International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 10 | October - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

NLP BASED TEXT SUMMARIZATION USING BART MODEL

Ms. Sangavi N[1], Ms. Umamaheswari M[2], Ms. Subasri V[3]

[1]Assistant Professor Level – I, Computer Science and Engineering & Bannari Amman Institute of Technology
[2] UG Scholar, Electrical and Electronics Engineering & Bannari Amman Institute of Technology
[3] UG Scholar, Computer Science and Engineering & Bannari Amman Institute of Technology

---

Abstract - Text summarization plays a major role which can be used to summarize articles, papers, documents, etc. It greatly helps the students, business executives, librarian, teachers and many people in different profiles to quickly get the gist of research papers, latest news, trends in industry, summarize books, articles, etc. Summarizing provides a concise overview of a document which helps us to make better decisions and communicate information more effectively. In this project we are going to implement an nlp based text summarization using bart model and finally deployed using django framework. Bart hugging face transformer model is one of the nlp models. Its input and output are in the form of sequence. The high-dimensional representation of the input is learnt by the encoder and then mapped to the decoder. The bart model is first pre-trained on a big text (like book corpus or wikipedia). Pretraining ensures that the model "understands the language" and has a solid foundation from which to learn and how to carry out further tasks. Since it will determine how successfully the model can be trained for tasks like text classification or text summarization, the model's capacity to understand language is more effective. The pre-trained weights and weights in the bart model are fine-tuned on question answering, text summarization, sequence classification, etc. Therefore, using bart model we can able to generate text in both extractive and abstractive way, by which we can get the simpler summarization with the important sentences and also more fluent and accurate summarization by understanding the meaning of the text. Finally, to provide a user-friendly interface, by which the users can easily use the text summarization, the project will be deployed using django framework. It allows us to easily build web applications.

Key Words: *BART, NLP, summarization, Wikipedia, pretrained*

1.INTRODUCTION

Natural Language Processing (NLP) has developed significantly over the past several years, completely altering how we interact with and comprehend textual information. The Bidirectional Encoder Representations from Transformers (BERT) and its variations are among the numerous NLP models that have become quite popular for their capacity to understand the context of words in a phrase. However, BERT models by themselves are not the best choice for text summary. It is at this point that the BART (Bidirectional and Auto-Regressive Transformers) model is used.

The BART model is a potent and adaptable transformer architecture developed for a variety of NLP applications, including text summarization, and it was launched by Facebook AI. It combines the advantages of auto-regressive and bidirectional transformers for text generation and context awareness. BART has the special ability to comprehend the original material and provide logical summaries, which makes it a prime contender for text summarizing. We explore the BART model's capabilities and possible uses in this project as we dig into the area of NLP-based text summarization. It is particularly skilled at producing high-quality textual content summaries thanks to the design of the model, which is closely connected to the Transformer model and combines both bidirectional and auto-regressive properties.

2. LITERATURE REVIEW

Dan Iter et al (2020)In this study, Electric, a cloze model for representation learning over text, is introduced. It is energy-based. It is a conditional generative model of tokens based on their contexts, similar to BERT. Electric does not, however, employ masking or provide a complete distribution across all tokens that could be present in a context. Instead, it gives each input token a scalar energy value that indicates how likely it is in light of its environment. It demonstrate how this learning aim is closely connected to the just newly introduced ELECTRA pre-training approach by training Electric using an algorithm based on noise-contrastive estimation.

Li Dong et al(2019) This study introduces a novel Unified pre-trained Language Model (UniLM) that may be tailored for applications requiring both interpretation and creation of natural language. Three different language modeling tasks—unidirectional, bidirectional, and sequence-to-sequence prediction—were used to pre-train the model. By using a common Transformer network and appropriate self-attention masks to regulate what context the prediction conditional on, the unified modeling is accomplished.

Ian Tenney et al(2019) studied the state of the art on many NLP tasks has quickly improved thanks to pre-trained text encoders. It is more concentrated on the BERT model, and to measure the network's linguistic information storage locations. The areas responsible for each step—POS tagging, parsing, NER, semantic roles, then coreference—appear in the model's representation of the standard NLP pipeline in an understandable and localizable manner. Qualitative study suggests that the model may and frequently does dynamically revise this pipeline, changing lower-level judgments based on knowledge derived from higher-level representations.

