Grenze International Journal of Engineering and Technology, June Issue

Abstractive Text Summarization using Deep Learning Techniques

Raksha Aruloli[1] and Dr J Uma Maheswari[2]
[1]UG Student, Computer Science and Engineering Vellore Institute of Technology Chennai, India
Email: raksha.aruloli2020@vitstudent.ac.in
[2]Associate Professor, SCOPE Vellore Institute of Technology Chennai, India
Email: umamaheswari.j@vit.ac.in

***Abstract*— It is essential to be able to extract valuable insights from large amounts of textual data in this era of information overload. A potent natural language processing (NLP) task called abstractive text summarization involves producing brief, logical summaries that effectively convey the main ideas of the source text. This research uses state-of-the-art deep learning techniques to produce summaries that resemble those of a human. Our project intends to revolutionise textual information condensing by utilising the power of advanced neural networks, specifically Long Short-Term Memory (LSTM) networks and attention mechanisms. Through comprehension and acculturation to complex patterns and relationships found in the text, the model will be able to provide abstractive summaries on its own. The research makes use of cutting-edge pre-trained language models, like T5 (Text-To- Text Transfer Transformer), to improve abstractive summarization's efficacy and efficiency. Additionally, our approach emphasizes multilingual capabilities. After incorporating T5 and Seq2Seq models, we ensure that summaries can be generated in multiple languages according to user preferences.**
The ultimate objective is to provide a comprehensive solution for users to extract crucial insights efficiently from large volumes of textual data, irrespective of language, thus advancing the frontier of multilingual abstractive text summarization.

***Index Terms*— LSTM, T5, Transformer, abstractive, textual, multilingual**

I. INTRODUCTION

In today's information age characterized by an exponen- tial growth in data production, the demand for intelligent systems capable of extracting meaning from vast amounts of textual content has become increasingly imperative. Among the various approaches to this challenge, Abstrac- tive Text Summarization emerges as a promising solution, offering a nuanced understanding and representation of underlying meanings beyond mere word condensation.

Traditional extractive summarization techniques often fall short in conveying the intricacies and context of orig- inal texts.This research endeavors to push the boundaries of abstractive summarization by seamlessly integrating deep learning with natural language processing while embracing multilingual support for a globally inclusive approach.

Central to the project's methodology is the utilization of sophisticated deep learning methods, with particular

Grenze ID: 01.GIJET.10.2.556_3

emphasis on attention mechanisms and Long Short-Term Memory (LSTM) networks. These components are pivotal in endowing the model with the capacity to compre- hend both sequential and contextual aspects of language, thereby ensuring that the generated summaries are not only grammatically precise but also semantically enriched. Furthermore, the project integrates multilingual support to cater to diverse linguistic backgrounds, facilitating efficient communication across language barriers.

A significant aspect of the project lies in the adoption of pre-trained language models, such as T5, which furnish a comprehensive understanding of linguistic nuances. These models, having undergone extensive training on diverse language tasks, enhance the summarization model's lin- guistic prowess through techniques like tokenization and embeddings, operating at a heightened level of abstraction. Throughout the project, evaluations will be conducted on the developed abstractive summarization model's per- formance and adaptability. By furnishing a robust and adaptable solution, this project endeavors to usher in a new era of efficient information distillation, making valuable insights easily accessible to individuals across diverse linguistic backgrounds.

II. Related Work

[1] The paper titled "NLP based Machine Learning Ap- proaches for Text Summarization" from Delhi Technological University reviews recent advancements, focusing on methods such as Machine Learning, Neural Networks, reinforcement learning, and sequence-to-sequence model- ing. Utilizing datasets like CNN corpus and DUC2003, the study highlights the effectiveness of combining these approaches, demonstrating improved summarization accu- racy compared to individual methods.

[2] The paper by Widyassari et al. examines automatic text summarization techniques, utilizing Gigaword and CNN datasets. They propose a Hybrid Method incor- porating MMI, PSO, and Fuzzy techniques, alongside an MS-Pointer Network to address word and semantic inaccuracies. While outperforming other ML approaches and achieving high Rouge scores, it struggles with sentence redundancy.

