Text Encapsulator

Automatic Text Summarizer

Yashovarddhan Malu
Second-Year
Electronics and Telecommunication Department

Vishwakarma Institute of Technology, Pune, Maharashtra, India, 411037

Abstract: In today's world, we don't have the time to sit and read long texts and articles and sometimes wish there was a way to understand the gist of it quickly. Text Encapsulator does just that. It provides a way to summarize long paragraphs and Wikipedia articles in as many sentences as we want using extractive summarization. The intention is to create an articulate summary highlighting the main points outlined in the document. To make the life of the user easy it provides various input methods like direct text entry, uploading text files and images or direct Wikipedia links and also provides a spell checker and dictionary if they have trouble understanding the summary.

Keywords - Dictionary, Extractive Summarization, Spell Checker, Summary, Wikipedia

I. INTRODUCTION

Text summarization is basically the task of compressing a piece of text to a shorter version. Reading long articles and texts is not a convenient activity in today's world as it costs a lot of manual labor and time. So, automatic text summarization is the need of the hour and therefore interests many students and intellects to research in this field. Text summarization finds important applications in many NLP related tasks like text classification, question answering, legal texts summarization, news summarization, and headline generation [1]. Moreover, we can integrate this feature of summarization as an intermediate step so as to output a more precise and shorter gist of the original version. In this era, there has been an eruption in the amount of text data from various sources on the net. Text data of this magnitude is a limitless source of knowledge and information which needs to be properly summarized in order to be useful. This has resulted in an increase in the research in the field of automatic text summarization.

It is a very challenging task, because when humans summarize any piece of text, we usually read it, develop our understanding and then write a summary with the main points. But since machines lack human knowledge and understanding of languages, it makes automatic text summarization very difficult and a non-trivial task.

Generally, there are two different types of approaches we follow to summarize texts automatically:

1. Extractive
2. Abstractive

EXTRACTIVE TEXT SUMMARIZATION

In this approach, we create a summary by directly picking up sentences from the document based on a scoring function to form a meaningful summary. [2] This usually works by identifying important sections in the text, cropping them out and then joining them together to produce a compressed version. Most of the summarization research in today's era has been focused on this approach because it is easier and gives us a summary which is naturally grammatically correct. Moreover, it contains the most important sentences of the input. Some studies have used Latent semantic analysis (LSA) to identify semantically important sentences. [3–5]

Recent studies have applied deep learning as well [6-9]. For example, Sukriti proposes an extractive approach for factual reports using a deep learning model which explores various features to improve the set of sentences selected to produce the final summary [9]. Yong Zhang proposed a summarization framework based on CNN to learn sentence features and perform sentence ranking jointly using a CNN model to rank sentences [10].

ABSTRACTIVE TEXT SUMMARIZATION

In this approach, we paraphrase the main contents of the input text using a vocabulary which is different from the original one [11]. This is very similar to what humans do. We create a semantic representation of the text in our brains and pick words from our general vocabulary and create a short summary. Developing this kind of summarizer is difficult as it requires advanced understanding of NLP. This approach was first proposed in a paper by Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, Bing Xiang from IBM [12]. The term "sequence to sequence models" is used as these models create an output sequence of words from the input text.

II. METHODOLOGY EXPERIMENTAL

A. BLOCK DIAGRAM

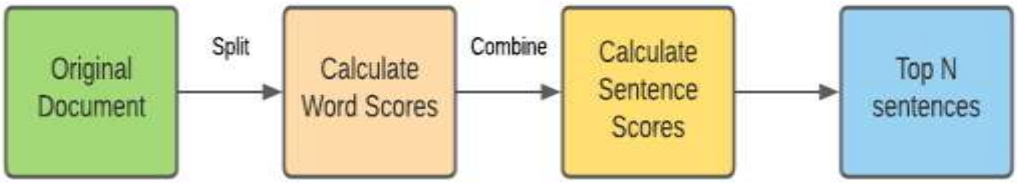


Figure 1: Block Diagram

We follow the extractive approach to create an automatic summary of our input text. We follow three basic steps:

1. Split out input text into all the words and calculate the occurrence frequency of each word and then divide each frequency with the maximum frequency of a word in order to obtain the word scores.
2. Calculate the sentence score for each sentence by adding the word frequency of each word in the sentence.
3. We create a table of the sentence and its corresponding score in descending order and pick the top n sentences as per the user input to create our final summary.