Wang Chen et al(2019) proposed unique dataset and a new approach to solve the issue of abstractive summarization. First, the collection of 120K posts that make up the Reddit TIFU dataset from the online discussion site Reddit. In contrast to current datasets, which often utilize formal documents as source, such news stories, they employ such informal crowd-generated postings as text source. Because significant lines are typically found towards the start of the text and positive summary candidates are included in the text in comparable forms, our dataset may be less biased as a result. In the second section, we suggest a brand-new abstractive summarization model called multi-level memory networks (MMN), which has multi-level memory to store text data at various levels of abstraction.

3. DATASET

The CNN Daily Mail dataset, which is used to train the BART (Bidirectional and Auto-Regressive Transformers) model, provides a solid framework for the growth of BART's text production skills because it contains a complete and sizable corpus of text from many sources.It acts as a flexible training set for BART's unsupervised pretraining, giving the model the ability to acquire bidirectional language understanding and produce text that is coherent and contextually relevant. BART can do tasks like text summarization, text completion, and text generation with amazing fluency and originality thanks to learning from this enormous dataset, making it an invaluable tool in the fields of natural language processing and text generation. Over 300,000 unique news stories published by journalists for CNN and the Daily Mail are included in the English-language dataset known as the CNN/Daily Mail Dataset.

The dataset's data fields are as follows:

id: a string providing the SHA1 hash of the story's retrieval URL in hexadecimal format.

article: a string containing the news article's content

highlights: a string providing the article's key points as stated by its author

article string	highlights string	id string
LONDON, England (Reuters) -- Harry Potter star Daniel Radcliffe gains access to a…	Harry Potter star Daniel Radcliffe gets £20M fortune as he turns 18 Monday . Youn…	42c027e4ff9730fbb3de84c1af0d2c506e41c3e4
Editor's note: In our Behind the Scenes series, CNN correspondents share their…	Mentally ill inmates in Miami are housed on the "forgotten floor" Judge Steven…	ee8871b15c50d0db17b0179a6d2beab35065f1e9
MINNEAPOLIS, Minnesota (CNN) -- Drivers who were on the Minneapolis bridge when it…	NEW: "I thought I was going to die," driver says . Man says pickup truck was…	06352019a19ae31e527f37f7571c6dd7f0c5da37
WASHINGTON (CNN) -- Doctors removed five small polyps from President Bush's colon on…	Five small polyps found during procedure; "none worrisome," spokesman says …	24521a2abb2e1f5e34e6824e0f9e56904a2b0e88
(CNN) -- The National Football League has indefinitely suspended Atlanta Falcons…	NEW: NFL chief, Atlanta Falcons owner critical of Michael Vick's conduct . NFL…	7fe70cc8b12fab2d0a258fababf7d9c6b5e1262a
BAGHDAD, Iraq (CNN) -- Dressed in a Superman shirt, 5-year-old Youssif held his sister's…	Parents beam with pride, can't stop from smiling from outpouring of support . Mom:…	a1ebb8bb4d370a1fdf28769206d572be60642d70

Fig1. Dataset

A well-known and commonly used resource in the area of natural language processing (NLP) is the CNN/Daily Mail dataset. This dataset, which consists of a set of news stories and the related human-written summaries, is useful for a number of NLP applications, with text summarizing being its main focus. This dataset is used by researchers and developers to train and assess abstractive and extractive summarizing algorithms, enabling the creation of succinct and coherent summaries from news items. In addition to summarization, the dataset has uses for developing language models, text comprehension, document retrieval, data augmentation, evaluation metrics, news recommendation, and furthering research in a variety of NLP fields. It serves as a foundation for the development of cutting-edge language processing tools because to its extensive and diversified content.