[3] Giamblanco and Siddavaatam introduce a unique ap- proach for Keyword and Keyphrase Extraction, leveraging Newton's Law of Universal Gravitation with SemEval data. Their innovative Newtonian method showcases su- periority over traditional methods like RAKE and TfIDF, offering promising avenues for more effective keyword extraction techniques.

[4] Ozsoy and Alpaslan delve into Text Summariza- tion through Latent Semantic Analysis, focusing on the Duc2002 dataset. They propose a cross approach within LSA, showcasing superior performance compared to other LSA-based methods and competitive results against alter- native algorithms. Notably, their approach shines particularly in summarizing longer documents, while revealing limitations for shorter texts, suggesting potential avenues for refinement and optimization.

[5] Allahyari et al. conduct a comprehensive survey on Text Summarization Techniques, highlighting varied approaches such as topic representation and indicator representation. Topic representation encompasses meth- ods like topic words, frequency-driven techniques (word probability, TFIDF), latent semantic analysis (LSA), and Bayesian topic models. Evaluation is performed using the ROUGE metric, comparing machine-generated summaries against human reference summaries, providing insights into the effectiveness of different summarization strategies.

[6]Abu Nada et al.tackle Arabic Text Summariza- tion utilizing the AraBERT model through an extractive approach. Leveraging summaries from Arabic language specialists, they employ the AraBERT model for sentence embedding and K-Means for clustering. While effec- tive, the approach faces challenges with dependency on sentence boundaries, warranting further investigation for
enhanced performance and robustness.

[7] Moawad and Aref propose a Semantic Graph Re- duction approach for Abstractive Text Summarization, demonstrated through a case study involving 7 sentences. The method involves generating a Rich Semantic Graph (RSG) representing document semantics, then reducing it using heuristic rules and WordNet relations to produce the final summary. Successfully reducing documents by 50%, the approach exhibits versatility across varied document sizes, showcasing its potential for effective summarization.

[8] Miller from Georgia Institute of Technology explores Extractive Text Summarization on lecture transcripts using BERT embeddings and K-Means clustering. Lever- aging BERT N-2 layer embeddings and clustering, the approach outperforms traditional methods like TextRank, offering a promising avenue for summarizing educational content effectively.

[9] Alomari et al. review Deep Reinforcement and Trans- fer Learning for Abstractive Text Summarization, focusing on CNN/DM dataset. They highlight challenges in defining rewards for RL and fine-tuning in TL, with

SimCLS standing out for short-document summarization across multiple datasets, indicating its efficacy in addressing summarization challenges.
[10] Nallapati, Zhou, and dos Santos present Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond, employing datasets such as Gigaword, DUC, and CNN/Daily Mail. Utilizing Bidirectional GRU-RNN and attention mechanisms, the approach shows improved Rouge scores, particularly with the switching genera- tor/pointer model on the Gigaword dataset.
[11] Suleiman and Awajan delve into Deep Learning Based Abstractive Text Summarization, exploring var- ious models such as pointer-generator networks, deep reinforced models with RL, and transformer-based mod- els. Their approach incorporating a pretrained encoder, achieves the highest ROUGE scores, underscoring the effectiveness of deep learning in abstractive summarization tasks.
[12] Akhil et al. explore Parts-of-Speech tagging for Malayalam using deep learning techniques, achieving the highest F-measure with a Bi-directional LSTM (BLSTM) model. Their study showcases the effectiveness of BLSTM with 64 hidden states in comparison to other architectures like RNN, GU, and LSTM.

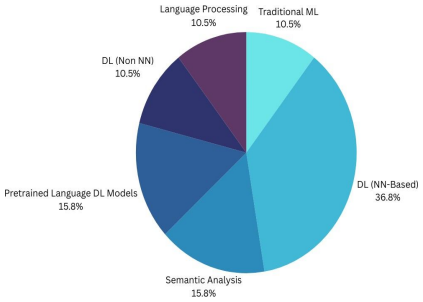

Figure 1: Methods used

III. Motivation

The explosion of digital content, particularly in the form of news articles, has created a pressing need for efficient methods of information extraction. Abstractive text summarization offers a solution to this challenge by condensing lengthy articles into concise summaries, enabling users to quickly grasp key insights. Additionally, the ability to generate summaries in multiple languages enhances accessibility and usability, making this research highly relevant and impactful.