B. SENTENCE SCORES

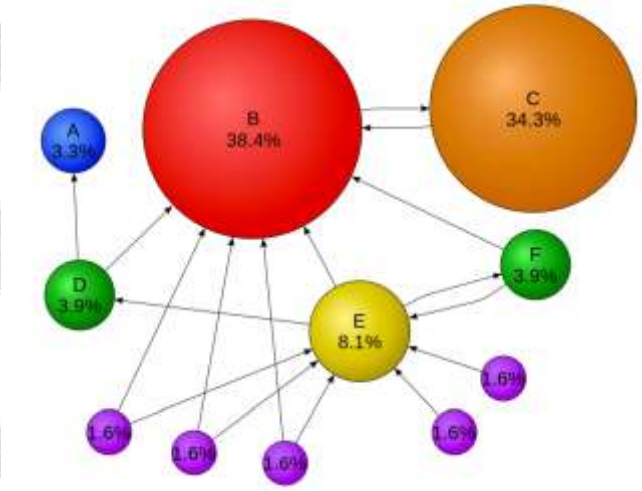


Figure 2: Sentence Scores

The most crucial part for creating a coherent summary in this approach is the calculation of sentence scores in order to pick the top n. To calculate the sentence scores, we first need to calculate the word scores so we split the text to obtain a list of all the words. Then we iterate through the list and calculate the frequency of occurrence of each word. Then we divide the frequencies by the highest frequency to obtain the word score of each word. We finally calculate the sentence scores by adding the word score of each word in the sentence. A table is created for each sentence and its respective score and then the top n sentences are selected.

C. INPUT METHODS/ADDITIONAL FEATURES

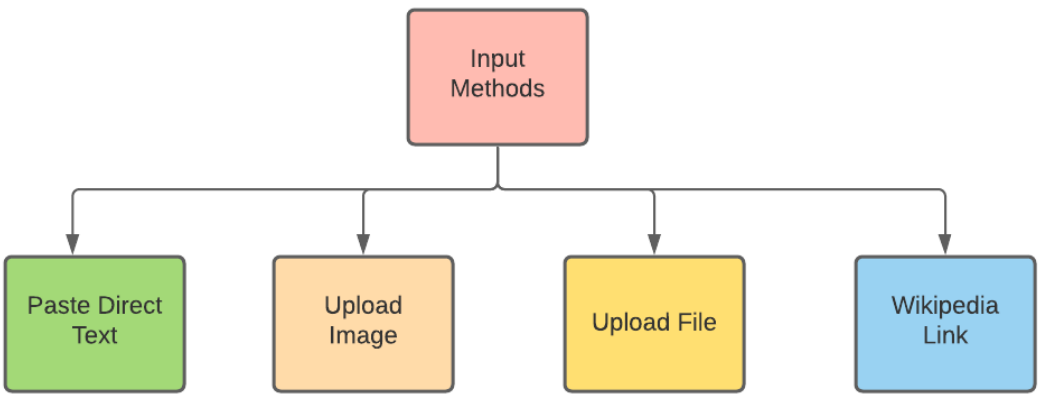


Figure 3: Input Methods

This project provides various input methods for the ease of the user. There are four different methods to input texts in the Text Encapsulator:

1. Direct Text – The user can directly paste the text he/she wants to summarize.
2. Upload Text File – The user has an option to upload the file he/she wants to summarize. The software will read the contents in the document and convert it into plain text before finally summarizing it.
3. Upload Image – Text Encapsulator also provides a way to upload images that contain texts that we want to summarize. A python module named pytesseract has been used to scan text from the image with 100% accuracy.
4. Wikipedia Link – We can directly input the link of the Wikipedia link that we want to summarize. Web scraping has been used to get information in the form of plain text from the webpage.

