Comparative Analysis of Different Text Summarization Models

Vijayshri Khedkar[1], Vedant Deshmukh[1], Priyanka Iyer[1], Ruchira Lokhande[1], Nevil Tanna[1], and Sonali Kothari[1]

[1]*Department of Information Technology, Symbiosis Institute of Technology, Pune, India*
E mail: vijayshri.khedkar@sitpune.edu.in

Abstract

Text summarization has long been a subject of conversation in academics. Despite the fact that various strategies for automatic text summarization have been developed in recent years, efficiency remains a challenge. The ever-increasing bulk of textual content needs the development of a method to retain the information in a condensed manner with little information loss. Given the increase in the size and number of papers available online, an efficient automatic news summarizer is an absolute necessity. This study proposes a pipeline of procedures for generating lossless summaries. There are 2 types of summaries: extractive & abstractive summaries. The extraction method identifies and extracts only relevant sentences from the original document. Abstractive summarization techniques, on the other hand, create the summary after knowing the source text, which makes it more complex. This study compares some transformers architecture based pre-trained models for summarization of text.

Keywords - Abstractive Text Summarization, BART, PEGASUS, ROUGE, Transformers, T5.

1. INTRODUCTION

If no one reads your work, it makes no difference how much information you include. Any summary can help a reader decide whether or not the subject is worth learning more about. Text summarization is a time-saving approach that may be combined with data extraction and filtering software. It is necessary to condense textual data into shorter, focused summaries containing crucial features to travel through it more effectively. Despite the fact that numerous methodologies for news summarization have been developed over this time, absolute efficiency has yet to be achieved. The amount of today's information repository is significantly bigger than one can readily and efficiently handle. Business transactions, news stories, satellite data, digital media, written reports and memoranda, and biological data are all examples of this. Furthermore, in current times, everyone wants to get more and more in a shorter amount of time. Therefore, reading large texts and then trying to understand them is not a smart idea. It is more worthwhile to read the synopsis of a large text while retaining the topic or key information provided within it. As a result, greater and greater data can be acquired in less time. The need for efficient data mining methods is increasing by the day. Hence, devising a technique for automatic text summarising that is not just time-saving for the reader, but also efficient, accurate, and feasible.

2. LITERATURE SURVEY

The research by Tadashi Nomoto provides a Bayesian model for text summarizing that explicitly encodes and uses information about how human judgements are dispersed across the text. Using test data from Japanese news texts, a comparison to non-Bayesian summarizers is done [1]. Amir

Jalilifard et. al propose STF-IDF, a unique connotation technique build using TF-IDF, is suggested for rating word weightage in a corpus of natural writings [2]. Mike Lewis et. al propose BART, a de - noising autoencoder used for the pre defined training of sequence-to-sequence models, is presented. BART is learned as first distorting the textual data with a random noising algorithm and later generating the model so as to retrieve the actual text. BART makes use of standard neural network translation architecture, which is based on transformers that, although being straightforward, may be thought of as generalizing BERT, GPT and then a variety of alternative recent pretraining approaches [3] Colin Raffel et. al in their paper delves into the area of transfer learning methodologies, by presenting single scheme which converts all text-based linguistic challenges in a text-to-text form. On hundreds of language understanding tasks, our structured research examines pretraining goals, model architectures, transfer methodologies, and many more other parameters [4]. Jingqing Zhang et. al introduce pre-training huge encoder-decoder models based on transformers, on "vast text corpora with a novel self-supervised aim. PEGASUS extracts/masked essential sentences deriving out of an input file & generates them like a single output series deriving out of the other phrases, comparable with an extractive type of summary. They put PEGASUS through its paces on 12 subsequent summarization exercises. Experiments show it delivers cutting-edge results on all the 12 datasets. The conclusion was formed on the basis of obtained ROUGE scores [5-8]. The final model developed by Lucy Vanderwende and Aria Haghighi, called HIERSUM, displays the contents particularly as a ranking of subject vocabulary arrangement. It was also suggested that HIERSUM may provide several "topical summaries" to facilitate content browsing and discovery [9-12].

Objectives

To identify and convey the most significant information from a particular text to end users.

To focus on data relevance, maximum information completeness, minimum information redundancy, and summary coherence.

To analyse a model generated summary against a set of reference materials using ROUGE Score.

To propose utilizing the Transformer paradigm to reevaluate NLP jobs The inputs & outputs will be strings of text in sus paradigm.