IV. Problem Domain

The problem domain revolves around the domain of natural language processing, specifically abstractive text summarization, with a focus on news articles. This domain encompasses various challenges such as understanding the context, capturing the essence of the article, and ensuring fluency and coherence in the generated summaries.

V. Problem Definition

The enormous amount of textual material that is avail- able in the modern digital age presents a major obstacle for users looking for rapid and insightful answers. Conventional methods of summarising frequently fail to convey the complexity and context found in complicated texts.
The worldwide nature of information transmission drives the choice to implement multi-language support, with the goal of removing language barriers and increasing the accessibility of quality content to a wider audience.

VI. Statement

This initiative addresses the challenge of extracting crucial insights from news articles. It aims to facilitate this process for users from diverse linguistic backgrounds.

VII. Innovative Content

Our project offers extensive multilingual support specifically designed for news articles. Leveraging the power of state-of-the-art natural language processing models, we provide seamless translation of summarized news

content into multiple languages. This unique feature is tailored to cater to the diverse needs of our users, facilitating cross- cultural communication, global outreach, and access to news articles in languages of their choice.

VIII. Problem Formulation Or Representation Or Design

A. SEQ2SEQ MODEL

1) Design Methodology

•Model Architecture

–Encoder-Decoder Structure: The Seq2Seq model consists of an encoder and a decoder. The encoder processes the input text and encodes it into a fixed-length vector representation. The decoder then generates the summary based on this representation.

–Embedding Layer: Both the encoder and decoder include embedding layers, which map words to dense vectors. This helps the model learn seman- tic representations of words.

– *LSTM Layers*: Long Short-Term Memory (LSTM) layers are used in both the encoder and decoder to capture temporal dependencies and handle sequences of variable length.
–*Dense Layer*: The decoder includes a dense layer with softmax activation to generate the output sequence.

• *Training Process*

– *Tokenization and Padding*: The input text and summaries are tokenized, and padding is applied to ensure uniform sequence length.
–*Model Compilation*: The model is compiled with appropriate loss function (sparse categorical cross-entropy) and optimizer.
–*Training*: The model is trained on the training data, with early stopping to prevent overfitting.

•Inference

–*Decoding Sequence*: During inference, the trained model is used to generate summaries for new input text.
–*Greedy Decoding*: The model uses greedy decod- ing to generate the most likely summary se- quence.

2) Mathematical model

•*Encoder Architecture* The encoder takes the input sequence X = (x1, x2, ..., xT) and produces a sequence of hidden states H = (h1, h2, ..., hT). Each hidden state ht is computed as follows:

$$ht = \text{Encoder LSTM}(xt, ht{-}1)$$

•*Decoder Architecture* The decoder takes the hidden states H from the encoder and generates the output sequence Y = (y1, y2, ..., yT′). Each output token yt′ is computed as follows:

–*Attention Mechanism*

Compute attention scores eti between the de- coder hidden state st and each encoder hidden state hi:

$$eti = \text{Attention}(st, hi) = vT \tanh(Wa[st; hi]+ba)$$

∗ Calculate attention weights αti using softmax function:

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{j=1}^{T} \exp(e_{tj})}$$

∗ Compute the context vector ct as the weighted sum of encoder hidden states:

$$c_t = \sum_{i=1}^{T} \alpha_{ti} h_i$$

– Decoder LSTM Cell Update the d sing the previous hidden state st 1, the previous output token yt 1, and the context vector ct:

$$st = \text{DecoderLSTM}(yt{-}1, st{-}1, ct)$$

– Output Layer Compute the logits yˆt for the next token using the decoder hidden state st:

$$\hat{y}t = \text{softmax}(Wost + bo)$$

• *Loss Function* The loss function used for training the Seq2Seq model is the sparse categorical cross-entropy loss, which measures the dissimilarity between the predicted probability distribution Yˆ and the true distribution Y . Given the true target sequence Y = (y1, y2, ..., yT′), the loss is computed as:

$$\text{Loss} = -\frac{1}{T'}\sum_{t=1}^{T'}\sum_{i=1}^{N} y_{ti}\log(\hat{y}_{ti})$$

where N is the vocabulary size and yti is the one-hot encoded vector for the true token at time step t.

