Text Optimization/Summarizer using Natural Language Processing

Mahesh Patil[1], Mayur Pawar[2], Yatin Rai[3], Prof. Satish Kuchiwale[4]

[1-3]Student, Computer Engineering, SIGCE, Navi Mumbai, Maharashtra, India
[4]Asst. Professor, Computer Engineering, Smt. Indira Gandhi College of Engineering, Navi Mumbai, Maharashtra

---***---

Abstract - *Computers have become a major medium for the exchange of information. Hence, we have the increasing need to do a thorough check of the information which the masses intend to exchange. The need for accurate grammar and spellings is not only a requirement for the formal business conversations, but also for the formal conversations taking place on the various platforms available for the same. The project aims at building an intelligent system to optimize English language. This system will perform functions like Auto Completion, Summarization, spell check, Grammar check.*

Key Words: Grammar, Business conversation, Optimize.

1. INTRODUCTION

Man's quest for making machines as smart as he is will go on forever. But the fact that we are able to design machines which react to human language by speaking like humans and passing the turing test with considerably fair amount of results is worth appreciating as far as the man-machine integration is concerned. The project aims at building intelligent system to optimize English language.

It will perform the following tasks:

1. Grammar Optimization
2. Spellcheck
3. Summarization
4. Sentence Auto Completion

Computers have become a major medium for the exchange of information. Hence, we have the increasing need to do a thorough check of the information which the masses intend to exchange. The need for accurate grammar and spellings is not only a requirement for the formal business conversations, but also for the formal conversations taking place on the various platforms available for the same. The project aims at building an intelligent system to optimize English language. This system will perform functions like Auto Completion, Summarization, Spell Check, Grammar Check

Let's define the job of a spell checker and an autocorrector. A word needs to be checked for spelling correctness and corrected if necessary, many a time in the *context* of the surrounding words. A spellchecker points to spelling errors and possibly suggests alternatives. An autocorrector usually goes a step further and automatically picks the most likely word. In case of the correct word already having been typed, the same is retained. So, in practice, an autocorrect is a bit more aggressive than a spellchecker, but this is more of an implementation detail — tools allow you to configure the behaviour. There is not much difference between the two in theory. So, the discussion in the rest of the blog post applies to both.

Grammar checkers accepts input in form of documents, paragraphs or sentence. However it then break down input into unit form, sentence. Corresponding language punctuation marks are used to identify completion of sentence. The sentence has to undergo some kind of preprocessing.

Text summarization, first we, have to know that what a summary is. A summary is a text that is produced from one or more texts, that conveys important information in the original text, and it is of a shorter form. The goal of automatic text summarization is presenting the source text into a shorter version with semantics. The most important advantage of using a summary is, it reduces the reading time. Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, concatenating them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then express those concepts in clear natural language. There are two different groups of text summarization: indicative and informative. Inductive summarization only represent the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. On the other hand, the informative summarization systems gives concise information of the main text .The length of informative summary is 20 to 30 percent of the main text paragraphs etc. from the original document and

1.1 OBJECTIVE:

This System aim to achieve the following through this project:

- Provide an intelligent and interactive system for interactive communications.
- This will be done by a high end and high computation processing system.
- Can be used on an individual level.
- Design a system with a highly usable UI.
- The system has an underlying objective of communicating with the user by masking itself as human.

1.2 SCOPE:

- The primary objective of the proposed system is to build a system which will intelligently optimize the English language inputs. This will be done with the help of Natural Language Toolkit (NLTK) which facilitates Natural Language Processing (NLP).
- The scope of the proposed system encompasses the Natural Language Processing (NLP) dimension in the field of man-machine integration.
- This can work as a stand-alone system or can be integrated with other systems to give increased number of functionalities.

2. LITERATURE SURVEY

Early experimentation in the late 1950's and early 60's suggested that text summarization by computer was feasible though not straightforward (Luhn, 59; Edmundson, 68). The methods developed then were fairly unsophisticated, relying primarily on surface level phenomena such as sentence position and word frequency counts, and focused on producing extracts (passages selected from the text, reproduced verbatim) rather than abstracts (interpreted portions of the text, newly generated). After a hiatus of some decades, the growing presence of large amounts of online text-in corpora and especially on the Web-renewed the interest in automated text summarization. During these intervening decades, progress in Natural Language Processing (NLP), coupled with great increases of computer memory and speed, made possible more sophisticated techniques, with very encouraging results. In the late I 990' s, some relatively small research investments in the US (not more than 10 projects, including commercial efforts at Microsoft, Lexis-Nexis, Oracle, SRA, and TextWise, and university efforts at CMU, NMSU, UPenn, and USC/lSI) over three or four years have produced several systems that exhibit potential marketability, as well as several innovations that promise continued improvement. In addition, several recent workshops, a book collection, and several tutorials testify that automated text summarization has become a hot area. Automatic Text Summarization gained attention as early as the 1950's. A research paper, published by Hans Peter Luhn in the late 1950s, titled "The automatic creation of literature abstracts", used features such as word frequency and phrase frequency to extract important sentences from the text for summarization purposes.

