International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 12 | Dec -2016 www.irjet.net p-ISSN: 2395-0072

Content Improvisation by Spell Checking, Grammar Checking, Tone Checking and Scoring

Soumya Singh*1, Sayali Nalawade*2, Saket Gonte*3, Mohd. Yusuf Quadri*4, Sachin Godse*5
Department of Computer Engg.
STES's Sinhgad Academy of Engineering, Kondhwa (Bk),
Pune, Maharashtra, India.

Abstract—Language is one of the medium for communication of thoughts and expressing our views. Among many languages which are spoken all over the world English covers the major area of the language landscape and it is accepted globally as lingua franca or the common language. So, it is important to write correctly.

We all know that the impact of the word on a person's mind is bigger than any actions performed. An incorrect content can leave a bad impression on the reader and can lack in communication of thoughts which the writer actually wanted to convey. A document can be incorrect if there are spelling mistake or there are grammatical errors or the tone used to write the document is not appropriate. Spelling mistakes can embarrass the person who is writing, grammatical mistake can change the entire meaning of the sentence whereas if the tone of the document goes wrong then it can not only change the perception of the reader but can also annoy him to a great extent. To combat these issues we are creating a Word plugIn, which will offer the writer to correct his/her document by detecting the spelling and grammatical mistakes and providing suggestions as well as correction for the same. We have added two extra functionalities to improve the content in all aspects, these are- scoring and tone checking. The main goal of this system is to make the work of the person smooth where they don't have to spend time in checking their document before sending or using. Manually checking the huge document is not only time consuming but a tedious job too and even after manually checking the entire document there can be many hidden errors.

Key words: **Grammar checking, JSON, N-Gram, POS-tagging, Rule-based, Scoring, Spell checking, Tone checking and Typographic errors.**

I. INTRODUCTION

English is not that easy language as it seems. There can be basically two types of error which people usually do while writing. First is the spelling mistake and the second one is the grammatical mistakes. Spelling mistakes are when the written word do not belong to the vocabulary and grammatical mistakes are those sentences which do not follow the rules stated by the language's grammar.

For improving the writing style of the user we have added two more features which are tone checking and scoring. Tone checking checks the tone of the content and inform the user about that. Tone of the content can be technical, general, happy, aggressive, etc. We included this feature as no one likes to read an aggressive document and a bad toned document can convey a wrong meaning and can lead to miscommunication.

Scoring is a feature which will score the content of the user according to mistakes. It will not only tell the amount of mistakes of the user but also encourage them to improve it which will improve their writing also.

II. APPROACH AND PROPOSED SYSTEN

No one is perfect, making errors while writing or typing is very common, but we cannot let these small mistakes cause us a huge loss. We are trying to make people's work a bit easier. People can enjoy hassle free writing and improve it at the same time. At the end English is a means of communication and it should convey the exact meaning or the thought of the writer.

A. Spell check

It is a software or a piece of code which will detect the spelling mistakes in the document. Spelling mistakes can be very embarrassing sometimes. Imagine making a spelling mistake while printing a flex where every letter costs and everyone sees it. The embarrassment, pain and the real sense of spelling mistake can be seen if someone got a tattoo with a wrong spelling.

The basic task in spell checking are:

1.Detection of the errors
2.Predict suggestions and correction of the errors.

A.1. Detection of error: Type of Error

Techniques are used or designed on the basis of spelling errors trends which are also called error patterns. Errors can be broadly classified into three types.

- Typographic errors

These errors occur when the correct spelling of the word is known by the user but the word is

mistyped. These errors do not follow any linguistic criteria as these are related to keyboards. The errors produced by this type are also called single-errors as it includes insertion of an extra letter or omission of a letter or writing other letter instead of the actual, e.g. writing 'a' instead of 'o' in 'occur' or switching two adjacent letters.

- Cognitive errors

These error occur when the correct spelling is not known, but the pronunciation of the misspelled word is almost near to the actual word. (E.g. receive -> receive).

- Slang word errors

In this social media loving era one basic type of spelling mistake which people do is using slang words or in technical term non word error.
Non-word error: These are the spelling errors which do not belong to the dictionary.[4]
These can be slang words used like gud, sry.
For example:
'I am sry.'
'I am sorry.'

A.2.1. Detection of error

In a sentence words are separated by space bar or punctuation marks and may be called a candidate words or tokens. A token or a candidate word is valid if it has a meaning otherwise it is a non-word. There are many techniques for error detection. The two most know and used are N-gram analysis and dictionary lookup. The error detection process checks whether the word is a dictionary word or not.

- Dictionary lookup

This is the most known and easy to implement method. In this each and every word is compared with the dictionary to find whether it is a misspelled word or not. In this method the cost of search is very high.[1][2]

- N-Gram Techniques

It is a process which detect wrongly spelled words. In this technique n-grams are used instead of comparing each and every word in a dictionary.

If n-gram is non-existent or a rare n-gram is found the word is marked as a misspelling, else not. A set of consecutive words taken from the sentence is called n-gram. N can be any number as 1, 2, 1-gram is called unigram, 2-gram is called bigram, etc.

