Survey of Spell Checking Techniques for Malayalam: NLP

Shahana Basheer[1], Sindhu L[2]

M Tech Student, Dept of Computer Science and Engineering, College of Engineering Poonjar, Kottayam, India

Head of the Department, Dept of Computer Science and Engineering,College of Engineering Poonjar,Kottayam, India[2]

Abstract- **Spell checking is a well-known task in Natural Language Processing. Nowadays, spell checkers are an important component of a number of computer software such as web browsers, word processors and others. Spelling error detection and correction is the process that will check the spelling of words in a document, and in occurrence of any error, list out the correct spelling in the form of suggestions. This survey paper covers different spelling error detection and correction techniques in various languages.**

Keywords- **NLP, Spell Checker, Spelling Errors, Error detection techniques, Error correction techniques.**

I. INTRODUCTION

Spelling error correction is a Natural Language Processing (NLP) problem, and it has recently become relevant because many of the potential NLP applications such as text summarization, sentiment analysis and machine translation etc take advantage of spelling error analysis.

An error detector that detects misspelled words, a candidate spelling generator that provides spelling suggestions for the detected misspelled word, and an error corrector that chooses the best correction out of the list of candidate spellings. All spelling checker tools use a dictionary as a database. Every word from the written text is looked up in the dictionary. When a word is could not find in the dictionary, it is detected as an error. To correct that error, a spell checker searches the dictionary for words that resemble the erroneous word most. These words are then listed as suggestions to the user who chooses the best word that was expected. There are two main steps in a spell checker, are Error detection and Error correction.

II. SPELLING ERRORS

Techniques for spelling error detection were designed on the basis of different spelling error trends these are also called error patterns. Studies were performed to analyze various trends in spelling errors.

Kukich (1992)[1] breaks the field of error detection into three increasingly broader problems:

- Non-word error detection: detecting the spelling errors that result in **non-words** (*bok* for *book*)
- Isolated-word error correction: correcting spelling errors that result in non-words. (correcting *bok* to *book*, but looking only at the word in isolation.)
- Context-dependent error detection and correction: using the context to help detect and correct **real-word errors**. (*dessert* for *desert* or *there* for *three*

According to Damerau [2] spelling errors are generally divided into two types Typographic errors and Cognitive errors.

A. Typographic errors

These errors are occurring 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.

B. Cognitive errors

These are errors occurring when the correct spellings of the word are not known. In these type of errors, the pronunciation of misspelled word is the intended as correct word.

III. GENERAL TECHNIQUES FOR ERROR DETECTION

Techniques to detect non word spelling errors in a text can be divided into two categories: dictionary lookup and n-gram analysis. A non word refers to a continuous string of characters and/or numbers that cannot be found in a given dictionary or that is not a valid orthographic word form. Dictionary lookup technique employs efficient dictionary lookup algorithms and/or pattern matching algorithms. N-gram analysis makes use of frequency counts or probabilities of occurrence of n-grams in text and in a dictionary, a lexicon or a corpus.

A. Dictionary Lookup Technique

Dictionary lookup technique checks every word of input text for its presence in dictionary. If that word present in the dictionary, then it is taken as a correct word. Otherwise it is put into the list of error words. The most significant dictionary lookup techniques are hashing, binary search trees and finite state automata.

1) Hashing: Hashing [3] is a technique used for searching an input string in a pre-compiled hash table via a key or a hash address associated with the word and retrieving the word stored at that particular address. In spell checking problem, if the word stored at the hash address is same as the input string there is a match. If the word stored in hash table is null the input word is indicated as a misspelling. This technique eliminates the large number of comparison required for lookups.

2) Binary Search Trees: Binary Search Trees are useful [3] for checking if a particular string, i.e. an input word exists within a large set of strings i.e. the dictionary. The main goal of binary search trees particularly median split tree is to make access to high frequency words faster than to low frequency words.

It is efficient compared to the lookup time of a linear search technique on a large data representation although it is slower compared to the lookup time of hashing.

3) Finite State Automata: Finite state automata used as a basis for string matching or dictionary lookup algorithms that locate elements of a dictionary within an input text. One specific form of FSA that has used for spell checking and correcting purposes is a trie data structure. Tries are also known as prefix trees [3]. Finite state approaches are used for spelling correction for agglutinating languages or languages with compound nouns.