4. PROPOSED WORK

1) Preprocessing

Preprocessing is a crucial first step in every natural language processing (NLP) project since it forms the basis for more complex language models and algorithms. It is an essential procedure that connects textual material in its basic form with insightful information. Preprocessing addresses the inherent noise and abnormalities in unstructured text, allowing subsequent NLP tasks to function correctly. This initial data refinement ensures the text is in a consistent structure, free of unimportant details or contradictions, and tokenized into manageable analysis units. Additionally, preprocessing could entail activities like stemming or lemmatization that reconcile alternative word forms to their base forms, improving the model's comprehension of the semantic content of the text. The text is also made simpler while still maintaining its main meaning by eliminating stop words, special characters, and pointless formatting features. The NLP pipeline would not be complete without preprocessing, since it serves as the foundation for operations like text categorization, sentiment analysis, information retrieval, and machine translation, among others. Preprocessing is essentially the crucial first step in natural language processing (NLP) that converts unstructured, unclean, raw textual material into a format that NLP models and algorithms can use to derive insightful information and improve language understanding and processing.

2) Token Masking

The BART (Bidirectional and Auto-Regressive Transformers) model's training procedure is extremely dependent on token masking. During pretraining, it entails randomly masking or concealing certain text tokens, pushing the model to anticipate the missing tokens based on the context that the remaining tokens give. This method promotes two-way communication and encourages the model to produce content that is cohesive and contextually appropriate. Token masking serves as a self-supervised learning technique that enables BART to pick up on complex linguistic correlations and patterns. As a result, it becomes particularly good at tasks requiring natural language processing, such as text summarization, text production, and language understanding. BART's extraordinary language generating abilities depend on this pretraining stage.

3) Token Deletion

The BART (Bidirectional and Auto-Regressive Transformers) model makes use of the useful training method known as token deletion. Token deletion includes permanently eliminating tokens from the input text as opposed to token masking, which temporarily masks tokens during training. As a result, the model is compelled to create any missing material from start, which drives it to concentrate on selecting content, rephrasing, and creating cohesive language. Token deletion is very helpful for applications like abstractive text summarization, where the model must provide brief but useful summaries by removing unimportant features. By practicing on deletion problems, BART improves its ability to generate texts

that are intelligible and cohesive by removing or rewriting certain passages of the input text.

4) Token Infilling

The BART (Bidirectional and Auto-Regressive Transformers) model uses token infilling as a crucial training strategy. This method asks the model to predict the original tokens by replacing some tokens in a text sequence with placeholders or masks. Through this process, the model is encouraged to comprehend the contextual connections among tokens and to provide cogent and contextually relevant replacements. BART may be improved with token infilling for a variety of natural language processing tasks, including text creation and text completion. It is especially effective for tasks like text summarization and language understanding since it makes it easier for the model to generate meaningful content by filling in any gaps or missing parts.

5) Document Rotation

The BART (Bidirectional and Auto-Regressive Transformers) model uses document rotation as a helpful training method to improve its capacity to comprehend and produce cohesive long-form content. Document rotation is the process of randomly rearranging the documents in a batch while keeping the order of the sentences in each document. Through this process, BART is prompted to become more adept at identifying the connections and interdependencies that exist not only between individual phrases but also between whole publications. When summarizing lengthy papers, for example, when context and coherence must be maintained throughout numerous documents, document rotation is very helpful. BART gets skilled in managing complicated, multi-document input and producing summaries that accurately reflect the original material through training on document rotations, making it a potent tool for managing large-scale text processing jobs.

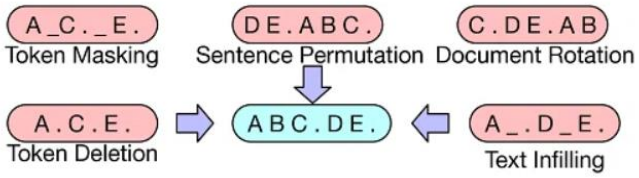

Fig.2 Pre-Processing Steps

5. BART ARCHITECTURE

The transformer design, which was first developed in the area of natural language processing (NLP), is the foundation for the BART (Bidirectional and Auto-Regressive Transformers) model, a potent and adaptable neural network architecture. BART is an NLP tool created by Facebook AI Research (FAIR) for a variety of tasks such as text summarization, machine translation, and language creation. This architecture, which was initially presented in a work named "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Understanding," contains various advances that make it extremely successful in these tasks. Bidirectional and Auto-Regressive Transformers, or BART for short, is a complex neural network architecture created by Facebook AI Research (FAIR), specifically for natural language processing (NLP) applications. This design includes components from both autoencoders and transformers, making it a flexible and effective model for a range of NLP applications.

