Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 9 Issue 01, January-2020

Natural Language Toolkit based Morphology Linguistics

Alifya Khan
Information Technology
Vidyalankar Institute of Technology
Mumbai, India

Pratyusha Trivedi
Information Technology
Vidyalankar Institute of Technology,
Mumbai, India

Karthik Ashok
Information Technology,
Vidyalankar Institute of Technology,
Mumbai, India

Prof. Kanchan Dhuri
Information Technology,
Vidyalankar Institute of Technology,
Mumbai, India

Abstract— **In the current scenario, there are different apps, websites, etc. to carry out different functionalities with respect to text such as grammar correction, translation of text, extraction of text from image or videos, etc. There is no app or a website where a user can get all these functions/features at one place and hence the user is forced to install different apps or visit different websites to carry out those functions. The proposed system identifies this problem and tries to overcome it by providing various text-based features at one place, so that the user will not have to hop from app to app or website to website to carry out various functions. The proposed system will provide users with various functions such as grammar checking, text extraction from an image, text summarization, translation of text into different languages, etc. which will help them in their daily life.**

I. INTRODUCTION

In this rapidly growing generation and increase in technology as well as productivity of man, now-a-days people want their work to be completed as soon as possible. Therefore, designing a system which will help many walks of life from the corporate sector to students and senior citizens who interact with such technologies on a regular basis.

Most commonly while working on any type of document the writer faces problems like incorrect grammar usage which leads to poor presentation of the document or a high count of lines which tends to miss the core point of the document. Sometimes, a document in a specific language needs to be translated to another language for better user interaction and understanding, this may make the user confused and unable to read the document. Due to this people tend to postpone their work leading to inefficiency.

Now, to get over this one has to use various applications or websites which can be time consuming. This system helps one to perform all the specified tasks at a single website and obtain finalized documents according to one’s needs.

II. PROBLEM STATEMENT

In the current situation, there are many websites and applications available which provide different text-based functionalities like extracting text from an image, translation of text to different languages, grammar check, etc. Generally, all these functionalities are used in tandem with each other. For e.g., to read a sign board in a different language, the first step is to extract text from that image and translate it to any respective language as required. Hence, to do this one has to switch from application to application which can be time consuming. To overcome this problem, an integrated environment is built where all these functionalities are available.

III. PROPOSED SYSTEM

The proposed system provides all the text related features at one place to ease the task of the user. With the help of this system the user will easily be able to enhance their sentences, extract text from images, summaries a whole essay or an article and even will be able to translate the sentences into different languages as per requirement. This will not only save the user’s time but also help them to provide an efficient output in their workspace.

The proposed system is divided into four main modules namely:

Text Extraction from an Image

This module focuses on extracting text from any kind of image and stores it in a document for usage.

Text Translation

This module is based on translation of text from a given language to any specified language as per user requirement.

Grammar Check

This module will help in correcting grammatical errors in English. Once a user inputs a sentence, grammatical errors are checked, and the corrected sentence is provided as output for further use.

Summarization of Article

This module is based on summarization of an article to specified number of sentences as mentioned by the user

IV. METHODOLOGY

As this system is totally based on text-based functionalities, each module represents different text functionalities which the user can use. Below different modules, each of variable functions are briefed as follows.

Text Extraction from An Image

This module focuses on extracting text from any kind of image. The extracted image is first converted to the grey scale value and pre-processing techniques are applied. This helps in identifying darker backgrounds which is then blurred based on the level of darkness. Blurring helps in identifying the text in an assorted background and makes the text clearer to perform extraction. Once extraction is performed the system is ready for output. As the output the system will display the text which has been extracted and will be provided in a convertible format of document.

Text Translation

This module is based on translation of text from English to Hindi, French, Spanish and German as per user requirement. The user will be prompted to enter the text in English. This text will be recognized by the system to verify if the text entered is in English. Further, the list of languages as specified will be chosen by the user and conversion of the text is performed by the system. The output will be displayed in the specified language by the user.

Grammar Check

This module will help in correcting grammatical errors in English. Once the user inputs a sentence, the sentence is tokenized which helps in identifying the part of speech. Further based on a generalized pattern the system identifies whether the sentence is grammatically correct.

Summarization of Article

This module is based on summarization of an article. An article generally consists of many lines and precise information are found in the beginning, middle and at the very end of the article. In order to obtain only significant lines is the motive of this module. The user first enters the article for summarization. Upon this the user enters the number of lines to summarize the article. The system then chooses the sentences based on the count of keywords and higher the frequency, higher is the chance of that sentence to be selected. The top keywords are chosen, and logical sentences are made based on the number of lines that the user has specified.

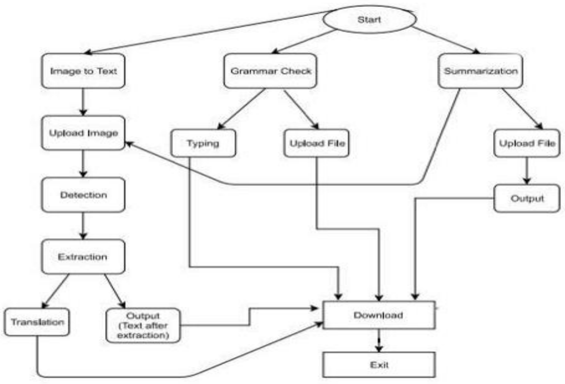

Fig 1. UML Diagram

V. DESIGN TOOLS

Optical Character Recognition:

Optical character recognition or optical character reader (OCR) is used to electronically or mechanically convert the images of typed, handwritten or printed text into machine-encoded text, whether from a scanned document, a photo or the text on signs and hoardings.