In Paper [1], A SURVEY OF TEXT SUMMARIZATION TECHNIQUES this proposed paper presents a Numerous approaches for identifying important content for automatic text summarization have been developed to date. Topic representation approaches first derive an intermediate representation of the text that captures the topics discussed in the input. Based on these representations of topics, sentences in the input document are scored for importance

Predictive text computer simplified keyboard with word and phrase auto-completion[2] in study a predictive text personal computer simplified keyboard with word and phrase auto-completion. It has a smaller keypad with each key representing several letters/characters so that only 9 keys are required to represent the entire alphabet of 26 characters

In IEEE Paper [3], Spell Checking Techniques in NLP: A Survey [Neha Gupta, Pratishta Mathur], Spell checkers in Indian languages are the basic tools that need to be developed. A spell checker is a software tool that identifies and corrects any spelling mistakes in a text. Spell checkers can be combined with other applications or they can be distributed individually. In this paper the authors are discussing both the approaches and their roles in various applications.

In [4], that IEEE that will be study Grammar checker is one of proofing tool used for syntactic analysis of the text. Various techniques are used for development of grammar checker. These techniques includes rule based technique, statistical based technique and syntax based technique. In this research article, all these three techniques have been discussed. Both advantages and disadvantages of these techniques have also been discussed at the end.

3. PROBLEM STATEMENT

The system should be smart enough to correct the errors in English Language and also summarize it. We will be using the NLTK tool available in Python. NLTK stands for Natural Language Toolkit. We will use different tools available in the Python NLTK for this purpose. Summarization feature will take input as a meaningful paragraph in English Language. Summarization functionality of the system will provide the "meaningful summary" of the paragraph which is taken as input. The output will be the summarized paragraph (where the original meaning will be retained).

The spell check feature will be comparing the input spellings and then suggest a correct one, if any word has been misspelt. The Grammar check feature will be used to find errors in the grammar and suggest the corrections, which are possible. The auto completion feature completes simple sentences automatically.

NLP can be integrated with a website to provide a more user-friendly experience. Features like spell check, autocomplete, and autocorrect in search bars can make it easier for users to find the information they're looking for, which in turn keeps them from navigating away from your site.

Automatic Text Summarization is one of the most challenging and interesting problems in the field of Natural Language Processing (NLP). It is a process of generating a concise and meaningful summary of text from multiple text resources such as books, news articles, blog posts, research papers, emails, and tweets.

The demand for automatic text summarization systems is spiking these days thanks to the availability of large amounts of textual data.

4. FLOWCHART

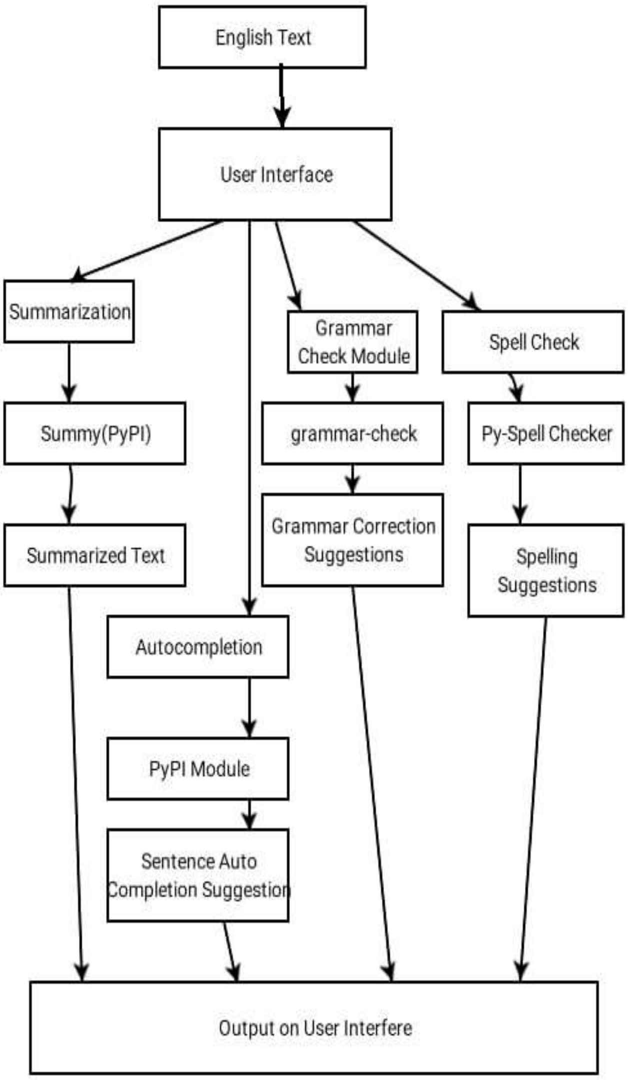

Fig -1: Flow chart of the system

4.1. ALGORITHMS

4.1.1 SUMMARIZATION:

In the process of summarization, the following steps are used:

