International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

Spelling and Grammar Checker and Template Suggestion

Adnan Shaikh[1],Gaurav Sarade[2], Nuwed Munshi[3] ,Rutvik Natu[4],Anand Dhawale[5]

[1] *Student, Dept. of Computer Engineering, MESCOE, Pune, India*
[2] *Student, Dept. of Computer Engineering, MESCOE, Pune, India*
[3] *Student, Dept. of Computer Engineering, MESCOE, Pune, India*
[4] *Student, Dept. of Computer Engineering, MESCOE, Pune, India*
[5]*Professor, Dept. of Computer Engineering, MESCOE, Pune,* India

---

Abstract - *In today's world, even a common man who is not well versed with the English language can use the computers or mobile phones easily. This is mainly because of two reasons, first is an easy user interface and the second one is auto grammar and spelling checker. There are a lot of options are available for the betterment of the user interfaces in mobile apps and other applications including voice controls. Whereas for spelling and grammar checking process still lot needs to achieve to provide more accurate and fast suggestions. Many methodologies are using the dictionaries to achieve the same and some methodologies are using learning techniques. Very few techniques are using the both methodologies to enhance the process of spelling and grammar checking process. So proposed methodology is using the technique of the decision tree and distance analysis technique to achieve the spelling and grammar process.*

***Key Words*: Levenshtein Distance, Spelling checking, Decision making, Dictionary.**

1. INTRODUCTION

There has been increased interest in researching and developing techniques for spelling and grammar checking due to the fact that English language rules and grammar is extremely difficult to ascertain in certain situations, even for seasoned writers. Most of the word processors have an inbuilt spelling checker which is good for basic editing. The spelling and grammar checkers that come as default are not as powerful and can only manage to detect minor mistakes.

To provide error correction in the English language, the words that are written are compared with a lexicon or a dictionary for finding mistakes in the spelling of certain words. This technique does not take into account the context of the word and its placement, but only the spelling. The words which are not present in the dictionary, they will not be processed or identified. Most of the errors that occur for the majority of the English writers is not the spelling mistake, which is present but there are very few and have more of grammatical errors also called real-word errors.

The real word errors can be defined with the help of the same dictionary-based spelling checking systems as the word which is present in the dictionary and has been spelled correctly but is not supposed to be used there in the sentence and in that form. This is a real word error and this cannot be detected by most of the spelling and grammar checkers. To identify such errors, the spelling checker has to understand the context of the sentence and intelligently evaluate the errors.

Spell Checking is one of the most fundamental concepts of Natural Language Processing. Spell checking can also be used for proofreading, information retrieval, etc. Natural Language Processing has been widely used in various applications, especially useful for checking spellings as it has the ability to breakdown the text which makes it very convenient for identifying the various inconsistencies and help correct them. As it is imperative for any writer to be able to convey his thoughts coherently with proper grammar and spelling.

There should always be a minimum amount of errors in a manuscript which entails a better understanding of the subject at hand. Reducing the grammatical and spelling errors have also had the added benefit of imparting the correct meaning and

knowledge to the reader and help reduce misunderstandings and misinformation being spread around. Therefore, it is a compulsion for a writer to have flawless grammar and spelling which is not usually possible as we are humans after and tend to make mistakes, therefore a technique that can determine various irregularities in the text is highly beneficial for writers and readers alike.

It has been used extensively in applications in correcting spelling issues. This technique is highly useful to determine the differences between two sets of sequenced strings. The technique utilized to achieve it compares the two strings and outputs the minimum number of edits that should be done to achieve the correct version in the shortest time. This is highly useful for word processors that have a built-in spelling checker, which will identify the wrong spelling and appropriately suggest the edit.

This technique was developed in the late 1960s and was named after its author Vladimir Levenshtein. He provided the equation that achieves the matching between the two words and also determines the dissimilarities between the two sequences. This is done to extract the number of single-character edits, which can be anything ranging from substitutions, deletions or even insertions that are needed to transform or correct the target string into the string provided as a matching reference. This technique is highly useful in achieving very accurate and fast results that can effectively provide valuable edits and a great addition to the spell checking paradigm.

This research paper dedicates section 2 for analysis of past work as literature survey, section 3 deeply elaborates the proposed technique and whereas section 4 evaluates the performance of the system and finally section 5 concludes the paper with traces of future enhancement.

2. LITERATURE SURVEY

This section of the literature survey eventually reveals some facts based on thoughtful analysis of many authors work as follows.