Text Encapsulator also provides a spell checker and a dictionary to cross check the summaries and words. For the spell checker, a python API called "spellchecker" has been used which provides functions to check the spelling mistakes in the text provided as input. The dictionary also uses a python API namely "PyDictionary". It provides various functions to find the meaning of the word, its synonyms and antonyms.

D. FRONT-END

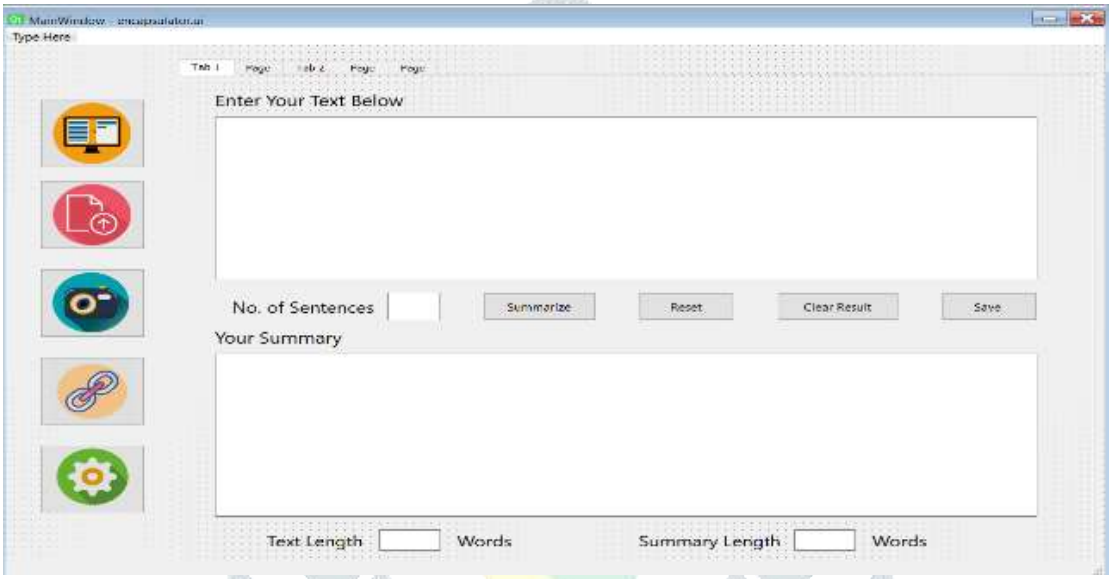


Figure 4: Front-End using QT Designer

The front-end has been created using Qt Designer which is a software that helps us create sophisticated and impressive GUIs for softwares coded in Python. We use the "pyqt5" python module to load the .ui file with the back-end and make changes in the front-end as per requirements.