B. T5 (Text-To-Text) MODEL

1) Design Methodology

• *Data Module Creation*: The NewsSummaryData- Module class is designed to create data loaders for training, validation, and testing. It encapsulates the training and testing datasets and provides Dat- aLoader instances for each split.

• *Model Architecture*: The NewsSummaryModel class defines the architecture of the T5- based summarization model. It utilizes the T5ForConditionalGeneration model from the Hugging Face Transformers library. The model is structured to take input text and generate summaries using the T5 architecture.

• *Training and Evaluation*: The Lightning Trainer from PyTorch Lightning is employed for training the model. During training, the model's performance is monitored using metrics such as loss. Additionally, the model is evaluated on a separate test set using metrics like BLEU, GLEU, and ROUGE scores.

2) Mathematical Model

• *Tokenization*

$$\text{input\_ids, attention\_mask} = \text{Tokenizer(text)}$$

• *Encoder-Decoder*

$$\text{en\_output} = \text{Encoder(input\_ids, attention\_mask)}$$
$$\text{de\_output} = \text{Decoder(en\_output, target\_ids)}$$

• *Attention Mechanism*

$$\text{attn\_weights} = \text{Attention(en\_output, de\_output)}$$

• *Loss Function*

$$\text{LCE} = -\text{target\_ids}_i \log(\text{softmax(de\_output)}i)i$$

• *Optimization*

$$\text{updated\_params} = \text{Optimizer(LCE)}$$

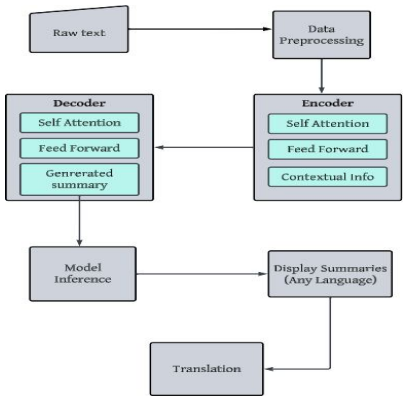

Figure 2: Proposed system T5 model

C. MULTILINGUAL ACCESSIBILITY

1) Translation Functionality: Leveraging state-of-the- art translation APIs, such as Google Translate, the module seamlessly translates the summarized text into the selected target language. The translated text is then displayed to the user, facilitating better comprehension and accessibility.

2) Error Handling: In the event of any translation er- rors or issues, the module provides informative feedback to the user, ensuring transparency and reliability throughout the translation process:

D. Evaluation Metrics

1) ROUGE (Recall-Oriented Understudy for Gisting Evaluation)

• *ROUGE-1*: Measures the overlap of unigram (individual words) between the generated summary and reference summary.

$$\text{ROUGE-1} = \frac{\text{Overlapping Unigrams}}{\text{Total Unigrams in Reference Summaries}}$$

• *ROUGE-2*: Measures the overlap of bigrams (pairs of consecutive words) between the generated summary and reference summary.

$$\text{ROUGE-2} = \frac{\text{Overlapping Bigrams}}{\text{Total Bigrams in Reference Summaries}}$$

• Measures the longest common subsequence between the generated summary and reference summary.

$$\text{ROUGE-L} = \frac{\text{Longest Common Subsequence (LCS)}}{\text{Total Words in Reference Summaries}}$$

2) BLEU (Bilingual Evaluation Understudy)

• Measures the similarity between the generated summary and reference summary based on n-gram precision.

$$\text{BLEU} = \text{BP} \times \exp\left(\sum_{n=1}^{N} \frac{1}{N} \log \text{precision}_n\right)$$

where BP is the Brevity Penalty, and precisionn is the n-gram precision.