Shashank Singh focuses on the challenges faces by the researchers while handling the errorsin the realworld. Spellcheckers are one of the integral parts of the word processors that highlight the wrong spelling to correct it with the right word. [1] In recent years there have been many applications developed for spell checking and to check the framing of the sentences. This paper detects and corrects the words that are wrongly spelled and also analyzes the method which is used worldwide. It also faces real word error problems faced by the researchers.

Leena J. Alhussaini elaborates the human languages into three parts there is language,then the second one is meaning and the third part is context. The language is studied in the form of the words which is purely defined by building blocks of affixes, suffixes, roots, etc. The accurate purpose of the communication is the correction of the spelling to be delivered [2] the proposed system contains the basic parts where one is an input system which is the decision-making system which manages the words and behavior system which is a dynamic system.This dynamic system is used to manage the output of the ranked suggestion list for user misspelled data.

Eranga Jayalatharachchl explains spell checking as identifying the errors in language and indicates the incorrect words in the document written in the natural language. For spelling checking the most effective technique is detecting the string and finding those words in the dictionary or in the wordlist. Thus, correcting the spelling is the second task which isan important and challenging part due to the fact that the system has to locate and suggest the correct spelling.[3] This research is to implement quality of the subhasa based data on the spell checker used in the technique called as minimum edit distance technique, the system is available on the internet.

S.Hossein introduces the content-based method in the real word spell checking. These days due to an increase in extensive text editing software, spell-checkers are one of the most widely used tools within natural language processing. [4] Recently, the Microsoft word has been regarded as the most influential editing software in spell checkers. But spellchecker in the processing software is a very challenging part and it also faces a problem in the Persian language. So, in the proposed paper, the researcher tried to cover up the problem by using the content-based method.

S.Singh explains spell checker is a software or program or idea which is designed to identify misspellings in a text while either automatically corrects them or suggest different spelling options. Spell Checking and Grammar Checking are a very important part of the process of writing.[5] English is a language that is spoken by around 420 million people on this earth and understanding it is not at all easy. The meaning of the sentence is changed as the tone of the speaker. Thus, the

paper presents the English language processing system for Spell Checker and a Grammar Checker.

B.QasemiZadeh[6] elaborates adaptive, language independent, and 'built -in error pattern free' spell checker. Ternary Search Tree data structure is used with a nondeterministic traverse is used to suggest misspelled words in the proposed paper. New language is equal to adding a lexicon and a COT matrix so the proposed system is easy to implement. The method proposed by the research in the present paper is one of the flexible methods and the accuracy rate of the paper is high in comparison to other proposed methods.

I.Zhuang describes the new technique called OCR i.e. Optical Character Recognition is a process where the system observes the character images to automatically translate the text informationOCR is nothing but the ICR(Independent Character Recognition) analyzes characters in every position of an image and give possible results. Thus, the OCR is of the most successful and the effective spelling approach and multi-knowledge based statistical language model [7] The proposed method as the high accuracy rate of 79.3 to 91.9%.

M.Kim proposes a statistical and context-sensitive spelling correction approach using the confusion sets to correct context-sensitive spelling errors that occur due to typographical errors.The confusion set is one of the advances which can correct the context-sensitive spelling errors.[8] Result of applying the proposed approach on all 5 confusion sets showed higher precision and recall than the baseline. The study includes diverse statistical techniques to increase recall while maintaining precision at 100%.

Z.Wint explains that in the field of word errors there are two types of errors such word errors and non-word errors. Word errors are the words that can be found in the dictionary and the same words are misspelled by the user. Non-word errors are the words that cannot be found in the dictionary these are the words where the users have accidentally pressed the wrongkey.[9] The proposed paper works on the spell checker technologyand corrector to check word errors in the social media datasets, which will be used in message filtering systems.

M.Go elaborates on the implementation of Gramatika, Grammar checkers added inside word processing software such as Microsoft Word and Google Docs. These grammar checkers are able to identify run-on sentences, misspellings, disagreements between subjects, and other error types that need to be corrected. [10]The accuracy rate of this grammar checker is 64% by giving the correct suggestions. And the final result of the proposed paper is as high accuracy rate of error detection as well as correction and giving suggestion.

N.Ehsan explains that the Grammar checking techniques can be divided into different types: syntax-based, statistical and rule-based.[11] In the syntax-based approach, a text is completely parsed and if the parsing is not completed the text is considered incorrect. The second is a statistical approach, a part-of-speech annotated corpus is used to build a list of POS tag sequences; and the last is rule-based approaches, a set of rules are matched against a text which has at least been POS tagged and the rules are developed manually. Thus, the accuracy rate of the proposed paper is between 70% and 83%.

