Volume 5, Issue 3, March 2015 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Spell Checker for Non Word Error Detection: Survey

Hema P. H.*, Sunitha C
Computer Science and Engineering
Vidya Academy of Science and Technology
Thalakkotukara, Kerala, India

Abstract— Spell checker is a software tool which is used to detect the spelling errors in a text document. A spell checker can also provide suggestions to correct the misspellings. The error can be either non word error or real word error. Detecting real word error is really difficult task and requires advanced statistical and Natural Language Processing (NLP) techniques. Currently we have many methods for detecting non word errors in a text document. This paper mainly deals with the non word error detection methods for various languages.

Keywords— error detection, error correction, non word error, real word error, N gram analysis

I. INTRODUCTION

Spell checking applications are important part of several fundamental applications such as word processor, Electronic dictionary, editors and search engines. It is a software tool which is used to detect and correct the misspelled words in a text document. Misspelled word can be a word that exists in the existing dictionary that is not correctly spelled or in shortened form. A Spell checker consists of two main components - error detection and suggestions prediction. Error detection component identifies the errors. After detecting the error, the suggestions are provided by the suggestion prediction component that are closer to the misspelled word. The first and far most requirement for any spell checker is a dictionary of different possible words of that language which will act as a corpus. Basically, if the dictionary of a spell-checker is big then higher will be the error detection rate, otherwise misspellings might pass undetected. Mainly the errors related to misspelled words can be categorized into two basic classes: 1) Non word errors, where the misspelled word is not a valid word in the language.2) Real word errors, where the word in question is correct yet inappropriate in that scenario.

Detecting Real word errors is very difficult and requires advanced statistical and Natural Language Processing (NLP) techniques. Human typing errors leading to non word errors can arise due to three major facts; typographic errors, are related to keyboard mis-punches(e.g. the–th). These errors are happening when correct spelling known but typed incorrectly. Mainly the typographic errors fall into one of the following categories: 1)Substitution Error, It happened due to single letter is substituted by other letter.2) Deletion Error, When at least one character is deleted in a word. 3)Insertion Error, due to the insertion of an extra character in to a word. 4)Transposition Error, When two characters in a word is transposed. 5) Run-on error, When a space is missing between two words. 6) Split word Error, it is just opposite to Run-on error. It happened due to an extra space between two words. cognitive errors are caused by the writers mis conceptions(e.g. Receive-recieve) ; phonetic errors are a result of substituting a phonetically equivalent sequence of letters(e.g. separate–seperate).

The first issue in developing spell checker is the error pattern developed by text generating media such as type writer, computer keyboard, Handwriting, Machine Printing, OCR system etc. The error pattern in one media will be different from the other. The error pattern can be formed due to insertion,substitution, Transposition error, deletion, single versus multiple character error, word length effect, positional bias, run on and split word error, character shape effect, heuristic tendencies phonetic similarity effect etc.

II. ERROR DETECTION

Error detection is the important task of a spell checker. The most common technique for the non word error detection is Dictionary-lookups and N-gram analysis [2]. Many works has been done for the error detection and correction for English and related languages.

- ***Dictionary Lookups***

Dictionary is a list of unique words. The non-word errors can be detected by checking each word against a dictionary. A purely dictionary based approach is not practical in the case of any language. It just matches the word in the given text with the words in the dictionary, if the word is not in there in the dictionary; even if it is a correct word the spell checker will detect it as an error. If the dictionary is too small then it will give many false rejections, if too large it can accept a high number of valid low-frequency words. There comes the difficulties of keeping such a dictionary up to date, and sufficiently extensive to cover all the words in a text. We can add many techniques along with the dictionary for better performance.

© 2015, IJARCSSE All Rights Reserved

A Rule cum dictionary based method can give better performance than purely dictionary based approach. In the dictionary cum rule based approach it generates a corpus which lists the words under suffixes, route words and post positions. By incorporating a morphological analyzer the spell checker identifies the words comes under the lists from the raw text and compares it with the corpora. If the word is not there in the list the spell checker will detect it as an error and it will provide a list of words as suggestion.