3) GLEU (Google-BLEU):

• Evaluates the fluency and grammaticality of the generated summary.

$$\text{GLEU} = \left(\frac{\sum_{n=1}^{N} \min(\text{Coun}_n, \text{Ref}_n)}{\sum_{n=1}^{N} \text{Coun}_n}\right)^{\frac{1}{N}}$$

where Counn is the count of n-grams in the candidate summary, and Refn is the count of n-grams in the reference summary.

IX. Results And Sensitivity Analysis

A. Experimental Results

We conducted experiments using both the seq2seq and T5 models for abstractive text summarization on a dataset of news articles that contains the headlines and the respec- tive summaries for it. The performance of each model was evaluated using the below metrics.

B. Sensitivity Analysis

To assess the robustness of our models, we conducted sensitivity analysis by varying parameters such as batch size, learning rate, and model architecture. The results indicate that both models exhibit stable performance across different parameter settings, with the T5 model demonstrating slightly better resilience to variations in hyperparameters. =

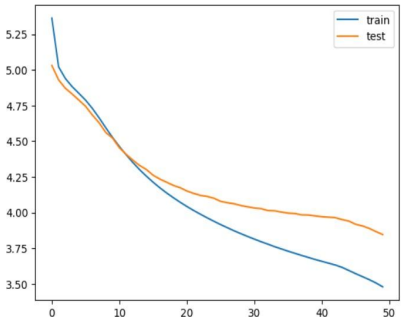

Figure 3: Epochs vs Validation Loss

TABLE I: COMPARISON OF RESULTS

Measure	Seq2Seq model	T5
ROUGE-1	0.2392	0.3804
ROUGE-2	0.0142	0.1732
ROUGE-L	0.2371	0.3455
BLEU	0.2327	0.1276
GLEU	0.3143	0.1536

TABLE II: ORIGINAL TEXT, SUMMARY, AND TRANSLATED SUM- MARY

Original Text	Summary
A CRPF jawan was on Tues- day axed to death with sharp- edged weapons by Maoists at a local village fair in Chhattis- garh's insurgency-hit Bijapur district. As per preliminary information, Maoists attacked the jawan while he had gone to visit the fair along with his family, a police official said. A combing operation was launched to nab the assailants, he added.	a CRPF jawan was attacked by Maoists at a local village fair. he was killed with sharp-edged weapons. a combing operation was launched to nab the as- sailants.
Translated Summary	
एक सीआरपीएफ जवान पर एक स्थानीय गाँव के मेले में माओवादियों द्वारा हमला किया गया था।वह तेज धार वाले हथियारों से मारा गया था।हमलावरों को नाब करने के लिए एक कं घी ऑपरे शन शुरू किया गया था।	

X. DATA MODEL

The system utilizes two primary datasets: ”newssum- mary.csv” and ”newssummarymore.csv”. Together, these datasets contain a total of 102,918 rows, providing a diverse range of news articles and their corresponding summaries. Each dataset includes essential information such as author, date, headlines, full text, and summary, facilitating comprehensive training and evaluation of the abstractive text summarization model.

XI. COMPARISON OF RESULTS

Figure 4: Comparison of metrics

XII. JUSTIFICATION OF THE RESULTS

Based on the provided results, it is evident that the T5 model outperformed the seq2seq model across ROUGE evaluation metrics, achieving higher ROUGE scores for ROUGE-1, ROUGE-2, and ROUGE-L.
However, it is important to note that the BLEU and GLEU scores were lower for the T5 model compared to the seq2seq model. BLEU and GLEU metrics provide valuable insights into the grammatical correctness of the generated text, they may not fully capture the semantic accuracy and informativeness of the summary, which are crucial aspects of text summarization tasks.