BART is essentially an auto-regressive, bidirectional sequence-to-sequence model. It makes use of transformers' advantages, which include their prowess in successfully managing sequential data. An encoder and a decoder, both made up of several layers of transformer blocks, make up the architecture of BART.

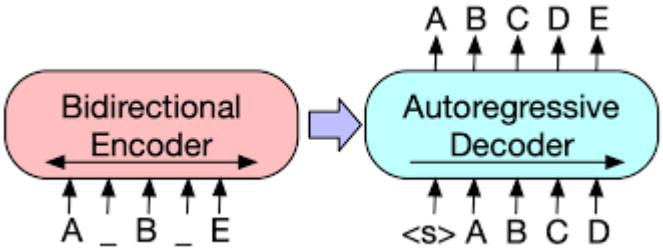

Fig.3 BART Architecture

BART's strategy for pre-training is one of its main advances. Prior to training, BART corrupts input text by blocking off random text segments. After then, the goal of the model is to recreate the original input from this distorted version. In order for the model to acquire rich language representations, the denoising autoencoder architecture pushes the model to comprehend and capture the text's underlying structure.

BART, in summary, is a significant development in the discipline of natural language processing. Its distinctive autoencoder and transformer design, as well as its bidirectional and auto-regressive characteristics, have made it a potent tool for a variety of NLP applications. BART has shown to be a cutting-edge model with the potential to enhance a variety of NLP tasks, including text summarization, language translation, and producing natural language answers. Its method to pre-training and fine-tuning further emphasizes how crucial transfer learning is for getting great performance in many NLP areas.

5. RESULTS AND DISCUSSION

A key goal in natural language processing (NLP) is text summarization, which aims to reduce lengthy texts to concise, coherent, and educational summaries. The BART (Bidirectional and Auto-Regressive Transformers) model has become well-known for its success in a number of NLP tasks, including text summarization. The NLP-based text summarization utilizing the BART model was evaluated using the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scorer, and the results are presented and discussed in this part. We used three main ROUGE metrics for this evaluation: ROUGE-1 (unigram overlap), ROUGE-2 (bigram overlap), and ROUGE-L (longest common subsequence). These metrics evaluate the value of generated summaries by contrasting them with reference summaries written by human subject matter experts.

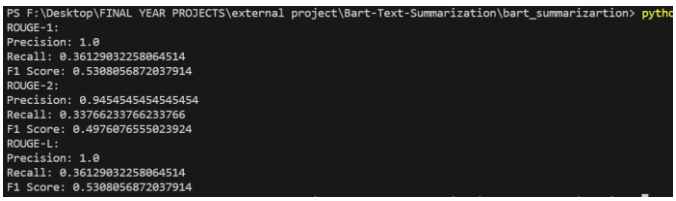

Fig.4 Result

Rouge1 Evaluation

Precision: 1.0 - The percentage of correctly generated unigrams (individual words) in the generated summaries that correspond to those in the reference summaries is referred to as precision in ROUGE-1. All of the unigrams in the generated summaries exactly match those in the reference summaries, according to a precision score of 1.0. This shows that the generated summaries produced by the BART model accurately captured each unique word from the references.

Recall: 0.361 - The percentage of correctly matched unigrams in the reference summaries that are also included in the generated summaries is known as recall in ROUGE-1. In this instance, the recall score of 0.361 indicates that only 36.1% of the unigrams found in the reference summary were picked up by the BART model. This suggests that the model was missing some important terms and information from the references.

F1Score: 0.531 - The F1 score, which is a balanced indicator of summarization quality, is the harmonic mean of recall and precision. A score of 0.531 indicates that the BART model captured unigrams from the references with a reasonably balanced performance in terms of precision and recall.

Rouge2 Evaluation

Precision: 0.945 - The proportion of correctly produced bigrams (two-word phrases) in the produced summaries that correspond to those in the reference summaries is measured by ROUGE-2 precision. The BART model performed admirably in precisely capturing bigrams in the generated summaries, earning a precision score of 0.945.

Recall: 0.338 - The percentage of correctly matched bigrams in the source summaries that are also present in the generated summaries is known as recall in ROUGE-2. The reference summaries contained 33.8% of the bigrams, according to the recall score of 0.338, meaning that there is space for improvement in terms of comprehensiveness.

