Automatic Error Detection and Correction in Malayalam

Ambili T
Department of Computer Science & Engineering
Mar Athanasius College of Engineering Kothamangalam, Kerala

Panchami K S
Department of Computer Science & Engineering
Mar Athanasius College of Engineering Kothamangalam, Kerala

Neethu Suhash
Department of Computer Science & Engineering
Mar Athanasius College of Engineering Kothamangalam, Kerala

Abstract

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. Spell checking is a well-known task in NLP. 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. The proposed method develops a system for spelling error detection and correction in Malayalam. The proposed system uses a spell checker that detects the error by a dictionary lookup approach and error correction is done through N-gram based technique. In dictionary method, it checks each word of input for its presence in the dictionary. If the word is present in the dictionary then it is a correct word else put into the list of error words. N-gram based technique corrects error by finding similarity between words and computing a similarity coefficient. Due to morphological richness of Malayalam, error detection and correction is a challenging task, however the proposed system meets this challenge and has high accuracy compared to other existing approaches.

Keywords: Natural Language Processing (NLP), corpus, N-gram, Tokens, Lexicon

I. INTRODUCTION

NLP is a form of human-to-computer interaction where the elements of human language, be it spoken or written, are formalized so that a computer can perform value adding tasks based on that interaction. The goal of the Natural Language Processing (NLP) group is to design and build software that will analyse, understand, and generate languages that humans use naturally, so that eventually we will be able to address our computer as though we were addressing another person. Language is one of the most important capabilities of human for communication. Natural language cannot be absent in human communication either spoken communication or written text. As we known, word is the fundamental semantic unit in the most languages; it plays an essential role in natural language processing. Since the word is the building block for natural language processing, the spelling error or typos usually cause negative effects in word for computer applications. Spelling error correction is a Natural Language Processing 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. The problem of devising algorithms and techniques for automatically correcting words in text has become a perennial research challenge. Work began as early as the 1960s on computer techniques for automatic spelling correction and automatic text recognition, and it has continued up to the present. There are good reasons for the continuing efforts in this area. Although some excellent academic and commercial spelling checkers have been around for some time, existing spelling correction techniques are limited in their scope and accuracy. As a consequence, many current computer applications are still vulnerable to costly text, code, and data entry mistakes, although some good commercial text recognition devices are available today, they perform optimally only under ideal conditions in which input consists of clean text set in a standard type font. Furthermore, even a character recognition accuracy rate as high as 99% yields only a 95%. Word recognition accuracy rate, because one error per 100 characters equates to roughly one error per 20 words, assuming five character words.

One study [Cushman 1990] found that in order for optically scanned documents to contain no more residual errors than typed documents, at least 98% character recognition accuracy coupled with computer assisted proof reading would be required. Evolving human-computer and computer communications technologies have opened the door for a host of new applications that will require better word recognition and error correction capabilities. Pen based interfaces will enable users to provide handwritten input to computers; text recognition devices will make scanning of printed material feasible under everyday conditions; voice synthesis (text-to-speech) devices will make textual material audible; and voice recognition (speech-to-text) technology will eventually allow voice input to computer systems. But none of these applications will become practical until significant improvements are made in the area of word recognition and correction.
Some of the other applications that will benefit from such improvements include more sophisticated software tools for text and code editing, computer-aided authoring, machine translation, language learning, computer-aided tutoring, and database interaction, as well as various voice input and voice output business applications and aids for the disabled such as fax-to voice devices and phonetic-transcription services. Early on, researchers working within the paradigms of automatic spelling correction and automatic text recognition proceeded somewhat independently using different techniques. Over time the various techniques began to migrate between the fields so that today numerous hybrid approaches and many reasonably successful systems exist. A distinction must be made between the tasks of error detection and error correction. Efficient techniques have been devised for detecting strings that do not appear in a given word list, dictionary, or lexicon. But correcting a misspelled string is a much harder problem.

II. RELATED STUDIES

As 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. Surveying different spell checkers we came to know about that in Malayalam spell checker rule cum dictionary based approach is used, so far the error detection of words is almost complete for any form of agglutination that can come in Malayalam vocabulary. Only standard words are checked in this version and we have further plans to cater the non-standard words too. As far as suggestion generation is considered, only single character errors are taken into account and advanced versions will be taking care of multiple character errors also. Paper [2] describes all techniques of spell checker already available and also spells checkers of different 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. On [8] the researcher’s main focus is to develop a Spell checking System for Punjabi language. In paper automatic keywords extraction for Punjabi language they include various phrase like removing stop words, identification of Punjabi nouns etc. which is used for information retrieval, classification, clustering. 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. Egyptian Arabic data is considered for this. With the help of these papers we came to know about different techniques of spell checker and also about spell checker of different language. We had made a survey to check that what kind of errors a human being make while writing a word in Punjabi. 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.

III. PROPOSED METHODOLOGY

In this section, we want to illustrate the proposed system framework to detect and correct the spelling errors. Our goal is to find the locations and correct the corresponding error character in input Malayalam sentences. For more clear presentation, here in, the system framework is divided into three parts: architecture, error detection and correction phases as described in Section A, B and C respectively.