III. RESULTS AND DISCUSSIONS

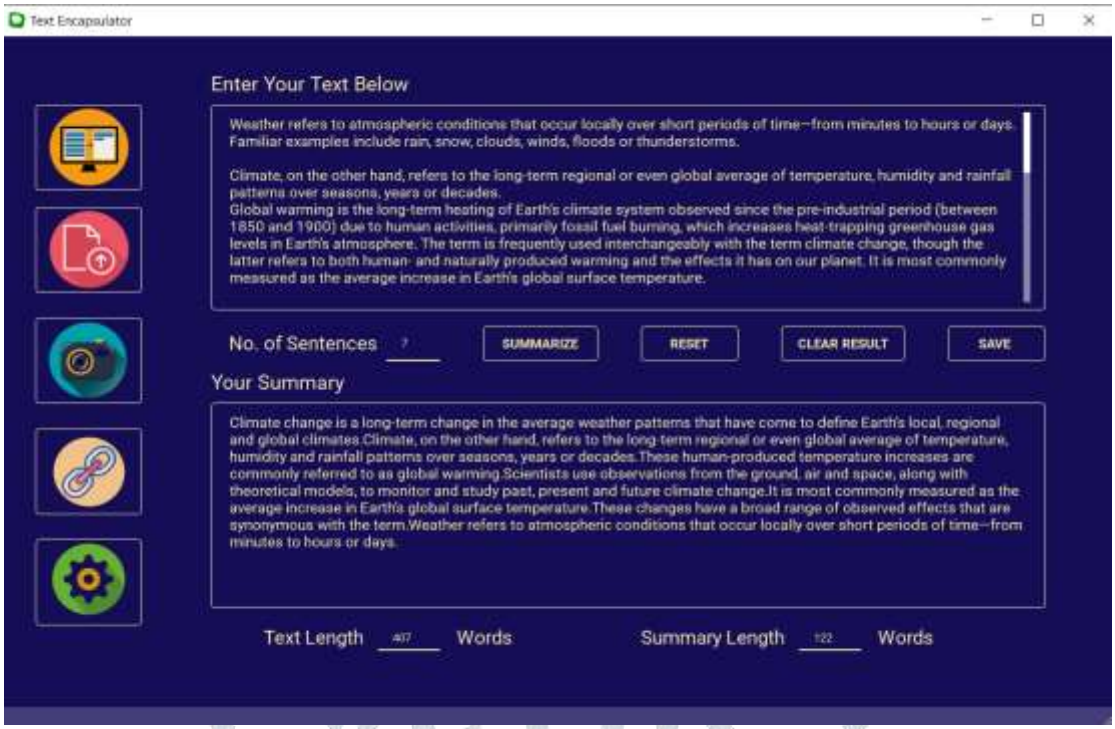


Figure 5: Final Software

Text Encapsulator is a desktop software that summarizes text automatically for us in a very short span of time. Its backend has been coded in Python and frontend made using Qt Designer and PyQt5 in Python. It provides various input methods for the user in which he/she can give the text. Input text can either be directly pasted in the text box, uploaded in the form of a text file or an image or a direct Wikipedia link we might want to summarize. Additionally, it also provides a spell checker and a dictionary.

The created summary is formed using the extractive approach as it is easy to implement and arguably faster than the abstractive approach. We also don't have to worry about grammatical mistakes here as it takes sentences as it is from the input texts.

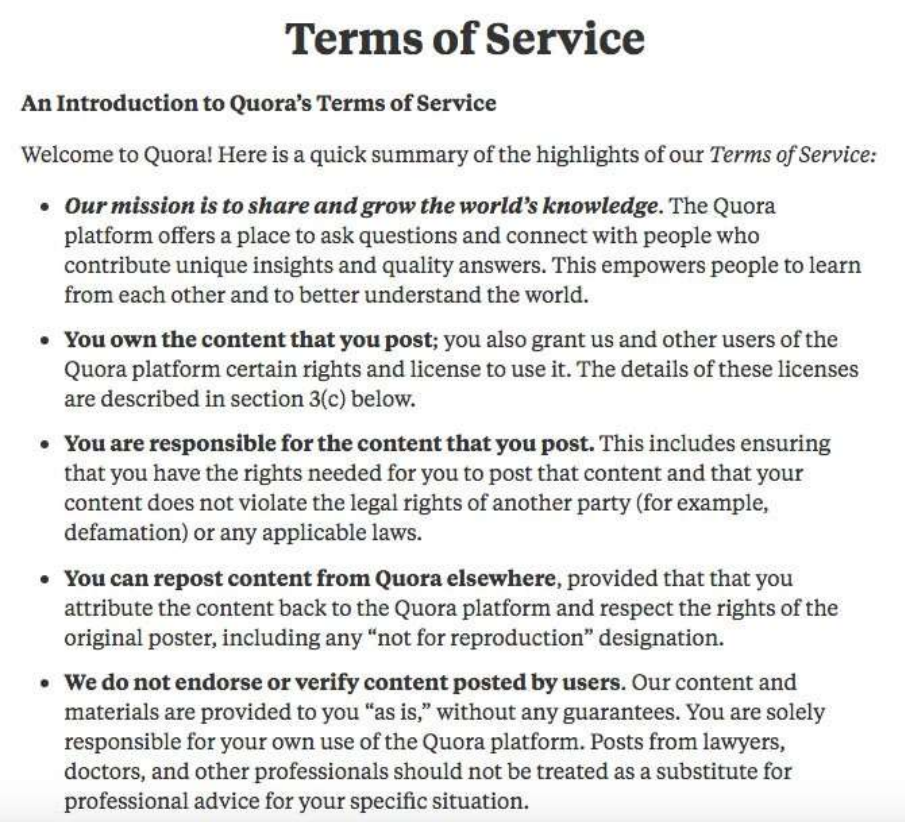
Below are some of the summaries that we get from the various input methods available in the software.

