

SUMMARISING PATIENT-PHYSICIAN CALLS USING FINE-TUNED BART MODEL

Charankumar P C

Student
Dept. of AI&DS
R.M.K Engineering College
Chennai, India
char20110.ad@rmkec.ac.in

Sunil B

Student
Dept. of AI&DS, ,
R.M.K Engineering College,
Chennai, India
suni20218.ad@rmkec.ac.in

Nishant Y P

Student
Dept of AI&DS
R.M.K Engineering College,
Chennai, India
nish20227.ad@rmkec.ac.in

Sumanth Vamsi

Student,
Dept. of AI&DS,
R.M.K Engineering College,
Chennai, India
suma20217.ad@rmkec.ac.in

Dr.M.Hemalatha

Associate Professor,
Dept. of AI&DS
RMK Engineering College,
Chennai, India
mhl.ad@rmkec.ac.in

Abstract—In the healthcare field, the summarization of medical conversations between physicians and patients is deemed a challenging, time-consuming, and error-prone task. To tackle this issue, the fine-tuned Bidirectional AutoRegressive Transformer (BART) model, a robust Natural Language Processing (NLP) tool trained on GPT-3 synthetic data, renowned for its excellent summarization capabilities, is employed. The solution involves the generation of concise summaries of these conversations, encompassing vital details and treatment plans. Artificial intelligence (AI) analysis is utilized to assess satisfaction scores, identify key conversation keywords and sentences, and perform topic modeling for the creation of personalized lifestyle plans. The approach, employing the BART-large-CNN model, consistently outperforms other state-of-the-art models like Pegasus and t5-large, as indicated by performance metrics, including ROUGE-1, ROUGE-2, and ROUGE-L. Furthermore, user engagement is enhanced by incorporating FrameVR for the creation of immersive virtual reality experiences, resulting in comprehensive environments. To ensure the accurate transcription of audio and video files, reliance is placed on Whisper AI, an advanced Speech-to-Text model with multilingual support. Additionally, cloud synchronization is implemented to maintain data integrity and prevent unauthorized alterations to stored transcripts and summaries.

Keywords—BART, ROUGE, FRAMEVR, Whisper AI, Language model.

I. INTRODUCTION

As technology advances, a lot of advancements are being made in the field of machine learning. One significant area of development is Natural Language Processing (NLP), which is receiving a lot of interest in the healthcare sector due to its amazing ability to analyse and comprehend vast amounts of patient information. By applying complex algorithms and machine learning techniques, NLP offers the capability to extract crucial concepts and insights from data that were previously concealed within textual data.

The ability of NLP to transform unstructured data into meaningful information is causing its significance in the healthcare sector to grow. This ground-breaking technology has the potential to greatly raise the bar of care, simplify medical procedures, and ultimately enhance patient outcomes. Medical records analysis presents a difficult problem because of their frequent plain language, usage of specialised terminology, and absence of a clearly defined structure. Various kinds of knowledge are needed for this assignment. The findings could, among other things, drive decisions about hospital administration and classify medical therapies. The method of document summarization has a lot of potential in fields like information

retrieval and online text searches because of the continual increase of digital textual resources.

II. LITERATURE SURVEY

The pretrain-then-finetune technique has greatly improved performance and supplanted the conventional paradigm in the field of NLP since the advent of ELMo [1] and BERT [2]. Domain-specific datasets can be used to train language models, which can improve their performance on tasks unique to particular domains [3]. Researchers have created pre-trained language models, including BioBERT [4] and PubMedBERT [5], specialised for the biomedical domain using the vast and freely available corpora from PubMed.

Natural Language Generation (NLG) problems, including topics like dialogue systems [6] and question answering [7], are of utmost significance in the field of biological artificial intelligence research. Natural language understanding is increasingly being tackled as NLG tasks, even in more general domains [8,9]. Constrained natural language generation approaches, for instance, can be used to successfully handle tasks like entity retrieval [10].

Despite the significant benefits that NLP offers, it is still difficult to analyse medical records because to their complexity. This complexity results from the frequent use of simple language, the incorporation of technical jargon, and the lack of a clear framework. For the latent potential of this data to be properly unlocked, such complexity necessitates insights from other domains.

In order to overcome these obstacles, the suggested strategy incorporates cutting-edge technologies to promote more effective and informative doctor-patient interactions. To quickly record talks, the Whisper Automatic Speech Recognition (ASR) model is used. The Facebook/BART-large-CNN model's hyperparameter adjustment also improves call interaction summarization. By utilising these features, the FRAME VR platform facilitates doctor-patient interactions in the metaverse through augmented reality (AR).

In light of the challenges posed by complex medical records and the critical need for improved doctor-patient interactions, this approach is tailored to address these issues. The integration of Whisper ASR and summarization models, coupled with the utilization of VR and cloud technologies, presents a comprehensive solution to enhance healthcare communication, accessibility, and data management.

III. METHODOLOGY

A. Summary of the Proposed Method

The proposed method involves the users to interact with the physicians through the metaverse platform where the session will be recorded and will be uploaded into the application and also the users are provided with an interface where they can also upload their pre-recorded calls which they had with their physicians. Once the audio recordings are uploaded the transcripts of the audio are generated using Whisper AI, an ASR model and once the transcripts are generated it will then be passed to the BART-large model for the summarization task.

Once the summaries are generated it will be displayed to the user. Here we provide an additional cloud feature that provides the continuity of care for the patients. Because once the user submits their call recordings, the call recording, the corresponding transcripts and the summaries will be uploaded to the cloud under the specific patient ID. This can help the Physician to look into the previous summaries and understand the patient's history.