3. MATERIALS AND METHODOLOGY

Dataset

The dataset utilized in this study for comparing the transformer models is the CNN-DailyMail dataset. A little over 3,000,000 unique news stories in English, published by CNN and Daily Mail writers, make up this dataset. The data fields contain 3 columns:

- ID - It includes a string that contains the SHA1 hash of the URL where the article was retrieved, formatted in hexadecimal.
- Articles - The article's body is contained in the article column.
- Highlights - The highlight of the piece, as written by the author, is contained in the Highlights column.

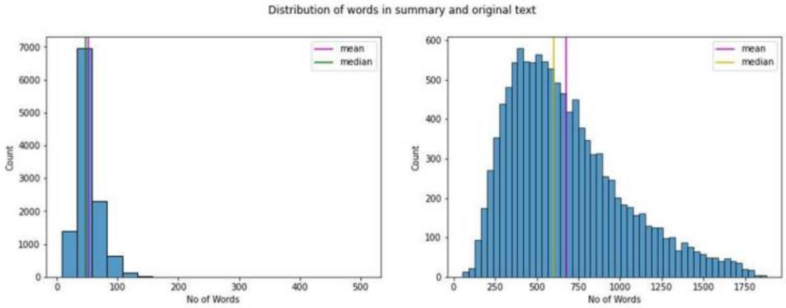

Figure 1: Distribution of the Words in the original article and the summary

The Basic Transformer Model

The various attention levels that comprise a transformer system for text summarization represent its fundamental foundation. It is based on attention layers & directional encoding to recall the sentences in an input pattern. The overall global reliance generated by using several attention layers help and assist in the concurrent computation of input pre-processing [13].

Encoder Block:

Computers do not understand words. Computers, on the other hand, deal with matrices, vectors, or integers. As a result, one must turn the words into vectors. By utilizing the embedding space for this, which is similar to an open area or a dictionary in that words with similar meanings are clustered together [14]. Each word in this system is mapped and assigned a value depending on its meaning. As a result, the encoder block convert our words into vectors. Positional encoders in this block provide context based on where a word is in a phrase. This concludes our input, which is subsequently forwarded to the encoder block.

Multi-headed Attention Layer:

It is up to the reader to determine how essential a word is in relation to the other words in a sequence of words. Relative position of words is much important in linguistic models so as to make sense out of the sentence. It is exhibited as an attention vector. Regardless of the fact that each word in the phrase has significantly more weight, this research work is interested in how each word in that sentence interacts with the other phrases. As a result, this layer compute numerous attention vectors for each word before utilizing a weighted average to calculate the final attention vector for each word. This strategy is also known as the multi-head attention block since the layer is employing many attention vectors. In decoder & encoder layers, layer having multi-head attention employs a process known as self-attention. Value, key & query vectors are generated from the inputs by routing it into 3 internally connected layers. Then 'n vectors' are formed by the division of these 3 vectors.

$$Attention = \text{softmax}\left(\frac{\mathrm{QK}^t}{\sqrt{d_k}}\right)V$$

Feed forward Layer:

Each attention vector is subjected to a basic feed-forward network. The network of digital neurons turns information into format which is suitable for the next encoder or decoder layer.

Multi-headed Masked Attention Layer:

Attention in the Masked Multi-Head Attention Layer is focused on tokens up to the current position (index till which the transformer forecasts) rather than future tokens (which have not yet been predicted).

Linear Layer:

A Logits vector is a large vector formed from decoder stack. A fully connected neural network forms a linear layer which does this job of converting the vectors.

