Error detection and automatic or manual correction for miss spelled words, is taken care successfully by Oriya Spell Checker. 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.

E. Annam (Tamil Spell Checker)

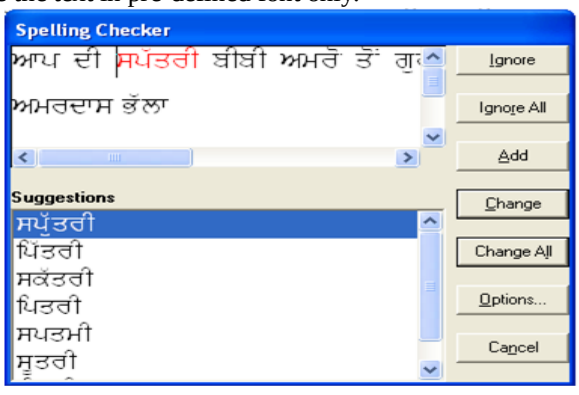
Tamil spell checker is used as a tool to check the spelling of Tamil words. It is available both offline and online. It provides possible suggestion for misspelled words. User has the provision to select the suggestion among the list, ignore the suggestion or add the particular word to the dictionary. 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. 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.

F. Malayalam Spell Checker-

It is a software sub system that can be executed with Microsoft word as a macro or the Malayalam editor, developed by CDAC, Thiruvananthapuram, 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.

G. 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.



VI. CONCLUSION AND FUTURE WORK

In this paper we have surveyed the spell checker techniques and websites of available spell checker in regional language. We have also discussed about importance of Punjabi language .In future I will make a Punjabi spell checker with the help of these techniques with efficient database. In future we will combine concept of AI (artificial intelligence0 and NLP (natural language processing) to create a Punjabi spell checker so that user can get appropriate suggestions for misspelled words.

REFERENCES

[1] Rupinderdeep Kaur and Parteek Bhatia, "Design and Implementation of SUDHAAR-Punjabi Spell Checker," International Journal of Information and Telecommunication Technology, Vol. 1, Issue 15 May, 2010.
[2] S. Dasgupta, C.H. Papadimitriou, and U.V. Vazirani, 'Algorithms', p173, available at http:/ / www.cs.berkeley.edu/~vazirani/algorithms.html.
[3] Neha Gupta &PratisthaMathur, "Spell Checking Techniques in NLP: A Survey," International Journal of Advanced Research in Computer Science and Software Engineering, Vol. 2, Issue 12, December 2012.
[4] Gurpreet Singh Lehal, "Design and Implementation of Punjabi Spell Checker", International Journal of Systemics, Cybemetics and Infomatics, 2007.
[5] G S Lehal&MeenuBhagat, "Spelling Error Pattern Analysis of Punjabi Typed Text", In Proceedings of International Symposum on Machine Translation, NLP and TSS, pp. 128-141, 2007.
[6] Jesus Vilares& Manuel Vilares, "Managing Misspelled Queries in IR Application," Issue 8, October 2010. Vol. 5, No.3, May 2012]
[7] Daniel Jurafsky, James H. Martin, Speech and Language Processing, PEARSON, 2nd ed. [2] Dr.T.V Geeta, Dr.Rajani Parthearath, "Tamil spellchecker", Resource centre for Indian language Technology Solution, TDIL newsletter
[8] Dr. R.K Sharma, "The Bilingual Punjabi English spell checker," Resource center for Indian language Technology Solution, TDIL newsletter
[9] "F.J Damerau (1964)"A technique for computer detection and correction of spelling error", communication ACM.
[10] Http://www.Baraha.com
[11] Manisha Das, S.Borgohain, Juli Gogai, S.B Nair (2002),"Design and implementation of a spell checkers for Assamese", Language Engineering Conference.
[12] Mukand Roy, Gaur Mohan, Karunes K arora, "Compartive study of spell checker algorithm for building a generic spell checkers For Indian language C-DAC NODIA, India.
[13] Malayalam spell checker, Santhosh. T. Varghese, K.G. Sulochana, R. Ravindra Kumar
[14] Alexa.com