P.Jin elaborates that spelling errors are very common in the texts and these errors will confuse the readers and decrease the quality of the documents. In Government documents, the spelling errors are not tolerated.[12] Automatically finding these errors is significant to many organizations such as press and government and therefore, it has a long research history in the natural language processing field. The author suggests that they will use a tri-gram and a higher-order language model to improve performance.

K.Shaalan [13] describes that words are the main language element used for communicating within a language. Spellcheckers are mostly used in many software products for identifying errors in users' writings. Spelling error detection is concerned with identifying a word as an incorrect word. Methods used for spell checking in many applications include dictionary lookup techniques, which are commonly used where words are compared and located in a language dictionary. The proposed method was evaluatedextensively and promising results were achieved.

3. PROPOSED SYSTEM

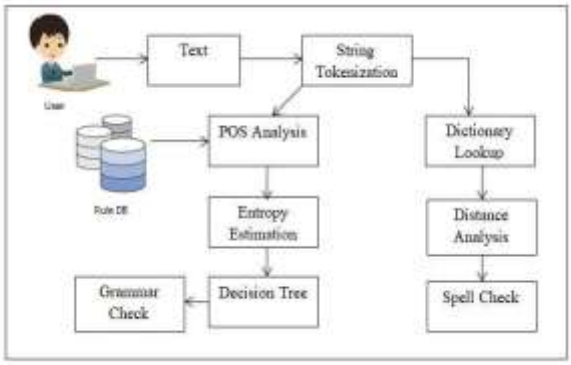

Fig-1: System Overview of Proposed System

The proposed model of spelling and grammar checking is depicted in the figure 1. The model is explained with the below mentioned steps.

Step 1: Dictionary Lookup- This is the initial step of the proposed model, where around 10,000 words of the English language are stored in a workbook sheet is fed to the system.

The stored words are read and uploaded using JXL external API into a static list. That can be used during the operation whenever it is required.

Step 2: Spell check through Distance Evaluation - To check the spelling of the given text, Initially the text is split on space character to tokenize all the string into an array of words. Then these each of the words are considered to compare with the other words of the Dictionary for their respective comparisons. To perform this operation of the spelling checking Levenshtein Distance is used. This Levenshtein distance is evaluated in between the input word and the dictionary words. If the yielded distance is 0 then it means the input word is present in the dictionary and there is no error in the word.

If the Levenshtein distance is other than the 0 that means the input word is having some spelling error. So the proposed model evaluates the distance of a word with all other words of the dictionary, then the word with the smallest distance is considered as the replacement word or correct word. This word is being replaced with the input word to correct the spelling.The pseudo code of this Levenshtein distance is depicted in the below mentioned pseudo code 1.

PSEUDO CODE 1: LEVENSHTEIN DISTANCE

Step 0: Start
Step 1: Set n as the length of the first word S_W
Step 2: Set m as the length of the second word S_T
Step 3: If n=0, then return m and exit
Step 4: If m=0, then return n and exit
Step 5: Initialize the 1st row from 0 to n
Step 6: Initialize the 1st column from 0 to m
Step 7: Examine each character of S_W(i from 1 to n)
Step 8: Examine each character of S_T (j from 1 to m)
Step 9: If $S_W[i] = S_T[j]$, Set cost as 0
Step 10: If $S_W[i] \neq S_T[j]$, Set cost as 1
Step 11: Set cell d[i,j] of the matrix equal to the minimum of:
  a. The cell immediately above plus 1: d[i-1,j] + 1.
  b. The cell immediately to the left plus 1: d[i,j-1] + 1.
  c. The cell diagonally above and to the left plus the cost: d[i-1,j-1] + cost.
Step 12: On finishing the iteration steps (3, 4, 5, 6) are complete, the distance is found in cell d[n,m].
Step 13: Return the distance
Step 14: Stop

Step 3: parts of speech Rules for grammar correction- Once all the spelling is checked, then the words of the input text are stored in an array to perform some grammar corrections as mentioned below.

Rule for *' Have '* - If the word next to ' have' is not from this list { been,a,an, to, the, no}, then that word should be converted into its past tense form.

Rule for 'we' - If the word next to ' we' is at present continues, then the word should be ' we are '.

Rule for 'they'- If the word next to 'they' is at present continues, then the word should be ' they are'.