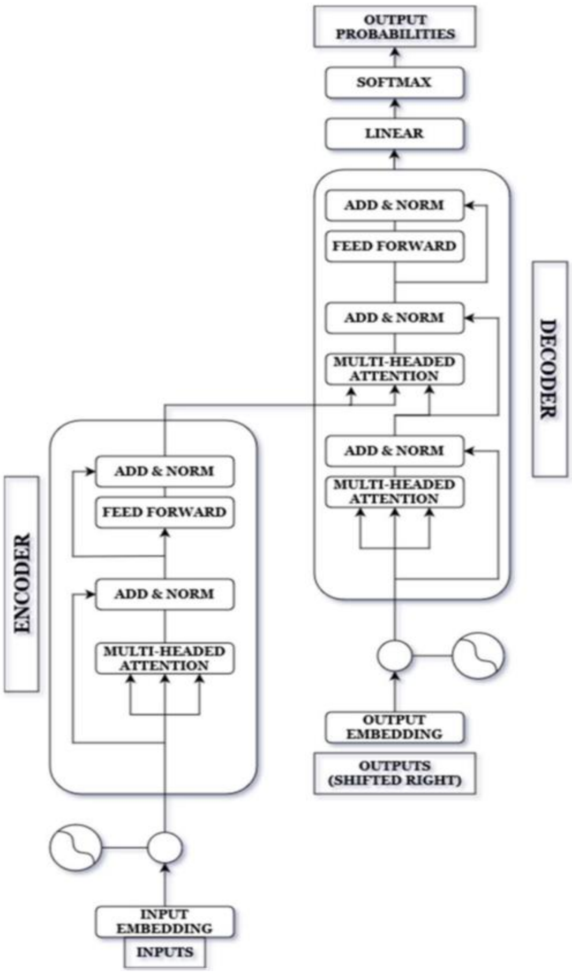

Figure 2: Transformer Architecture

SoftMax Layer:

The SoftMax layer transforms the input to a probability distribution that can be interpreted by humans. The three transformers T5, BART and PEGASUS are used for finetuning on our dataset.

A. T5:

The Text-to-Text Transfer Transformer is abbreviated as T5. T5 paradigm is based on a concept known as Transfer learning. After being trained in Transfer Learning on a task with a significant amount of text, this T5 model was finetuned on a downstream task to obtain broad day-to-day language abilities and knowledge that could be used for tasks like text summarization [15]. This transformer model utilizes a sequence-to-sequence generation method, where encoded input is delivered to the decoder layer via the model's cross-attention layers [16]. The decoder's output is autoregressive, which means that it will report the next future words, values on where encoder receives tokens in sequential manner as inputs & translates those to a collection of embeddings.

B. BART:

BART is a model based on sequence-to-sequence denoising autoencoder. It makes use of a standard seq2seq/NMT structure alongside bidirectional encoder & left-to-right decoder [3]. It implies that a BART model which is finely-tuned on a dataset may have an input text series & create new text sequence from it. BART combines both the concepts from GPT and BERT, being bidirectional like BERT, and the auto regressive one like GPT. The argument or the thinking behind it is that BERT's bidirectional nature, which is the auto encoder paradigm is beneficial for some NLP works like classification, which require details and data about complete sentence [17]. Therefore if one have classification tasks, it's not necessarily an advantage to predict one word at a time, it's more about understanding the whole the whole sentence at once in a sense. The downside of something like BERT is it's is poor at handling generation NLP work, where the generated word should only hnag on to the formerly prediscted word [18].

PEGASUS:

Pre-training with Extracted Gap-sentences for Abstractive Summarization Sequence-to-Sequence Models is abbreviates as PEGASUS [5]. PEGASUS by Google improves the state-of-the-art (SOTA) outputs for the abstractive mode of summarization, particularly with limited supply of computational power, by utilizing prior researched and finding in NLP [19]. To be more specific, PEGASUS, unlike previous models, allows us to get results that are almost equivalent to SOTA utilizing 1,000 samples rather than hundreds of thousands of training sets [21]. To teach sequences from sequences, PEGASUS employs an encoder-decoder framework.

Table 1: Models along with their parameters and checkpoints used

Models	Parameters	Checkpoints
T5	11 Billion	T5 – Base
BART	140 Million	BART – large-cnn
PEGASUS	568 Million	PEGASUS – cnn_dailymail

4. EVALUATION AND RESULTS

ROUGE is an abbreviation that means Recall-Oriented Understudy for Gisting Evaluation. It provides ways to evaluate the excellence of a synopsis mechanically by matching it to the rest of the (ideal) summaries generated from individuals. This metrics counts amount of overlying subsequence, like n-grams, text patterns, & pairs of words, that exist in computer generated summaries & the optimal summary by humans [21].

ROUGE metrics are classified into various segments, like the ROUGE-1, ROUGE-2, ROUGE-L & others.

ROUGE-1 distinguishes uni-grams in the computer-generated & manual referential summaries.

ROUGE-2 relates to bigram overlapping between the system and referential synopsis.

ROUGE-L calculates the sequence of words with longest matching subsequence with the help of Longest Common Subsequence (LCS). LCS seems to provide the ease of just needing in-sequence comparisons that capture sentence level grammatical structure, rather than consecutive matches. Since it mechanically includes the greatest in-sequence related n-grams, no specified n-gram length is required.

