Text Summarization using NLP

A Project Report in partial fulfilment of the degree

Bachelor of Technology

in

Computer Science & Engineering / Electronics & Communication Engineering

By

19K41A0434	Aravelli Tejaswi
19K41A0570	Gorantla Sathvika
19K41A0573	Koutam Sudeeptha

Under the guidance of

Dr. D. RAMESH

Assistant Professor School of CS & AI SRU

Submitted to

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

S.R. ENGINEERING COLLEGE (A), ANANTHASAGAR, WARANGAL

(Affiliated to JNTUH, Accredited by NBA) May-2022.

SR Engineering College
Innovation . Creativity . Entrepreneurship

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

This is to certify that the Project Report entitled "TEXT SUMMERIZATION USING NLP" is a record of bonafide work carried out by the student(s) Aravelli Tejaswi, Gorantla Sathvika, Koutam Sudeeptha bearing Roll No(s) 19K41A0434, 19K41A0570, 19K41A0573 during the academic year 2022-23 in partial fulfillment of the award of the degree of **Bachelor of Technology** in **Computer Science & Engineering / Electronics & Communication Engineering** by the S.R. ENGINEERING COLLEGE, Ananthasagar, Warangal.

Supervisor **Head of the Department**

External Examiner

ABSTRACT

In this new era, where tremendous information is available on the Internet, it is most important to provide the improved mechanism to extract the information quickly and most efficiently. It is very difficult for human beings to manually extract the summary of a large documents of text. There are plenty of text material available on the Internet. So there is a problem of searching for relevant documents from the number of documents available, and absorbing relevant information from it. In order to solve the above two problems, the automatic text summarization is very much necessary. Text summarization is the process of identifying the most important meaningful information in a document or set of related documents and compressing them into a shorter version preserving its overall meanings. We have developed using BART.

TABLE OF CONTENTS

1. INTRODUCTION

Artificial Intelligence (AI) is the part of computer science that focuses on designing intelligent computer systems that show the traits we re-late with human intelligence like comprehending languages, learning problem-solving, decision making, etc. One of the significant contributions of AI has remained in Natural Language Processing (NLP), which glued together linguistic and computational techniques to assist computers in understanding human languages and facilitating human-computer interaction. Machine Translation, Chat bots or Conversational Agents, Speech Recognition, Sentiment Analysis, Text summarization, etc., fall under the active research areas in the domain of NLP. However, in the past few years, Sentiment analysis has become a demanding realm. Nowadays, Artificial Intelligence has spread its wings into Thinking Artificial Intelligence and Feeling Artificial Intelligence (Huang and Rust 2021). Figure 1 shows the sub domain of artificial intelligence. Thinking AI is designed to process information in order to arrive at new conclusions or decisions. The data are usually unstructured. Text mining, speech recognition, and face detection are all examples of how thinking AI can identify patterns and regularities in data. Machine learning and deep learning are some of the recent approaches to how thinking AI processes data.

AI has made a big impact on the globe. AI was reintroduced in a significant manner in the twentieth century, and it inspired researchers to perform in-depth studies in domains like NLP, and machine learning. However, the domains of NLP remain ambiguous due to its computational methodologies, which assist computers in understanding and producing human-computer interactions in the form of text and voice. Text summarization refers to the technique of shortening long pieces of text. The intention is to create a coherent and fluent summary having only the main points outlined in the document. Automatic text summarization is a common problem in machine learning and natural language processing (NLP).

Text Summarization is one of those applications of Natural Language Processing which is bound to have a huge impact on our lives. It is a process of generating a concise and meaningful summary of text from multiple text resources such as books, articles etc. applying text summarization reduces reading time, accelerates the process of researching for information, and increases the amount of information that can fit in an area. Text summarization is an interesting machine learning field that is increasingly gaining traction. As

research in this area continues, we can expect to see breakthroughs that will assist in fluently and accurately shortening long text documents. Extractive text summarization involves the selection of phrases and sentences from the source document to make up the new summary. Techniques involve ranking the relevance of phrases in order to choose only those most relevant to the meaning of the source. Abstractive text summarization involves generating entirely new phrases and sentences to capture the meaning of the source document. This is a more challenging approach, but is also the approach ultimately used by humans. Classical methods operate by selecting and compressing content from the source document.

2. LITERATURE REVIEW

There are few papers on which we studied and did out literature reviews on. Many did their study on different models like reinforcement learning, LSTM, Bi-LSTM, aequenxe to sequence models.