Rule for ' I '- If the word next to 'I" is at present continues, then the word should be 'I am '.

Rule for ' it '- If the word next to 'it ' is at present continues, then the word should be ' it is'.

Rule for 'many'- If the word next to next of 'many' is equal to 'are' and if the word next to ' many' is in its singular form, then convert it into plural form.

Rule for 'every' - If the word next to ' every' is a plural then convert it into singular. And then the word next to the converted singular word should be 'is'. For example - ' Every students are going to trip' should be ' Every student is going to trip'.

Step 4: Entropy and Decision Tree Rules - Here the rules are defined to correct the grammar based on the distribution factors of the word in a sentence. These rules can be defined as below.

Vowel Rules- For this rule first each every word in the list is checked for its beginning with the vowels, If it is then the previous word should be 'a'. Then that a is converted into 'an'.

Rule for 'then' - If the word before ' then' is except 'and' or 'is', then the past word should be appended with a comma character (Example - ,then).

Rule for Sentence Formation - Here the sentence is corrected by having an upper case character at its beginning. And then every sentence should start with a Capital letter by leaving a space for the past sentence's termination.

After following all the rules the given text is free from the spelling and grammar for the defined constraints.

Step 5: Template access - The proposed model also provides some templates for the keywords by evaluating the matched count for the keywords. This is done by analyzing the match count with respect to the content of the template file names.

International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 06 | June 2019 www.irjet.net p-ISSN: 2395-0072

Table-1: Mean Square Error

Experiment Number	Number of Words	Expected Grammatical Mistakes (xi)	Corrected Grammatical Mistakes (yi)	Difference (xi-yi)	Mean Square Error
1	174	12	10	2	4
2	199	9	9	0	0
3	277	15	13	2	4
4	387	17	16	1	1
5	210	21	17	4	16
6	256	26	23	3	9
7	388	13	13	0	0
8	345	11	9	2	4
9	469	16	15	1	1
10	512	19	17	2	4
				Average MSE	3.9

The proposed model includes some templates under the file name like Balance cover, Business apology, Formal Complaint, Formal Invitation, Formal Job application, Formal Resignation, Leave Letter, Letter of intent for job, Scholarship letter and finally Warning letter for negligence in duty.

3. RESULT AND DISCUSSIONS

The presented technique for the Spell and Grammar checking paradigm has been implemented successfully on a machine running on aWindows operating system. The system is powered by a Core i5 Central Processing Unit with 6GB of physical memory. The proposed methodology has been coded in Java programming language on an Integrated Development Environment called NetBeans. The technique has been tested extensively to measure its performance and accuracy.

The effectiveness of the proposed system has been calculated using the versatile Root Mean Square Error technique. Root Mean Square Error has the capabilities to ascertain the variability of two correlated continuous entities. The central idea of the Root Mean Square Error is that it utilizes standard deviation and extracts residuals. These residuals are nothing but the errors, as they determine the distance between the residuals and the line of

$$\mathbf{RMSE}_{fo} = [\sum_{i=1}^{N} (z_{f_i} - z_{o_i})^2 / N]^{1/2}$$

regression. The RMSE is evaluated from these residuals which elaborate on how distributed the residuals are within each other.

Therefore, the two entities in this technique are the actual number of expected spell and grammar mistakes and the other entity is the number of accurate suggestions/corrections for which the Root Mean Square Error is being calculated. The Equation given below shows the process in much detail.

Where,

$\sum$ - Summation

$(Zfi - Z_{oi})_2$ - Differences Squared for Expected Grammar Mistakes and Corrected GrammarMistakes.

N - Number of samples or Trails

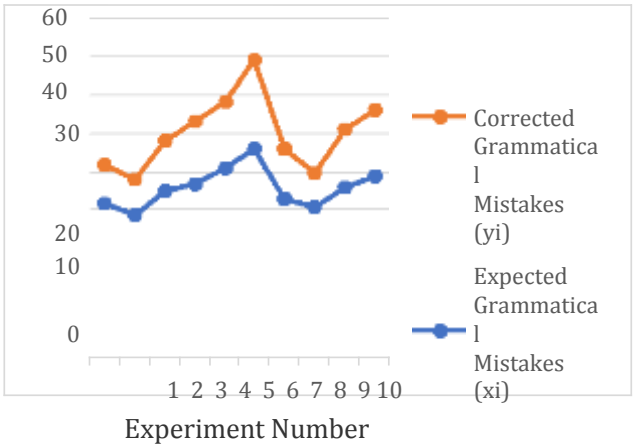

Fig-2: The plotted difference between Expected no. of Grammatical Mistakes v/s Corrected Grammatical Mistakes

The results have been tabulated above indicate that the proposed system achieves the Mean Square Error as 3.9. This value of Mean Square Error has been utilized further to generate the Root Mean Square Error value which is 1.97 that is obtained from the 10 experiments conducted. Thus, the result concludes that the methodology discussed in this paper is highly accurate and effective.