A. INPUT TEXT

"Climate change is a long-term change in the average weather patterns that have come to define Earth's local, regional and global climates. These changes have a broad range of observed effects that are synonymous with the term. Changes observed in Earth's climate since the early 20th century are primarily driven by human activities, particularly fossil fuel burning, which increases heat-trapping greenhouse gas levels in Earth's atmosphere, raising Earth's average surface temperature. These human-produced temperature increases are commonly referred to as global warming. Natural processes can also contribute to climate change, including internal variability (e.g., cyclical ocean patterns like El Niño, La Niña and the Pacific Decadal Oscillation) and external forcing (e.g., volcanic activity, changes in the Sun's energy output, variations in Earth's orbit). Scientists use observations from the ground, air and space, along with theoretical models, to monitor and study past, present and future climate change. Climate data records provide evidence of climate change key indicators, such as global land and ocean temperature increases; rising sea levels; ice loss at Earth's poles and in mountain glaciers; frequency and severity changes in extreme weather such as hurricanes, heatwaves, wildfires, droughts, floods and precipitation; and cloud and vegetation cover changes, to name but a few."

Summary - "Climate change is a long-term change in the average weather patterns that have come to define Earth's local, regional and global climates. Scientists use observations from the ground, air and space, along with theoretical models, to monitor and study past, present and future climate change."

B. OCR SCANNED IMAGE



Terms of Service

An Introduction to Quora's Terms of Service

Welcome to Quora! Here is a quick summary of the highlights of our *Terms of Service:*

- ***Our mission is to share and grow the world's knowledge.*** The Quora platform offers a place to ask questions and connect with people who contribute unique insights and quality answers. This empowers people to learn from each other and to better understand the world.
- **You own the content that you post;** you also grant us and other users of the Quora platform certain rights and license to use it. The details of these licenses are described in section 3(c) below.
- **You are responsible for the content that you post.** This includes ensuring that you have the rights needed for you to post that content and that your content does not violate the legal rights of another party (for example, defamation) or any applicable laws.
- **You can repost content from Quora elsewhere**, provided that that you attribute the content back to the Quora platform and respect the rights of the original poster, including any "not for reproduction" designation.
- **We do not endorse or verify content posted by users.** Our content and materials are provided to you "as is," without any guarantees. You are solely responsible for your own use of the Quora platform. Posts from lawyers, doctors, and other professionals should not be treated as a substitute for professional advice for your specific situation.

Figure 6: Terms and Conditions for Quora provided as input

Summary - You own the content that you post; you also grant us and other users of the Quora platform certain rights and license to use it. You are responsible for the content that you post. Our mission is to share and grow the world's knowledge. The Quora platform offers a place to ask questions and connect with people who contribute unique insights and quality answers.

C. WIKIPEDIA LINK

Link - https://en.wikipedia.org/wiki/Automatic_summarization

Summary - An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. A promising line in document summarization is adaptive document/text summarization. The idea of adaptive summarization involves preliminary recognition of document/text genre and subsequent application of summarization algorithms optimized for this genre. Another important distinction is that TextRank was used for single document summarization, while LexRank has been applied to multi-document summarization. For example, in document summarization, one would like the summary to cover all important and relevant concepts in the document. At a very high level, summarization algorithms try to find subsets of objects (like set of sentences, or a set of images), which cover information of the entire set. Pattern-based summarization was the most powerful option for multi-document summarization found by 2016.Intra-textual methods assess the output of a specific summarization system, and the inter-textual ones focus on contrastive analysis of outputs of several summarization systems.