P (A/B) = P (B/A) P (A) / P (B)
Where:
P (A/B): Probability of A given B
P (B/A): Probability of B given A
P (A): Total probability of A
P (B): Total probability of B

A.2.2. Suggestions and Corrections

- Edit Distance

It is a simple technique. It is based on the assumption that the person usually makes few errors in the spelling while typing the letters, therefore for each dictionary word the least number of the basic editing operations (insertion, deletions, substitutions) necessary to convert a dictionary word into a non-word ,the higher the probability that the user has made such errors. Keyboard input error, slang word errors can be solved using Edit Distance method.[2]

- Rule-based Techniques

The spelling is checked using the grammar of English language. In this we match the text with the given set of grammatical rules like tenses of the word. The major drawback of this method is that there can be some errors which are not detected by it, rest it has many advantages over other methods.[3]

- Frequency-based Suggestion

Suggestions are displayed using this method. It finds the distance between the non-word and dictionary word and rank each found word according to its usage frequency. [1]

B. Grammar check

It is a software or a program which checks the grammar of the document. Grammatical errors are introduced when we violates the rule of grammar of a given language. Poor grammar can annoy the reader too much.

Grammar is a vast domain. It is easier for human to understand it but making the system understand it is a very tough job. There are various techniques where the grammar is checked and the corrections are suggested.

Tense are the first thing which is checked while checking the grammar. There are three tenses in English language and each have four sub parts.

1. Present tense
2. Past tense
3. Future tense

The rules of the tenses are stored in a file, from there the rules are compared and checked if the tense is valid or not.

The basic grammar checking process includes POS Tagging of the words then parsing them and generating the parse tree, then checking the tense from the file where rules are given and at the end suggesting corrections.

B1. Rule Representation

The rules are written and stored in a file. Rules should be written in a way which is easy and quick to read. JSON file is the most common option to represent the rules. [1]

B2. POS Tagging

Part-of-speech tagging, it is also known as grammatical tagging. Through it we can assign part of speech tag to a word which tell the system the meaning of the word and reduces disambiguation of the words. [3] It is important to perform as there are many words in the English language which have more than one meaning. For example: 'bark', bark of a tree or a dog's bark.

Example- She is eating ice-cream.

Pos tag- She|PRP is|VBZ eating|VBG ice-cream|NN .|.

So the sentence is POS tagged using pos tagger.

B3. Parser

A sentence have a hidden level of hierarchy in it. Parser analyses that hierarchy and forms a parse tree after which it is easier for the system to understand the sentence on the basis of grammar.

Example- She is eating ice-cream.

Parse tree is given below.

```
(S
(NP (PRP She))
(VP (VBZ is)
(VP (VBG eating)
(NP (NN ice-cream))))
(. .)))
```

Once the tree is generated we can find the singulars and the plurals, now we can match the tense rules and check whether the given sentence is grammatically correct or not.

B4. Matching with the tense rules

In this we select the root from the file where rules are stored and start comparing each level of the node. If the match is found then we go to the next node. We repeat the process till we reach the leaf node. If leaf node is reached then the grammar is correct otherwise the input sentence is grammatically incorrect.

C. Tone check

Tone of the document is an important tool as few adjustments or changes in the tone can change the people's perception about the document and improve its effectiveness.

Emotional tones are computed using Tone analyzer, in addition to social and writing style tones. Tone analysis is less about analyzing the feeling of the person, and more about analyzing how you are coming across to others.

Tone checking can be divided into two major streams. Where one is analyzing current state of the person and the attitude of the writer towards the context and what kind of impact it can have on the readers. Second is classifying the document for some category of people depending upon the audience of the document, for example document written for technical group of people consists of all the minute details whereas document written for general people will consist of terms which are easier for the people to understand.

Tone can be checked at the 'Document level' for the overall detection of the tone of the document and tone can be checked at the 'Sentence level' which will identify sentence having stronger tone.

D. Scoring

Scoring is giving score on the document of the person on the basis of spell checking, grammar checking and tone analysis. A high score means the document is almost perfect in all aspects and is ready to deliver. It will not only tell the performance level of the person but it will also encourage them to improve their writing style and gain higher marks. It will act as a guide of the person. The person will be aware of their mistakes and learn from it to avoid committing the same mistake in the future.

III. CONCLUSION

We have suggested to develop an overall content improvisation system which checks the spelling, checks the grammar and also checks the tone of the document and after analysis the gives score. This includes almost everything what a content needs to be improvised.

IV. FUTURE SCOPE

The accuracy of the system can be further increased by using better techniques for analyzing Spellings, Grammar and by refining the existing ones. Grammar check is not only confined to tense checking, as told earlier grammar is a huge domain to work on. It consists of Passive voice and Active voice, Direct and Indirect Speech, etc. The system can be made more intelligent by adding a tutor function which will teach the person minute but significant rules of grammar which we tend to forget or get confuse with. There are a huge list of functionalities which can be added like segmentation, etc.

V. REFERENCES

[1]"Shashi Pal Singh, Ajai Kumar, Lenali Singh, Mahesh Bhargava, Kritika Goyal, Bhanu Sharma (2012). Frequency based Spell Checking and Rule based Grammar Checking: International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT)."

[2] "Pratistha Mathur, N. G. (2012). Spell Checking Techniques in NLP: A Survey. International Journal of Advanced Research in Computer Science and Software Engineering. ISSN: 2277 128X"

[3]"Naber, D. (2003). A Rule-Based Style and Grammar Checker"

[4] "Amit Sharma, P. J. (2013). Hindi Spell Checker"

[5] "https://tone-analyzer-demo.mybluemix.net/"