[1] Dima Suleiman and Arafat Awajan 2020. In recent years, the volume of textual data has rapidly increased, which has generated a valuable resource for extracting and analyzing information. To retrieve useful knowledge within a reasonable time period, this information must be summarized. The Giga word dataset is commonly employed for single-sentence summary approaches, while the Cable News Network (CNN)/Daily Mail dataset is commonly employed for multi sentence summary approach. They have used the LSTM and Seq2Seq models.

[2] Y.M. Wazery 2022. In this paper, an abstractive Arabic text summarization system is proposed, based on a sequence-to-sequence model. +is model works through two components, encoder and decoder. Our aim is to develop the sequence-to-sequence model using several deep artificial neural networks to investigate which of them achieves the best performance. Different layers of Gated Recurrent Units (GRU), Long Short-Term Memory (LSTM), and Bidirectional Long Short-Term Memory (BiLSTM) have been used to develop the encoder and the decoder.

[3] Mengli Zhang 2022 Automatic text summarization (ATS) is becoming an extremely important means to solve this problem. .e core of ATS is to mine the gist of the original text and automatically generate a concise and readable summary. Recently, to better balance and develop these two aspects, deep learning (DL)-based abstractive

summarization models have been developed. At present, for ATS tasks, almost all state-of-the-art (SOTA) models are based on DL architecture. However, a comprehensive literature survey is still lacking in the .eld of DL-based abstractive text summarization.

[4] Sara Tarannum, Piyush Sonar 2021. This paper, describes a unique approach for extractive summarization with sentiment analysis for two-level text summarizing from online news ources. At first, important sentences are extracted and individual summaries are prepared and later it is extended to sentimental analysis.

[5] Qicai Wang, Peiyu Liu 2019. Firstly, we convert the human-written abstractive summaries to the ground truth labels. Secondly, we use BERT word embedding as text representation and pre-train two sub-models respectively. Finally, the extraction network and the abstraction network are bridged by reinforcement learning. To verify the performance of the model, we compare it with the current popular automatic text summary model, and use the ROUGE metrics as the evaluation method.

[6] Tarun Miran and 2017. This paper is based on 2 level summarization in first level -the URLs are fetched as an input and two/three summaries are generated primarily from news articles by applying extraction-based method. Firstly it is applied on classifying tweets and reached an accuracy of 75% with SVM and later 90%accuracy with Decision and RF. Sentimental analysis is done to get a positive ,negative ,neutral opinion on the article.

[7] Ji Eun Lee, Hyun Soo Park and 2013. In this experiment, experimental data is collected from graduated students using news articles, by collecting URL of an article selected by users for the summarization and extract attributes from the collected data . Then ,applied data mining techniques using data mining tool WEKA to eliminate the ads and a decision model is being generated. They have used the Navie Bayes Theorem.

[8] Abu Kaiser Mohammed Marum and 2019. Main purpose of this paper is to create an short, fluent, abstractive summary of text. In this experiment, successfully reduced the training loss with a value of 0.036 and abstractive text summarizer able to create a short summary of English to English text. Model used in this are the sequence to sequence model with a two-layered bidirectional RNN's. On the input text and two layers RNN's, each with an LSTM using on the target text to produce an extensive summary.

3. DESIGN

a. REQUIREMENT SPECIFICATION (S/W & H/W)

Hardware Requirements

- ✓ **System** : Intel Core i3, i5, i7 and 2GHz Minimum
- ✓ **RAM** : 4GB or above
- ✓ **Hard Disk** : 10GB or above
- ✓ **Input** : Keyboard and Mouse
- ✓ **Output** : Monitor or PC

Software Requirements

- ✓ **OS** : Windows 8 or Higher Versions
- ✓ **Platform** : Jupyter Notebook, Google Colab
- ✓ **Program Language** : Python