- ***N gram Analysis***

N-Grams are n letter sub sequences of words or strings where n usually is one two or three. N-gram with one letter is called unigram or monogram. Two letter n-grams are bigram and three letter n-grams are called trigram. N-gram method is an easy approach to find the correct spelling in a text document. 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 ngram frequencies are stored. In the n-gram method there will be a pre compiled word unigram and syllable bigram, trigram frequencies. The spell checking based on n-gram statistics is relatively inexpensive to construct without deep linguistic knowledge. Any person can use the dictionary without deep linguistic knowledge. And also many letter errors can be identified by using this n-gram method.

III. ERROR CORRECTION

Correcting spelling errors in a text document is an important problem. The error correction mainly include two steps, they are generating candidate words and ranking the candidate words. The spelling correction mainly focuses on the isolated words, and will not consider the context in which the word appears. Today we are having the following types of error correction algorithms; Minimum edit distance method, Soundex algorithm, Rule based Technique, N gram based technique, Probablistic method, Neural net techniques

- ***Minimum Edit Distance Method***

Minimum edit distance is the number of operations required to transform one text to other. The operations can be insertion, deletion, substitution etc. Most of the edit distance algorithm gives the integer distance scores. Edit distance is useful for correcting errors formed from keyboard input. It is not preferable for correcting phonetic errors.

- ***N gram based Technique***

N-Grams are n letter sub sequences of words or strings where n usually is one two or three. This can be use for error correction in text document. N gram can be used along with dictionary or without a dictionary. Without dictionary N gram can be used to find the position where the misspelled word occurs. 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.

- ***Rule Based Technique***

The method will have a list of rules from common spelling error patterns, and it will be applied on the misspelled words, and will generate a list of candidate words for the error correction.

- ***Soundex Algorithm***

The soundex system is mainly used in the phonetic spelling correcting situations. The method in which a soundex code will be generated for all words in the dictionary. And the algorithm will generate a soundex code for a misspelled word and will retrieve words from the dictionary which having the same code.

- ***Neural net Techniques***

Neural networks is an important technique. 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.

- ***Probablistic method***

It is a simple method in which two common methods are used they are transition probabilities and confusion probabilities. Transition probability depends on the language. It gives the probability of a letter followed by another letter. The probability can be estimated by calculating n gram frequency for the words from a corpus. Confusion probability gives the value of how often a letter is mistaken or substituted for another letter. They are source dependent.

IV. AVAILABLE SPELL CHECKERS

Nowadays, spell checkers are an important component of a number of computer software such as web browsers, text processors and others. Today we have many spell checkers for detecting non word errors in text document. Different methods are used for various languages. The following part describes the various spell checking techniques developed for different languages.

© 2015, IJARCSSE All Rights Reserved

- ***A non word error spell checker for indonesian using morphologically analyzer and hmm*** **[3]**

This spell checker is built by using morphological analyzer and dictionary lookup as error detection method with two alternative optimization, binary search and hash. And for error correction, two alternative methods, namely forward reversed dictionary and probability of similarity is used. Forward reversed dictionary corrects the misspelled word by considering edit distance between the misspelled word and its candidates. Probability of similarity, which is the main proposed method for error correction, correct the mis spelled word by calculating its similarity to a candidate word, based on the value of optimum subsequence between them. Candidate sorting was accomplished through the use of HMM (Hidden Markov Model), where the word is considered as observed state and the candidates as hidden state. By using HMM, the system does not only consider the similarity of the candidate word with misspelled words, but also consider the sequence of words in sentences where the word is located. The experiment result proves that sorting candidates by using HMM increase the precision accuracy. As for correction method, the result showed that using probability of similarity has better correctness accuracy than forward reversed dictionary.