Table 2: ROUGE Scores for Transformer Models

Models	ROUGE - 1	ROUGE - 2	ROUGE - L
PEGASUS	0.013042	0.000968	0.000968
BART	0.376068	0.069565	0.290598
T5	0.487804	0.247933	0.406504

5. DISCUSSION

This research work used transformers fruitfully and examined them by making use of a standard evaluation criterion ROUGE. Our research led us to the outcome that finely calibrated transformers layered on top of previously trained language models generated tremendous success and a logical and flowing summary of a specified text's material. For comparative purposes, this research derived ROUGE scores for every model's forecasting and deduced that the T5 model outperformed every other model, notably BART and PEGASUS. Based just on the ROUGE score, the study may infer that T5 comes out on top, accompanied by BART and then PEGASUS. In the future, attention should be focused on constructing higher trustworthy models. The Transformer model could be employed to build more efficient models that provide accurate and clear summaries and appear more genuine, and human-generated. A combination of the aforementioned models, as well as hybrids, could be employed to enhance the precision, readability, and clarity of the summaries

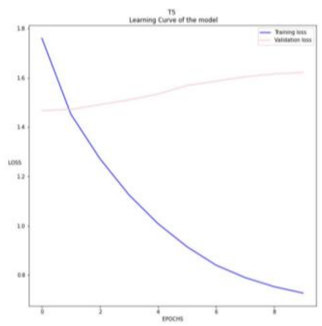

Figure 3: Training Loss vs Validation Loss curve for T5

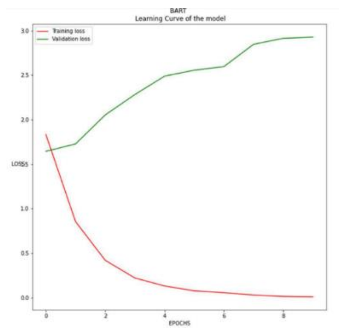

Figure 4: Training Loss vs Validation Loss curve for BART

6. REFERENCES

[1] N. Tadashi. 2005. Bayesian Learning in Text Summarization. In Proceedings of Human Language Technology Conference, pages 249– 256, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

[2] A. Jalilifard, V. F. Caridá, A. F. Mansano, R. S. Cristo and F. P. da Fonseca. (2020). Semantic Sensitive TFIDF to Determine Word Relevance in Documents. arXiv. https://doi.org/10.1007/978-981-33-6977-1

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O.Levy, V. Stoyanov, & L. Zettlemoyer. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv. https://doi.org/10.48550/arXiv.1910.13461

[4] Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., & Liu P. J. (2019). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. arXiv. https://doi.org/10.48550/arXiv.1910.10683

[5] J. Zhang, Y. Zhao, M. Saleh, & P. J. Liu (2019). PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. arXiv. https://doi.org/10.48550/arXiv.1912.08777

[6] L. Wang, C. Cardie. (2013) Domain-independent abstract generation for focused meeting summa-rization. In: Proceedings of the 51st annual meeting of the association for computational linguistics. Association for Computational Linguistics, Sofia, Bulgaria, pp. 1395–1405

[7] C. Li, X. Qian, Y. Liu. (2013) Using supervised bigrambased ILP for extractive summarization.In: Proceedings of the 51st annual meeting of the association for computational linguistics.ACL, Sofia, Bulgaria, pp. 1004–1013

[8] X. Qian, Y. Liu. (2013) Fast joint compression and summarization via graph cuts. In: Proceed-ings of the conference on empirical methods in natural language processing. ACL, Seattle,Washington, USA, pp. 1492– 1502

[9] A. Haghighi. and L. Vanderwende. 2009. Exploring Content Models for Multi-Document Summarization. In Proceedings of Human Language Technologies: The 2009 of the North American Chapter of the Association for Computational Linguistics, pages 362–370, Boulder, Colorado. Association for Computational Linguistics.

[10] A. Gupta, & D. Chugh, A. Anjum, & R. Katarya. (2021). Automated News Summarization Using Transformers.

[11] P. Bhattacharjee, A. Mallick & M. S. Islam (2020). Bengali Abstractive News Summarization (BANS): A Neural Attention Approach. arXiv. https://doi.org/10.1007/978-981-33-4673-4_4

[12] Y. Liu and M. Lapata, "Text Summarization with Pretrained Encoders," in the 2019 Conference on Empirical Methods in Natural Language Processing and The 9th International Joint Conference on Natural Language Processing, 2019, pp. 3728–3738.

[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin (2017) Attention is All you Need. arXiv preprint arXiv:1706.03762.