B.N-gram analysis

N-gram analysis [5] is used to detect incorrectly spelled words in a mass of text. Here instead of comparing the complete word in a text to a dictionary, only n-grams are compared with dictionary because comparing each single word with dictionary is a time consuming process. It uses an n-dimensional matrix, where the actual n gram frequencies are stored is used for spell checking.

If a non-existent or rare n-gram is detected the word is flagged as an error or misspelled, otherwise not. An n-gram is a set of consecutive characters taken from a string with a length of n. If the value of n is set to one then it is called unigram, if n is two then it is a Bigram, similarly if n is three then the term is trigram. Every string that is involved in the comparison process is split up into sets of adjacent n-grams. The major advantage of n-grams algorithms are that they require no knowledge of the language that it is used with and so it is often called language independent algorithm.

IV.GENERAL TECHNIQUES FOR ERROR CORRECTION

Spell correcting refers to find the subset of dictionary or lexical entries that are similar to the misspelling in some way. Spell checking can be categorized by isolated word error correction and context dependent error correction. Isolated word error correction thus refers to spell correcting without taking into account any textual or linguistic information in which the misspelling occurs. A context dependent corrector would correct both real word errors and non word errors involving textual or linguistic context.

Kukich [4] pointed out that 80% of spelling errors tend to be single-letter errors, such as insertions, deletions, substitutions, and transpositions. Spelling error correction relies on some approximate string matching technique to find a set of correctly spelled words in the dictionary that satisfy a similarity relation. This involves the association of a misspelled word with one word or a set of correctly spelled words in the dictionary that satisfy a similarity relation [3].

According to [2] 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.

A. Minimum Edit Distance

Most studied technique for spelling correction. Here the minimum number of editing operations (insertions, deletions, substitutions and transpositions) required to transform one string into another. Wagner [3] introduced the notion of edit distance for spelling correction. Minimum edit distance has different algorithms are Levenshtein algorithm, Hamming, Longest Common Subsequence [6].

B. Similarity Keys

A key is assigned to each dictionary word and only these keys are compared with the key computed for the non word. The words for which the keys are most similar are listed as suggestions. Such an approach is speed effective only if the words with similar keys have to be processed with a good transformation algorithm. This method can handle keyboard errors.

This technique has two different approaches Soundex Algorithm and The SPEEDCOP System [6].

C. Rule-based Techniques

Rule-based methods are interesting. This technique having a set of rules that collect common spelling and typographic errors and applying these rules to the misspelled word. 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 calculating 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.

D. N-gram-Based Techniques

N-gram based technique can be used in two ways, either together with a dictionary or without having a dictionary. Used without a dictionary, n-grams are employed to find in which position in the misspelled word the error occurs. 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.

E. Probabilistic techniques

This technique is based on some statistical features of the language. Two common methods are confusion probabilities and transition probabilities. Transition probabilities are similar to n-grams. This give the probability that a given letter or sequence of letters is followed by another given letter.

Transition probabilities are not very useful when we have access to a dictionary or index. Given a sentence which has to be checked, 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.

And, a word-bigram model and a certain algorithm are used to determine the best scoring word sequence for the sentence [6].

F. Neural Networks

Neural networks are also an interesting and promising technique. The current methods are based on back-propagation networks, it uses one output node for each word in the dictionary and an input node for every possible n-gram in every position of the word, and where n is one or two. 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. In the learning phase the time requirements are too big on traditional hardware. For training a neural net Back Propagation Algorithm is the most widely used one.

V.RELATED WORKS

Different techniques and methodologies are used in development of spell checking system in various languages. In English efficient spell checkers are available. But in highly inflectional languages like Malayalam researches are taking place to build an efficient system for this**.** Here we discuss different approaches for spelling error detection and correction in various languages.

In [7] authors discuss different stages involved in the development of a Malayalam spell checker. This is developed at ER&DCI (T). Here Spell checking is done with the aid of language specific rules and a dictionary. This system consists of two main modules Language and Engine module. But this is implemented with some limited functionalities. Due of the morphological richness of the language subparts like POS tagger and Morphological Analyzer are not so efficient.

A generalized discriminative model for spelling error correction which targets character-level transformations is presented on [8]. It uses supervised learning to map input characters into output characters in context. Character-level Corrections are learned at the character-level1 using a supervised sequence labeling approach. Here only the language independent features are taken. Authors make a brief description of existing works in this field through this paper. Egyptian Arabic data is considered for this. A detailed explanation of GSEC approach is contained in this paper. The results and evaluation matrices shows that this model outperforms in case of out of vocabulary words in Egyptian Arabic data. Studies can be extended to other languages also.