- ***Spell checking in assamese***

The non word error in Assamese are detected by looking up document words in a dictionary of valid words. A hash table has been used as a lexical lookup datastructure. Three methods are used for generating suggestions compraising of the soundex, edit distance and morphological processing. This method maps every words in to a key so that similarly spelled words have similar keys.

- ***Hindi spell checker [4]***

The spell checker uses a dictionary with word, frequency pairs as language model. The error word is detected by matching the word with the dictionary. And the suggestions provided by calculating strings at edit distance one and two from the identified erroneous string and further filter out those strings that are not present in the dictionary.The edit distance is found out by using Damerau-Levenshtein edit distance method.

- ***Kannada spell checker for non word error detection using morphological analyzer& dictionary lookup method*** **[5]**

It is the spell checker designed and implemented to detect the non word error in kannada language. The spell checker is developed by using morphological analyzer and dictionary lookup method. The tool handles all the text in Unicode format. User can input a block of text either manually or through a file. The block of text is then split into individual words and then each of these words is analyzed using a morphological analyzer, which makes use of separate root and suffix dictionaries. A one to one correspondence has been established between different categories of root and suffixes with the help of a mapping function. Morphological analysis helps in detecting whether word is spelled correctly or not. Further the erroneous words are displayed to the user with corresponding suggestions. The misspelled words can be corrected as per the suggestions selected by the user. As Kannada is a morphologically rich language, each root word can combine with multiple morphemes to generate huge number of word forms. A complete spell checker for Kannada language has not been developed yet. In the above method Multiple suggestions for a single misspelled word is provided with a combo box in which user can make a choice out of the listed possible words. To improve the performance of the tool multi-threaded approach is used.

- ***Automatic spelling correction tool to improve retrieval effectiveness based on Revised n-gram method [7]***

It is a language-independent spell-checker based on an enhancement of the n-gram model. It is an approach to detect the non word error in a text document. The spell checker works on basis of a multi spell algorithm. According to the algorithm it compares the keywords given from the user with the correct words contained in the dictionary. If the word detected as misspelled then the algorithm builds n-grams for the misspelled word. Then we select correction candidates from the dictionary. For the selected words the n-grams are computed and the similarity score is computed. The correction with a combo box in which user can make a choice out of the listed possible words. To improve the performance of the tool multi-threaded approach is used.

V. CONCLUSION

Spell checker is the important tool used for error detection and correction in a text document. Today we have many spell checkers for difeerent languages. This paper mainly discusses the non word error detection and correction methods used for spell checking application.

REFERENCES

[1] Neha Gupta, Pratistha Mathur ” Spell Checking Techniques in NLP: A Survey” , In International Journal of Advanced Research in Computer Science and Software Engineering 2 (12), December - 2012, pp. 217-22.S.

[2] T. Santhosh, K. G. Varghese, R. Sulochana, and R. Kumar,” Malayalam Spell Checker ” , in Proceedings of the International Conference on Universal Knowledge and Language - 2002, Goa, India, 2002.

[3] Soleh, M.Y. ; Dept. of Inf., Bandung Inst. of Technol., Bandung, Indonesia ; Purwarianti, A., ” A non word error spell checker for Indonesian using morphologically analyzer and HMM”, Electrical Engineering and Informatics (ICEEI), 2011 International Conference on 17-19 July 2011.

© 2015, IJARCSSE All Rights Reserved

[4] mit Sharma (10080) and Pulkit Jain (10543) CS365 Course Project under Dr. Amitabha Mukherjee "Spell Checker for Hindi".

[5] Rajashekara Murthy S, Vadiraj Madi2 , Sachin D, Ramakanth Kumar P, "a non-word kannada spell checker using Morphological analyzer and dictionary lookup Method", International Journal of Engineering Sciences & Emerging Technologies, June 2012. ISSN: 2231 – 6604

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

[7] Farag Ahmed, Ernesto William De Luca, and Andreas Nrnberger, "Revised N-Gram based Automatic Spelling Correction Tool to Improve Retrieval Effectiveness ", Polibits (40) 2009.

© 2015, IJARCSSE All Rights Reserved