On the other hand, ROUGE metrics are specifically designed for evaluating text summarization tasks, taking into account the overlap of content units between the generated summary and the reference summary. As such, ROUGE metrics are more aligned with the objectives of text summarization and are widely used in the literature for this purpose.
Therefore, this can be attributed to the T5 model's ability to leverage pre-trained knowledge and fine-tuning capabilities, enabling it to generate summaries that better reflect the content and structure of the input articles.
Furthermore, the language customization feature of the T5 model enables users to obtain summaries in their preferred language, enhancing accessibility and usability.

XIII. Conclusion

In conclusion, this paper presents a comprehensive in- vestigation into abstractive text summarization for news articles using deep learning techniques. We have demon- strated the effectiveness of Seq2Seq and T5 models in generating concise and coherent summaries, with the T5 model showing superior performance. The innovative feature of language customization enhances the usabilityof the summarization system, making it accessible to users worldwide.

Future Work

We could develop a system that analyzes the content of news articles and categorizes them into relevant topics or themes, such as politics, sports, technology, etc.
Additionally, we could introduce a personalized recom- mendation system that suggests news articles to users based on their interests, reading history, and preferences.

References

[1] Rahul, S. Adhikari and Monika, "NLP based Machine Learn- ing Approaches for Text Summarization," 2020 Fourth Inter- national Conference on Computing Methodologies and Com- munication (ICCMC), Erode, India, 2020, pp. 535-538, doi: 10.1109/ICCMC48092.2020.ICCMC-00099.

[2] Widyassari, Adhika & Rustad, Supriadi & Shidik, Guruh & Noersasongko, Edi & Syukur, Abdul & , Affandy & Se- tiadi, De Rosal Ignatius Moses. (2020). Review of Automatic Text Summarization Techniques & Methods. Journal of King Saud University - Computer and Information Sciences. 34. 10.1016/j.jksuci.2020.05.006.

[3] N. Giamblanco and P. Siddavaatam, "Keyword and Keyphrase Extraction using Newton's Law of Universal Gravitation," 2017 IEEE 30th Canadian Conference on Electrical and Computer Engineering (CCECE), Windsor, ON, Canada, 2017, pp. 1-4, doi: 10.1109/CCECE.2017.7946724.

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

[5] Allahyari, M. (2017, July 7). Text Summarization Techniques: A Brief Survey. arXiv.org. https://arxiv.org/abs/1707.02268

[6] Abu Nada, Abdullah and Alajrami, Eman and Alsaqqa, Ahmed and Abu-Naser, Samy. (2020). Arabic Text Summarization Using AraBERT Model Using Extractive Text Summarization Approach.

[7] I. F. Moawad and M. Aref, "Semantic graph reduction ap- proach for abstractive Text Summarization," 2012 Seventh In- ternational Conference on Computer Engineering and Systems (ICCES), Cairo, Egypt, 2012, pp. 132-138, doi: 10.1109/IC- CES.2012.6408498.

[8] Miller, D. (2019, June 7). Leveraging BERT for Extractive Text Summarization on Lectures. arXiv.org. https://arxiv.org/abs/1906.04165

[9] Ayham Alomari, Norisma Idris, Aznul Qalid Md Sabri, Izzat Alsmadi, Deep reinforcement and transfer learning for abstractive text summarization: A review,Computer Speech and Language,Volume 71,2022,101276,ISSN 0885-2308,https://doi.org/10.1016/j.csl.2021.101276.

[10] Nallapati, R. (2016, February 19). Abstractive Text Summariza- tion Using Sequence-to-Sequence RNNs and Beyond. arXiv.org. https://arxiv.org/abs/1602.06023

[11] Dima Suleiman, Arafat Awajan, "Deep Learning Based Ab- stractive Text Summarization: Approaches, Datasets, Evalu- ation Measures, and Challenges", Mathematical Problems in Engineering, vol. 2020, Article ID 9365340, 29 pages, 2020. https://doi.org/10.1155/2020/9365340

[12] K K, Akhil and R., Rajimol and S., Anoop. (2020). Parts- of-Speech tagging for Malayalam using deep learning tech- niques. International Journal of Information Technology. 12. 10.1007/s41870-020-00491-z.