A. Architecture:

The system includes mainly two components: error detector and error corrector. Initially a corpus of Malayalam words are created that will act as a dictionary. With the input given by the user the error detector detects misspelled words by comparing input word with the words in the dictionary. The error corrector includes a candidate spelling generator that provides spelling suggestions for the detected misspelled word and chooses the best correction out of the list of candidate spellings. This is achieved by performing a similarity measurement between the input word and the words in the dictionary and hence computing a similarity coefficient.
B. Error Detection:

The main task in spell checking is to detect the errors in written text. There are two techniques for error detection is N-gram analysis and dictionary lookup. The error detection process usually consists of checking to see if an input string is a valid dictionary word or not. Efficient techniques have been devised for detecting such types of errors. Spellcheckers rely mostly on dictionary lookup and n-gram techniques.

1) Dictionary Lookup:

A dictionary is a list of large number of correct words. Dictionary look up is one of the two principal ways of spelling error detection. Dictionary looks up technique which checks each word of input text for its presence in dictionary. If that word is present in dictionary, then it is a correct word otherwise it is put into the list of error words [6]. The main disadvantage is the need to devise a clever hash function that avoids collisions. After finding the word incorrect various handcrafted rules are applied to generate the correct spellings of the word by considering the linguistic features of the particular language.

2) N-gram Analysis:

It is a method to detection of incorrectly spelled words in a mass of text. Instead of comparing each entire word in a text to a dictionary, just n-grams are compared with dictionary because comparing each single word with dictionary is a time consuming process. 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 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 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. Each string that is involved in the comparison process is split up into sets of adjacent n-grams. 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 algorithm.

3) Error Correction:

When an input word is detected as an error in written text then spelling correction techniques are applied on erroneous word to correct the word or providing correct suggestions for that word. This process is called spell correction. N-gram models can be imagined as placing a small window over a text, in which only n words are visible at the same time. The model in which we only look at one word at a time is called unigram model. In similar fashion, a bigram shows two words at a time on a window. N-grams can be thought of n number of words under observation. N-grams can be used in either without a dictionary or together with a dictionary [3]. Used without a dictionary, n-grams are employed to find in which position in the incorrect word the error occurs. If there is a unique way to change the incorrect word so that it contains only valid n-grams, this is taken as the correction. The performance of this method is low. Its main advantage 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 be done by using several ways, for example check how many n-grams the misspelled word and a dictionary word have common, weighted by the length of the words.

IV. SYSTEM DESIGN

A. Input Design:

All spell checking tools use dictionary as a database. In Error detection and correction system it includes a corpus of different Malayalam words which will act as a dictionary. The dictionary includes around 10,000 words. So when an input word is given it
is compared with words in the dictionary. If the word does not have a match with any words in the dictionary it is indicated as a misspelled word. The system generates possible suggestions for this incorrect word, from which the most probable word can be selected by the user.

**B. Output Design:**

There is a button on the main window i.e. Spell Check and is clicked when the user wants to check the word entered in the text area for spelling errors. When the user clicks on the button, if the input word is present in the dictionary no error is indicated, else the word is marked by a red line indicating an error has occurred. The misspelled word when right-clicked gives a list of correct words from which the probable word can be selected.

**V. IMPLEMENTATION**

The Error detection and Correction system identifies the misspelling in a word and if any error is present it provides the correct word in the form of a list of suggestions of the most probable words. It is implemented as two phases.

**A. Phase I:**

First phase is the error detection phase in which, the system flags words in a document that may not be spelled correctly. Error detection is done through a dictionary lookup approach.

**Algorithm:**

Input: a word
Output: incorrect word

1) Start
2) Read the input word.
3) Compare the input word with words in the dictionary.
4) If the word is present,
   1) return no error
5) Else,
   1) return error
6) End.

![Phase I Flowchart]

**B. Phase II:**

In this phase error correction is done if any word is incorrectly spelled. Error correction is implemented through n gram based technique.
Algorithm:
Input: a word
Output: list of possible correct words
1) Start
2) Read the input word. 
3) Split the input word into tokens.
4) Compute the n-gram of input word with the word to which it has to be compared (n=2).
5) Remove the n-grams that are replicated.
6) Sort the unique n-grams alphabetically.
7) Calculate the unique n-grams that the two words share.
8) Compute the similarity coefficient.
9) Repeat the above steps for other words in the dictionary.
10) Return the words with high values of similarity coefficient.
11) End.

VI. EXPERIMENTAL RESULTS

The goal of experimental analysis is to calculate the accuracy of the proposed system can achieve. Parameters that can be used for in this proposed model is comparison performance and correct suggestion produced by it. The detailed description of dataset used for experiment and analysis of experiment results and their discussions are under here. A test dataset is considered for evaluating the proposed method. The test dataset is a dictionary contained 10000 words which are collected from Malayalam corpus. The experiment analyzed over datasets in order to evaluate the performance of proposed algorithm. The accuracy is a parameter; it is used to evaluate the efficiency of the proposed N-gram analysis algorithm over a rule based Malayalam spell checker. The accuracy is defined in this proposed based on number of words stemmed correctly.

In the proposed system the number of false positives and false negatives may occur in both detection and correction phase.

VII. CONCLUSION

In this spellchecker, so far the error detection of words is almost complete for any form of agglutination that can come in Malayalam vocabulary. Only standard words are checked in this version and we have further plans to cater the non-standard words too. As far as suggestion generation is considered, only single character errors are taken into account and advanced versions will be taking care of multiple character errors also. Algorithms for correcting the multiple character errors, which can happen when multiple agglutination occurs is under development. The dictionary also can be improved, which will obviously improve spell checker, by adding the commonly used English words in a Malayalam conversation or in the language. Right now these cases are not handled. Facility to add a temporary database, for adding commonly used proper nouns by a particular user also will be added to the system in the next versions.

REFERENCES