b. FLOW CHART

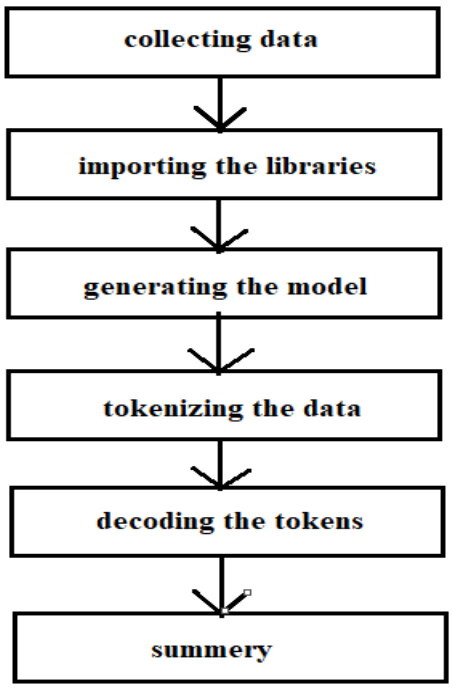

Figure 1: Flow Chart

4. METHODOLOGY

a. BERT

Summarization refers to extracting (summarizing) out the relevant information from a large document while retaining the most important information. BERT (Bidirectional Encoder Representations from Transformers) introduces rather advanced approach to perform NLP tasks. BERT (Bidirectional transformer) is a transformer used to overcome the limitations of RNN and other neural networks as Long term dependencies. It is a pre-trained model that is naturally bidirectional. This pre-trained model can be tuned to easily to perform the NLP tasks as specified, Summarization in our case. Being trained as a masked model the output vectors are tokened instead of sentences. Unlike other extractive summarizers it makes use of embeddings for indicating different sentences and it has only two labels namely sentence A and sentence B rather than multiple sentences. These embeddings are modified accordingly to generate required summaries. refers to the representation of words in their vector forms. It helps to make their usage flexible. Embedding even the *Google* utilizes the this feature of BERT for better understanding of queries. It helps in unlocking various functionality towards the semantics from understanding the intent of the document to developing a similarity model between the words.

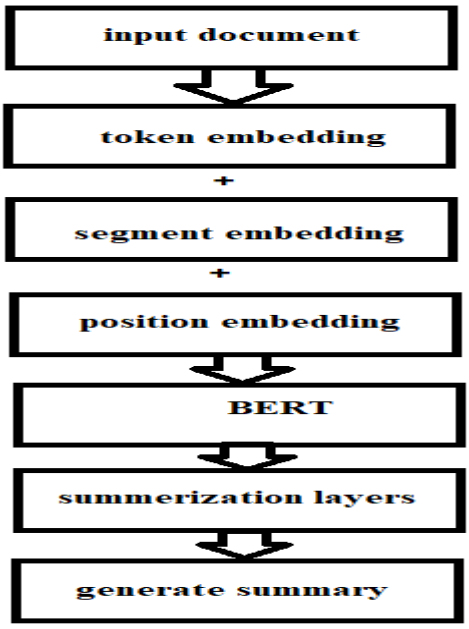

Figure 2: BERT flow chart

There are three types of embeddings applied to our text prior to feeding it to the BERT layer, namely:

- Token Embeddings - Words are converted into a fixed dimension vector. [CLS] and [SEP] is added at the beginning and end of sentences respectively.
- Segment Embeddings - It is used to distinguish or we can say classify the different inputs using binary coding.
- Position Embeddings - BERT can support input sequences of 512. Thus the resulting vector dimensions will be (512,768). Positional embedding is used because the position of a word in a sentence may alter the contextual meaning of the sentence and thus should not have same representation as vectors.

There are following two bert models introduced:

1. BERT base
   In the BERT base model we have 12 transformer layers along with 12 attention layers and 110 million parameters.

2. BERT Large
   In BERT large model we have 24 transformer layers along with 16 attention layers and 340 million parameters.

Transformer layer- Transformer layer is actually a combination of complete set of encoder and decoder layers and the intermediate connections. Each encoder includes *Attention layers* along with a RNN. Decoder also has the same architecture but it includes another attention layer in between them as does the seq2seq model. It helps to concentrate on important words.

b. BART MODEL

BART is a sequence-to-sequence model trained as a denoising autoencoder. This means that a fine-tuned BART model can take a text sequence (for example, English) as input and produce a different text sequence at the output (for example, French). This type of model is relevant for machine translation (translating text from one language to another), question-answering (producing answers for a given question on a specific corpus), text summarization (giving a summary of or paraphrasing a long text document), or sequence classification (categorizing input text sentences or tokens). Another task is sentence entailment which, given two or more sentences,

evaluates whether the sentences are logical extensions or are logically related to a given statement. Since the unsupervised pretraining of BART results in a language model, we can fine-tune this language model to a specific task in NLP. Because the model has already been pre-trained, fine-tuning does not need massive labelled datasets (relative to what one would need for training from scratch). The BART model can be fine-tuned to domain-specific datasets to develop applications such as medical conversational chatbots, converting natural text to programming code or SQL queries, context-specific language translation apps, or a tool to paraphrase research papers. BART was trained as a denoising autoencoder, so the training data includes "corrupted" or "noisy" text, which would be mapped to clean or original text. BART is constructed from a bi-directional encoder like in BERT and an autoregressive decoder like GPT. BERT has around 110M parameters while GPT has 117M, such trainable weights. BART being a sequenced version of the two, fittingly has nearly 140M parameters. Many parameters are justified by the supreme performance it yields on several tasks compared to fine-tuned BERT or its variations like RoBERTa, which has 125M parameters in its base model.