OCR is used as a “hidden” technology. OCR technology includes data entry automation, indexing documents for search engines, automatic number plate recognition, as well as assisting blind and visually impaired persons.

OCR technology has proven immensely useful in digitizing historic newspapers and texts that have now been converted into fully searchable formats and had made accessing those earlier texts easier and faster.

Open Source Computer Vision:

Open Source Computer Vision (OpenCV) is an open source computer vision and machine learning software library. It was designed to provide the same infrastructure for all computer vision applications. The library has optimized algorithms. These algorithms are used to detect and recognize faces, identify objects, classify human actions in videos, track moving objects, extract 3D models of objects, stitch images together to produce a high resolution image, remove red eyes from images taken using flash, follow eye movements, recognize scenery, etc.

Natural Language Toolkit:

Natural Language Toolkit (NLTK), is a set of libraries and programs for natural language processing (NLP) for English written in the Python programming language. NLTK includes graphical demonstrations, sample data and underlying concepts behind the language processing tasks supported by the toolkit. It provides easy-to-use interfaces to over 50 corpora and lexical resources such as WordNet along with text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities.

VI. FEASIBILITY STUDY

Technical Feasibility

Technical feasibility focuses on the technical resources (software and hardware) available for the project. It also helps to determine whether the team is capable of converting the ideas into working systems. The software required for the system are a basic text editor like Atom, Sublime Text for writing the code and using Python as the programming language. The hardware components are not required for this system.

Economic Feasibility

As this system is totally based on software does not require any hardware, the budget of the project is null. Hence, it is highly cost effective.

Operational Feasibility

This assessment involves a study to analyze and determine how well the organization’s needs can be met by completing the project. The main objective of the project is to firstly detect text from an image, translate a given text in English language to any specified language as per the user. Secondly, correct grammatical errors of any text in English language and summarize an article to reduce the number of lines and increase the reader's understanding and speed. As all these technicalities mentioned above are present in one system, it is very useful not only for students but for teachers and working professionals as well.

VII. CONCLUSION

In concluding this entire system, one can be assured that their basic text-based difficulties can be solved by this system. One is just a click away to overcome their problems and present a complete document of any size and top quality which will be legible and easy to understand. Most of our day to day text-based difficulties are overcome by this system. Most useful for students and teachers, this system will also prove to be helpful for corporate people and common people in using modules like extracting text from an image or translation. The user will operate in a user-friendly environment by just having a basic knowledge of computer technology.

VIII. FUTURE SCOPE

As all the modules and different applications provided by this system are text based, various text-based functionalities can be deployed together, one of them being handwriting recognition. In this functionality the system will be able to identify handwritten text and provide a text document of the same. The system will scan the handwritten document and identify each letter according to the English alphabets and provide the output as text document respectively. Further, this system can be integrated with various other features like for instance reading a document to the user. Here, the input document will be provided to the system where the system will recognize each word and prompt it to the user. Another feature that can be added is commanding the system to make changes in the document through voice instruction. For example, prompting “Copy this document”, the system will be able to recognize the voice and follow the command and perform it successfully. Finally, converting sentences to active and passive voice can also be added as an extra feature. This functionality will mostly be useful for students and teachers in the literary background as active and passive voice are basic grammatical rules in English language. Thus, including the modules will not only enhance the system but also provide useful and time saving functionalities.

REFERENCES

[1] http://cs229.stanford.edu/proj2018/
[2] https://towardsdatascience.com/build-a-handwritten-text-recognition-system-using-tensorflow-2326a3487cd5
[3] https://pyimagesearch.com/2018/08/20/opencv-text-detection-east-text-detector/
[4] https://github.com/ayesha92ahmad/NLP-image-to-tex
[5] https://github.com/Prashant047/text-extract-from-image
[6] https://www.youtube.com/results?search_query=text+extraction+from+image+using+python
[7] https://www.pyimagesearch.com/2018/09/17/opencv-ocr-and-text-recognition-with-tesseract/
[8] https://www.pyimagesearch.com/category/optical-character-recognition-ocr/
[9] https://www.nltk.org/book/ch08.html
[10] https://www.pyimagesearch.com/2017/07/10/using-tesseract-ocr-python/
[11] https://en.wikipedia.org/wiki/Sentence_diagram
[12] https://www.youtube.com/watch?v=nRBnh4qbPHI&feature=youtu.be
[13] https://realpython.com/natural-language-processing-spacy-python/
[14] http://www.danielnaber.de/languagetool/download/style_and_grammar_checker.pdf
[15] https://www.quora.com/What-kind-of-algorithms-is-Grammarly-using-to-grammar-check/answer/Khushal-Singh-1?ch=3&share=fad573d0&srid=pSniA
[16] https://www.youtube.com/watch?v=Dcvecpq7N0I
[17] https://www.kdnuggets.com/2017/09/machine-learning-translation-google-translate-algorithm.html