D. SPELL CHECKER

Input Text - As global communication expands worldwide, so does the need for a global language. English is a language that is used to communicate in many parts of the world. English is not only used as a mere casually communication tool, but it is also used in business, diplomatic policies, tourism, in entertainment and educational sectors, and to communicate between two non-English natoive nations. English is used everywhere. In this age of globalization, countries like China, Japan and Russia, who have passionately sustained and promoted their language, have today accepted the importance of the English language. In India, English is an ofeficial language used to communicate between different states having their language. It has become a medium of instruction in educational institutions. For higher education (e.g., medicine, engineering), only English is used to teach. Though English is a relatively easy language and is used as a global communication tool, one must never disregard their native language. We should all prleserve our language, culture and heritage even if we learn English. English is considered as a global language. Many non-native English countries have English as their second language. Though Chinese has the most numbers of native speakers, English is spoken in different parts of the world. English is derived from Latin, French

and some other European languages. In India, English is the official language. English is a relatively easy language for most learners. The English language has a simple letter system which is based on phonetics. English has a vast vocabulary that helps one to express things accurately. The source of English to become a global language is from the British colonialism and American Imperialism. English is used as a global communication tool.

Misspelled Words –

- prleserve
- ofeficial
- natoive

Suggestions –

- preserve
- official
- native

IV. FUTURE SCOPE AND CONCLUSION

This project holds a lot of potential for further advancement with all the research happening in the field of automatic summarization. The accuracy of the summary can be worked upon and better summarization models and frameworks could be introduced to make better summaries. Automatic text summarization is a very useful concept especially in today's fast paced world. It helps save time and labor of reading and understanding long and confusing texts or articles. It provides a gist of the input text in a more concise manner and helps the reader understand it quickly.

V. ACKNOLEDGEMENTS

I would like to thank the honorable Director of my institute Prof. (Dr.) R.M Jalnekar, Vishwakarma Institute of Technology, Pune and our professors and colleagues for giving me strong moral support and inspiration. I would also like to express my gratitude to my project guide Prof. Rupali Tornekar for helping me throughout this project.

REFERENCES

[1] V. Alwis, "Intelligent E-news Summarization," 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), 2018, pp. 189-195, doi: 10.1109/ICTER.2018.8615590.

[2] Text Summarization Techniques: A Brief Survey - Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assef, Saeid Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut

[3] Ozsoy, Makbule & Alpaslan, Ferda & Cicekli, Ilyas. (2011). Text summarization using Latent Semantic Analysis. J. Information Science. 37. 405-417. 10.1177/0165551511408848.

[4] Steinberger, Josef & Jezek, Karel. (2004). Using Latent Semantic Analysis in Text Summarization and Summary Evaluation.

[5] An Enhanced Latent Semantic Analysis Approach for Arabic Document Summarization - Kamal Al-Sabahi, Zuping Zhang, Jun Long, Khaled Alwesabi

[6] A Neural Attention Model for Abstractive Sentence Summarization EMNLP 2015 · Alexander M. Rush, Sumit Chopra, Jason Weston

[7] Neural Extractive Text Summarization with Syntactic Compression IJCNLP 2019 · Jiacheng Xu, Greg Durrett

[8] DebateSum: A large-scale argument mining and summarization dataset 14 Nov 2020 · Allen Roush, Arvind Balaji

[9] Extractive Summarization using Deep Learning - Sukriti Verma, Vagisha Nidhi

[10] Abstract Text Summarization with a Convolutional Seq2seq Model Yong Zhang, Dan Li, Yuheng Wang, Yang Fang and Weidong Xiao

[11] Abstractive Summarization of Spoken and Written Instructions with BERT KDD Converse 2020 · Alexandra Savelieva, Bryan Au-Yeung, Vasanth Ramani

[12] Abstractive Text Summarization Using Sequence-to-Sequence RNNs and Beyond - Ramesh Nallapati, Bowen Zhou, Cicero Nogueira dos santos, Caglar Gulcehre, Bing Xiang