[14] Z. Albert, P. Bahar, K. Irie, R. Schluter,& H. Ney (2019). A Comparison of Transformer and LSTM Encoder Decoder Models for ASR. 8-15. 10.1109/ASRU46091.2019.9004025.

[15] A. Mastropaolo. Studying the Usage of Text-To-Text Transfer Transformer to Support Code-Related Tasks. (2021) IEEE/ACM 43rd International Conference on Software Engineering (ICSE), 2021, pp. 336-347, doi: 10.1109/ICSE43902.2021.00041.

[16] K. Grover, K. Kaur, K. Tiwari, R. Goyal & P. Kumar (2021). Deep Learning Based Question Generation Using T5 Transformer. 10.1007/978-981-16-0401-0_18.

[17] B. Dang, D. -T. Do and L. -M. Nguyen (2022), BART: Abstractive summarization based on the joining of Topic modeling and BART. 14th International Conference on Knowledge and Systems Engineering (KSE), 2022, pp. 1-6, doi: 10.1109/KSE56063.2022.9953613.

[18] H. Batra, A. Jain, G. Bisht, K. Srivastava, M. Bharadwaj, D. Bajaj, U. Bharti (2021). CoVShorts: News Summarization Application Based on Deep NLP Transformers for SARS-CoV-2. 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), pp. 1-6, doi: 10.1109/ICRITO51393.2021.9596520

[19] T. H. Yang, C. C. Lu and W. L. Hsu (2021) More than Extracting "Important" Sentences: the Application of PEGASUS. International Conference on Technologies and Applications of Artificial Intelligence (TAAI), 2021, pp. 131-134, doi: 10.1109/TAAI54685.2021.00032.

[20] A. Sahu and S. G. Sanjeevi (2021). Better Fine-Tuning with Extracted Important Sentences for Abstractive Summarization. International Conference on Communication, Control and Information Sciences (ICCISc), 2021, pp. 1-6, doi: 10.1109/ICCISc52257.2021.9484944.

[21] C. W. Lin. (2004). ROUGE: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10.

Biographies

Vijayshri Khedkar is an Assistant Professor working at Symbiosis Institute of Technology and skilled in NLP, Applied Machine Learning, Data Analytics, Information Retrieval & Deep Learning. A life-long learner with a strong educational background holding two Master's Degrees (M.B.A. & M.E.) and pursuing Ph.D. in Computer Engineering (NLP) from Symbiosis International University, India. She is a life member of IAENG and Senior IEEE member. She has published around 30 papers at various reputed conferences and journals. She is active reviewer for IJECE journal indexed in Scopus and reviewed papers for various conferences and journals.

Vedant Deshmukh is an Information Technology student pursuing his B. Tech at Symbiosis Institute of Technology, Pune. His areas of interest include Cybersecurity, Deep Learning and NLP. He actively participates in volunteering activities and is also the Events Head of the ISR club of his college. Being a team player, he is good at leadership and managing people. He is an avid learner and is always looking for opportunities to utilize his skills for the betterment of society.

Priyanka Iyer is a Final Year Student at Symbiosis Institute of Technology pursuing her B.Tech degree in Information Technology. She is also studying Artificial Intelligence and Machine Learning as a major specialization in addition to Information Technology. She is also a core member of one of the clubs in her college. She is a well-organized and efficient individual who can encourage and direct her abilities and skills to achieve goals. She is eager to learn new skills and receive training, and she is adaptable in the workplace.

Ruchira Ravi Lokhande is an Information Technology student pursuing her B. Tech Degree from Symbiosis Institute of Technology, Pune. In addition to this she is pursuing a major specialization in the field of Artificial Intelligence and Machine Learning. She is dedicated towards gaining more knowledge in the fields of NLP, AI and ML. Along with this she is also the Head of Content and Research in three college clubs. She has good leadership qualities and aims to utilize her interpersonal skills to gain more knowledge and experience.

Nevil Tanna is a final year student pursuing his B.Tech at Symbiosis Institute of Technology, Pune, majoring in Information Technology. Along with his B.Tech, he has also completed a diploma in business management. He is also leading the ISR club of his college and core member of two other clubs.

Dr. Sonali Kothari is working as Associate Professor in Department of Computer Science and Engineering at Symbiosis Institute of Technology, Symbiosis International (Deemed University), Pune. She has completed her PhD in computer Engineering specialization from Sant Gadge Baba Amravati University, Amravati, She has published 35+ research articles in various International/National conferences and journals. She has more than 20 years of teaching and research experience. She is senior member of IEEE, Life member ISTE.