Anitha S Pillai on [9] presents an approach for automatic correction of spelling mistakes in Tamil document using the Finite State Automata (FSA). A 'Rule cum Dictionary' based approach is followed. The lexicon (dictionary) is stored in the form of finite state automata instead of using Hash tables. When the user types a word and if the word cannot be represented as a FSA an error is shown. This means that this particular word is not a correct one and so the FSA fails to move from the initial state to the final state. In such cases using minimum edit distance suggestions are displayed to the user to choose correct one from that list.

Naveen Sankaran and C V Jawahar on [10] propose a model which can detect errors for highly inflectional languages like Telugu and Malayalam with an F-Score comparable to some less inflectional languages like Hindi. In this study they investigate challenges involved in the development of error detection techniques for highly inflectional languages. Error detection experiments are conducted on Telugu and Malayalam languages. The result shows that statistical models are more efficient in these languages. According to evaluation matrices and calculated F-Score dictionary coupled with SVM (Support Vector Machine)s have better results comparing with others.

[11] Presents the construction of a spell checker for Sinhala. Due to its morphological richness, the language is difficult to enumerate completely in a single lexicon. The approach described is based on n-gram probability. Paper gives review of related works and description of the followed methodology. In this n-gram based spell checker for Sinhala by substituting phonetically similar characters in a given word, permutations are generated and sent to the best suggestion selection module. The best suggestion selection module uses three techniques for ranking the generated permutations. The three techniques are based on word unigram frequencies, syllable trigram frequencies and syllable bigram frequencies, which are pre-computed from a raw text corpus. Different algorithms for developing this are also discussed in [11].

On [12] the researcher's main focus is to develop a Spell checking System for Punjabi language. Implementation of such a system has so many challenges. Initially the study goes through some general Error Types, Application of Spell checking System and related works that conducted in this area. Finally certain conclusions are made to implement the system with hybrid approaches. Algorithm based on dictionary lookup and rule based techniques are developed for this.

VI.CONCLUSION

In this paper different Spelling error detection and correction techniques are discussed. The main objective of this survey is that to develop an accurate and efficient Spell checking system for Malayalam language. Which is the official language of Kerala and it belongs to Dravidian family. From these studies it seems that by the using certain techniques together, we can develop hybrid methods to achieve greater accuracy. The minimum edit distance method is the widely studied and used spell correcting techniques.

To develop a system for spelling error detection and correction in Malayalam so many challenges will be experienced. Due to the morphological richness of Malayalam some hybrid techniques are more efficient. Rule based error detection using some FSA lexicon can be an efficient system. Integrating it with some statistical models will give some accurate systems.

REFFERENCES

[1] Daniel Jurafsky, James H. Martin, *Speech and Language Processing*, Pearson

[2] F.J Damerau (1964) *A technique for computer detection and correction of spelling error*, Communications ACM.

[3] Hsuan orraine Liang *Spell Checkers and Correctors:A unified Treatment* , University of Pretoria,South Africa

[4] Karen Kukich *Techniques for Automatically Correcting Words in Text* ,Bellcore, South Street, Morristown

[5] Neha gupta, Pratistha mathur *Spell checking techniques in NLP: a survey*, Department of computer science, Banasthali vidyapith, India

[6] Ritika Mishra1, Navjot Kaur2 *A Survey of Spelling Error Detection and Correction Techniques*

[7] Santhosh. T. Varghese, K.G. Sulochana, R. Ravindra Kumar *Malayalam Spell Checker*

[8] Noura Farra, Nadi Tomeh, Alla Rozovskaya, Nizar Habash *Generalized Character-Level Spelling Error Correction*

[9] Anitha.S Pillai *Spell Checker for Tamil using Finite State Automata*, Hindustan University

[10] Naveen Sankaran and C. V. Jawahar *Error Detection in Highly Inflectional Languages* International Institute of Information Technology, Hyderabad, India

[11] Asanka Wasala, Ruvan Weerasinghe, Randil Pushpananda, Chamila Liyanage and Eranga Jayalatharachchi *A Data-Driven Approach to Checking and Correcting Spelling Errors in Sinhala*

[12] Amanjot Kaur, Dr. Paramjeet Singh, Dr. Shaveta Rani *Spell Checking and Error Correcting System for text paragraphs written in Punjabi Language using Hybrid approach*