4. CONCLUSION AND FUTURE SCOPE

The proposed model of spelling and grammar correction is designed on the basis of minimum constraints. The proposed model uses around 10000 words of the English language for the purpose of spelling checking. Levenshtein distance is used to measure the nearest word for the spelling correction. Grammar correction is done on some of the protocols as discussed in the proposed methodology section using the parts of speech and distribution rules. Decision tree plays a vital role in deciding the proper grammar for the given textual content. The proposed model is evaluated for the error rate precision using the Root mean square error. The experiment yields around 1.97 of RMSE which is quite low error rate and indicates the effectiveness of the proposed model.

In the future the process of spelling and grammar correction can be improved by using the huge vocabulary set and the Protocol set via Abstract syntax tree in the distributed web environment.

REFERENCES

[1] Shashank Singh Shailendra Singh "Review of Real-World Error Detection and Correction Methods in Text Documents", Proceedings of the 2nd International Conference on Electronics, Communication, and Aerospace Technology,2018

[2] Leena J. Alhussaini, "Application of Component Engineering to the Design of Holistic Spell-Checking Algorithm", Blueprint for a New Computing Infrastructure, Morgan Kaufmann, second edition, 2003.

[3] Eranga Jayalatharachchl, Asan a Wasa a, Ruvan eerasmg, "Data-Driven Spell Checking: The Synergy of two algorithms for Spelling Error Detection and Correction" The International Conference on Advances in ICT for Emerging Regions 2012.

[4] Samani, Mohammad Hossein, Rahimi, Zeinab, Rahimi, Sara, "A Content-based Method for Persian Real-Word Spell Checking", IKT2015 7th International Conference on Information and Knowledge Technology, 2013.

[5] Shashi Pal Singh, Ajai Kumar, Lenali Singh,Mahesh Bhargava, Kritika Goyal, Bhanu Sharma, "Frequency-basd Spell Checking and Rule-based Grammar Checking" International Conference on Electrical, Electronics and Optimization Techniques(ICEEOT)-2016.

[6] Behrang Qasemi Zadeh, Ali Ilkhani, Amir Ganjeii, "Adaptive Language Independent Spell Checking Intelligent Traverse on a Tree", IEEE Conference on Cybernetics and Intelligent Systems,2006.

[7] Li Zuang, Ta Bao, Xiaoyan Zhu, Chunheng Wang, Satoshi NAoi, "A Chinese OCR Spelling Chevk Approach Based on Statistical Language Models" IEEE International Conference on Systems, Man and Cybernetics 2004.

[8] Minho Kim, Jingzhi Jin, Hyuk-Chul Kwon, Aesun Yoon, "Statistical Context-sensitive Spelling Correction using Typing Error Rate ", IEEE 16th International Conference on Computational Science and Engineering 2013.

[9] Zar Wint, Th'eo Ducros, Masayoshi Aritsugi, "Spell Corrector to Social Media Datasets in Message Datasets in Message Filtering Systems", The Twelfth International Conference on Digital Information Management,2017.

[10] Matthew Phillip Go, Nico Nocon and Allan Borra, "Gramatika: A Grammar Checker for the Low-Resourced Filipino Language",Proc. Of the 2017 IEEE Region 10 Conference(TENCON), Malaysia, November 5-8,2017

[11] Nava Ehsan, Heshaam Faili, "Towards Grammar Checker Development for the Persian Language", Towards Grammar Checker Development for the Persian Language,2010

[12] Peng Jin, Xingyuan Chen, Zhaoyi Guo, Pengyuan Liu, "Integrating Pinyin to Improve Spelling Erraors Detection for the Chinese Language",IEE/WIC/ACM International Joint Conference Web Intelligence(WI) and Intelligent Agent Technologies.

[13] Khaled Shaalan, Ran Aref, Aly Fahmy,"Approach for Analysing and Correcting Spelling Errors for Non-native Arabic learners" Umm AL-Qura University, 2009.