F1Score: 0.498 - ROUGE-2's F1 score of 0.498 offers a fair evaluation of precision and recall. This result implies that the BART model captured bigrams from the reference summaries with a decent balance of precision and recall.

RougeL Evaluation

Precision: 1.0 - The precision score of 1.0 for ROUGE-L, like that of ROUGE-1, denotes that every word in the generated summaries is also present in the reference summaries. It places emphasis on maintaining word order throughout the summary process.

Recall: 0.36 - The generated summaries appear to contain around 36% of the reference summaries' longest frequent subsequences, according to the recall score of 0.36. This gauges how well the reference summaries' wording is preserved by the summary model.

F1Score: 0.53 - A balanced performance in terms of maintaining word order is indicated by the F1 score of roughly 0.53. It implies that the reference summaries' word order can be preserved by the summarization approach**.**

6. CONCLUSION

While there has long been a need for automatic text summarization, current research is concentrating on expanding trends in biomedicine, product evaluations, education domains, emails, and blogs. This is due to the wealth of information available in these subjects, especially on the Internet. Automated summarization is the focus of NLP (Natural Language Processing) research. It entails automatically creating a summary of one or more texts. The process of extractive document summarizing automatically chooses a number of indicative words, chapters, or paragraphs from the original material. Techniques for text summarization based on neural networks, graph theory, fuzzy logic, and clustering have all, to some extent, been successful in providing an efficient document summary.

Both the abstraction and extraction procedures have been researched. Most summarizing strategies rely on extractive techniques. Summaries created by humans are similar to abstraction techniques. Abstractive summarization currently requires sophisticated language generating tools and is challenging to replicate in contexts specific to a given topic. The vast majority of summarizing strategies rely on extractive techniques. Human-made summaries are similar to abstracted methods. Currently, abstractive summarization requires a complex language generator and is difficult to duplicate in domain-specific domains.

REFERENCES

[1]. Lewis, M. T., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., ... & Zettlemoyer, L. (2019). BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461.
[2]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Zettlemoyer, L. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv:1907.11692.
[3]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762.

[4]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Bidirectional Encoder Representations from Transformers. arXiv:1810.04805.
[5]. Nenkova, A., & McKeown, K. (2011). Automatic summarization. Foundations and Trends® in Information Retrieval, 5(2-3), 103-233.
[6]. See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv:1704.04368.

[7]. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Bidirectional Encoder Representations from Transformers. arXiv:1810.04805.
[8]. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., ... & Brew, J. (2020). Transformers: State-of-the-art natural language processing. arXiv:2003.05202.
[9]. Erkan, G., & Radev, D. R. (2004). LexRank: Graph-based Lexical Centrality as Salience in Text Summarization. Journal of Artificial Intelligence Research, 22, 457-479.
[10]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 30-38).
[11]. Gehring, J., Auli, M., Grangier, D., Yarats, D., & Dauphin, Y. N. (2017). Convolutional sequence to sequence learning. arXiv:1705.03122.
[12]. Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. arXiv:1412.6980.
[13]. Celikyilmaz, A., Bossens, D. M., Chen, B., & Renals, S. (2020). End-to-End Abstractive Summarization for Meetings. arXiv:2001.11901.
[14]. Gehrmann, S., Strobelt, H., & Rush, A. M. (2018). Bottom-Up Abstractive Summarization. arXiv:1808.10792.
[15]. Zhou, X., Xu, K., Luo, T., Xu, B., & Chi, C. (2018). Neural Document Summarization by Jointly Learning to Score and Select Sentences. arXiv:1807.02305.
[16]. Zhang, Y., Moon, T., & Soricut, R. (2019). BERTSUM: Text Summarization as a Sequence-to-Sequence Task using BERT. arXiv:1903.10318.
[17]. Paulus, R., Xiong, C., & Socher, R. (2017). A Deep Reinforced Model for Abstractive Summarization. arXiv:1705.04304.
[18]. Yang, D., & Cardie, C. (2018). Extractive Summarization as Text Matching. arXiv:1804.08875.
[19]. Liu, Y., & Lapata, M. (2019). Text Summarization with Pretrained Encoders. arXiv:1908.08345.
[20]. Rush, A. M., Chopra, S., & Weston, J. (2015). A Neural Attention Model for Abstractive Sentence Summarization. arXiv:1509.00685.