Model	Description	# params
`bart.base`	BART model with 6 encoder and decoder layers	140M
`bart.large`	BART model with 12 encoder and decoder layers	400M
`bart.large.mnli`	`bart.large` finetuned on `MNLI`	400M
`bart.large.cnn`	`bart.large` finetuned on `CNN-DM`	400M
`bart.large.xsum`	`bart.large` finetuned on `Xsum`	400M

Figure 3: parameters for different BART models

5. RESULT

TEXT:

Johannes Gutenberg (1398 – 1468) was a German goldsmith and publisher who introduced printing to Europe. His introduction of mechanical movable type printing to Europe started the Printing Revolution and is widely regarded as the most important event of the modern period. It played a key role in the scientific revolution and laid the basis for the modern knowledge-based economy and the spread of learning to the masses.

Gutenberg many contributions to printing are: the invention of a process for mass-producing movable type, the use of oil-based ink for printing books, adjustable molds, and the use of a wooden printing press. His truly epochal invention was the combination of these elements into a practical system that allowed the mass production of printed books and was economically viable for printers and readers alike.

In Renaissance Europe, the arrival of mechanical movable type printing introduced the era of mass communication which permanently altered the structure of society. The relatively unrestricted circulation of information—including revolutionary ideas—transcended borders, and captured the masses in the Reformation. The sharp increase in literacy broke the monopoly of the literate elite on education and learning and bolstered the emerging middle class.

figure 4: Text to be summarized

SUMMERY:

Johannes Gutenberg (1398 – 1468) was a German goldsmith and publisher. His introduction of mechanical movable type printing to Europe started the Printing Revolution. It is widely regarded as the most important event of the modern period. It played a key role in the scientific revolution and laid the basis for the modern knowledge-based economy and the spread of learning to the masses. The relatively unrestricted circulation of information—including revolutionary ideas—transcended borders, and captured the masses in the Reformation.

Figure 5: Summarized text

6. CONCLUSION

From this we can conclude that the process of text summarization can be made simpler with the help of the Artificial Intelligence and the Natural Language Processing models. We did this by using the BART models. This can also be developed by using various methods like Seq2Seq, LSTM, and any other up coming new models. We can also make this language specific by training the model in the required way and getting the summery only for the specific language texts.

7. REFERENCES

[1] L. M. Al Qassem, D. Wang, Z. Al Mahmoud, H. Barada, A. Al-Rubaie, and N. I. Almoosa, "Automatic Arabic summarization: a survey of methodologies and systems," Procedia Computer Science, vol. 117, pp. 10–18, 2017. View at: Publisher Site | Google Scholar

[2] A. B. Al-Saleh and M. E. B. Menai, "Automatic Arabic text summarization: a survey," Artificial Intelligence Review, vol. 45, no. 2, pp. 203–234, 2016. View at: Publisher Site | Google Scholar

[3] K. Sarkar, "Using domain knowledge for text summarization in medical domain," International Journal of Recent Trends in Engineering, vol. 1, no. 1, Article ID 200, 2009. View at: Google Scholar

[4] A. M. Azmi and N. I. Altmami, "An abstractive Arabic text summarizer with user controlled granularity," Information Processing & Management, vol. 54, no. 6, pp. 903–921, 2018. View at: Publisher Site | Google Scholar

[5] C. Sunitha, A. Jaya, and A. Ganesh, "A study on abstractive summarization techniques in indian languages," Procedia Computer Science, vol. 87, pp. 25–31, 2016. View at: Publisher Site | Google Scholar

[6] G. C. V. Vilca and M. A. S. Cabezudo, "A study of abstractive summarization using semantic representations and dis- course level information," Text, Speech, and Dialogue, pp. 482–490, 2017.

[7] W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: a com- prehensive survey," Expert Systems with Applications, vol. 165, Article ID 113679, 2021.

[8] S. Syed, Abstractive Summarization of Social Media Posts: A case Study using Deep Learning, Master's thesis, Bauhaus University, Weimar, Germany, 2017.

[9] D. Suleiman and A. A. Awajan, "Deep learning based ex- tractive text summarization: approaches, datasets and eval- uation measures," in Proceedings of the 2019 Sixth International Conference on Social Networks Analysis, Man- agement and Security (SNAMS), pp. 204–210, Granada, Spain, 2019.

[10] Q. A. Al-Radaideh and D. Q. Bataineh, “A hybrid approach for Arabic text summarization using domain knowledge and genetic algorithms,” Cognitive Computation, vol. 10, no. 4, pp. 651–669, 2018.