- Taking user input in the form of a paragraph
- Passing the user input to our summarizer.
- The summarizer then eliminates English language stop words from the extract presented by the user
- It then carries out the process of Stemming
- After stemming, it creates a frequency table which maintains a count of all the distinct words in the extract entered by the user
- Tokenization takes place on the sentences in the frequency table
- Assign weights to words in the frequency table after tokenization of sentences using neural networks
- Find the average score for these values
- Generate the summary based on these values

4.1.2 GRAMMAR CHECK:

Grammar check is carried out by the following steps:

- Taking a grammatically incorrect sentence/ paragraph from the user
- Using the language_check tool in python to check if the input sentence is following all the rules of English grammar
- If the rules are not being followed, then return the issues in the sentence/paragraph entered by the user

4.1.3 SPELL CHECK:

Spell check follows the following algorithm:

1. Takes a sentence/paragraph from the user
2. Checks for spelling of each word in the sentence/paragraph for correctness
3. Return the misspelt words so that they can be highlighted to the user

Spell check is a form of NLP that everyone is used to by now. It's unobtrusive, easy to use, and can reduce a lot of headaches for both users and agents alike.
Not every user is going to take the time to compose a grammatically perfect sentence when contacting a help desk or sales agent. Salesforce knows this, so they made sure their contact form was equipped with spell check to make users' lives easier.

This also makes their employees' lives easier, too. Error-ridden customer messages can be difficult to interpret, leading to miscommunication and frustration for all involved.

4.1.4 AUTO COMPLETE:

Auto complete is a neural networks based algorithm that works on user's previous data(of using the system) to give auto completion suggestions to the user Search autocomplete is another type of NLP that many people use on a daily basis and have almost come to expect when searching for something. This is thanks in large part to pioneers like Google, who have been using the feature in their search engine for years. The feature is just as helpful on company websites.

Salesforce integrated the feature into their personal search engine. Users interested in learning more about a topic or function of Salesforce's product might know one keyword, but maybe not the full term.

Search autocomplete will help them locate the correct information and answer their questions faster. This helps cut down on the likelihood that they'll become disinterested and navigate away from the site.

5. CONCLUSION

The proposed system as planned after extensive research during a literature survey includes the following features: Implementation of Data Mining algorithm for summarization of a given English Language Paragraph. It will also enable the user to perform spell check and Grammar check on the user's inputs to the system in English language.

ACKNOWLEDGEMENT

The success and final outcome of this research required a lot of guidance and assistance and we are extremely privileged to have got this all. All that we have done is only due to such supervision and assistance and we would not forget to thank them.

We respect and thank Prof. Satish Kuchiwale, for providing us insight and expertise that greatly assisted the research. We are extremely thankful to him for providing such a nice support and guidance.

REFERENCES

[1] Ani Nenkova, Kathleen McKeown "A SURVEY OF TEXT SUMMARIZATION TECHNIQUES"

[2] David Gikandi "Predictive text computer simplified keyboard with word and phrase auto-completion"

[3] Neha Gupta, Pratishta Mathur "Spell Checking Techniques in NLP: A Survey"

[4] Blossom Manchanda, Vijay Anant Athavale ,Sanjeev kumar Sharma "Various Techniques Used For Grammar Checking"

[5] N.Moratanch, S.Chitrakala "A Survey on Extractive Text Summarization.

[6] Haoran Li, Junnan Zhu, Cong Ma, Jiajun Zhangand Chengqing Zong "Read, Watch, Listen and Summarize: Multimodal Summarization for Asynchronous Text, Image, Audio and Video

BIOGRAPHIES

Mahesh Sunil Patil, Pursuing the Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering (SIGCE), Navi Mumbai. His current research interests include Web Designing & Machine Learning

Mayur Anand Pawar, Pursuing the Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering (SIGCE), Navi Mumbai. His current research interests include Web Designing & Machine Learning

Yatin Sitaram Rai, Pursuing the Bachelor degree (B.E.) in Computer Engineering from Smt. Indira Gandhi College of Engineering (SIGCE), Navi Mumbai. His current research interests include Web Designing & Machine Learning

Prof. Satish Lalasaheb Kuchiwale, Obtained the Bachelor degree (B.E. IT) in the year 2007 from Rajarambapu Institute of Technology (RAIT), Rajaramnagar,Sakharale, and Master degree (M.E. Computer) from Lokamanya Tilak College of Engineering(LTCE), Navi Mumbai. He is Asst. Professor in Smt. Indira Gandhi College of Engineering of Mumbai university and having about 12 yrs. of experience.