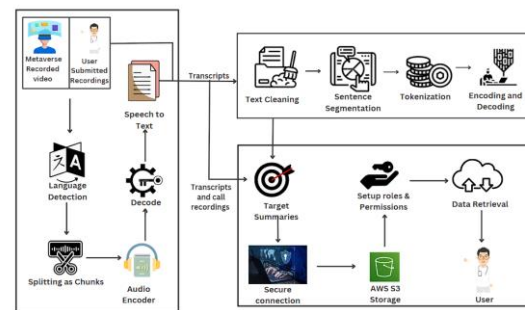


Fig.1 Patient-Physician Calls architecture diagram

B. Whisper AI

The proposed application allows the upload of either pre-recorded calls or recordings from the metaverse platform and seamlessly integrates with it using the Django framework. After a successful upload, the system uses the state-of-the-art Whisper Automatic Speech Recognition (ASR) model from OpenAI to transcribe the audio or video files. The encoder-decoder approach used in the Whisper ASR architecture as shown in Fig.2 is the encoder-decoder Transformer. After being broken up into 30-

second chunks, the audio is input into an encoder, where it is converted into a log-Mel spectrogram. A decoder is trained to anticipate the corresponding text caption using special tokens that instruct the single model to perform tasks like language identification, phrase-level timestamping, multilingual speech transcription, and to-English speech translation.

C. Summarization Module

The transcribed transcripts obtained from Whisper ASR are then channelled into the powerful BART summarization model, provided by Facebook. This sophisticated model employs advanced natural language processing techniques to distill the extensive transcripts into concise and coherent summaries. The target summaries extracted encapsulate the key points and core information from the original content, enhancing its accessibility and digestibility.

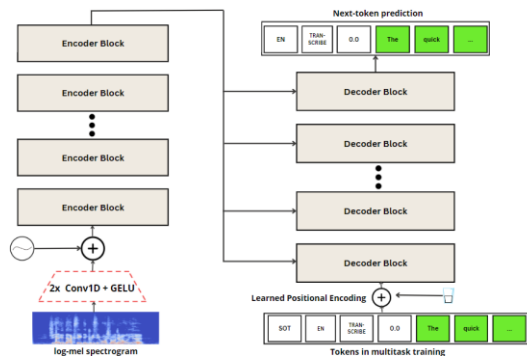


Fig.2 Whisper AI Model Architecture Diagram

D. Cloud Working

The application's design is centred on the needs of the user. Uploading audio and video files is simple for authorised users, and the files are then safely stored in the cloud as shown in Fig.1. This cloud integration ensures robust data management, scalability, and dependability. The user interface of the application is created to provide a seamless experience, enabling users to easily browse through the uploaded content, transcripts, and generated summaries.

Within the architecture of the Django application, the orchestration of data transfer between the application and an Amazon Web Services (AWS) Simple Storage Service (S3) bucket is achieved through the utilization of the Boto3 Python library—a powerful interface to AWS services. When a user

triggers an action necessitating data transmission, pertinent information such as patient identities and corresponding contact details is extracted from the Django application's model layer, reflecting a coherent embodiment of the application's data schema. Serialization of this extracted data into a suitable format, such as JSON, is then facilitated, aligning it for seamless integration with the AWS infrastructure. As part of the data lifecycle management, the transcripts and corresponding summaries are persistently stored within our cloud infrastructure.

Users with the appropriate permissions can conveniently access their archived summaries, enabling them to efficiently retrieve and review their historical content. This streamlined process fosters a sense of continuity and enables users to leverage previous insights without undue effort.

E. Fine-tuned Bart-large-cnn model

The proposed model, as discussed in this section, serves as a pivotal component of our application, as delineated in the methodology. Its expertise in abstractive text summarization is highly pertinent, especially in healthcare, where the succinct distillation of patient interactions is paramount. This model enriches our application by enhancing the accessibility of patient call transcripts, enabling users to glean actionable insights from historical data. This underscores the impactful integration of cutting-edge technology in healthcare communication and knowledge management.

A large dataset of 1200 records were used in this paper, and each record contained the three crucial columns "id," "dialogue," and "summary." Along with the seamless integration of cutting-edge technology, meticulous data preprocessing was carried out. The 'facebook/bart-large-cnn' model from the HuggingFace library was integrated and served as the main building block of our project. This model is renowned for its skill in abstractive text summarization.

This model, which consists of both encoders and decoders, uses a bidirectional approach to ensure a firm understanding of language semantics and contextual cues. Decoders create clear, concise summaries that capture the essence of patient call interactions as encoders process the input text to identify intricate patterns as shown in Fig 3. Utilising the 'rouge' metrics, a well-known metric for quantifying the quality of machine-generated

summaries, was a crucial part of our model integration. The effectiveness of our summarization outputs is guaranteed by this stringent evaluation framework, enabling accurate evaluations of the performance of our model.

The creation of an "AutoTokenizer" class, a crucial element from the Transformers library, marked the beginning of the process of tuning this potent model. This class made it easier to create a tokenizer instance by loading a tokenizer that had already been trained to work with the "facebook/bart-large-cnn" model. The model's linguistic intelligence was harnessed through this instantiation, allowing for efficient tokenization of input dialogues and target summaries.

Important parameters were precisely defined to prepare the ground for tuning. Notably, the 'max_input_length' was purposefully set to 512 characters, while the 'max_target_length' was limited to 128 characters. These settings optimised the delicate balancing act between output succinctness and comprehension, an essential aspect of abstractive summarization.

The creation of a preprocessing function was crucial to the process of fine-tuning. During this stage, the dialogues were tokenized to create labels, or target summaries, suitable for training. Truncation was configured as "True" for both dialogues and summaries as a strategic decision to ensure compatibility with the model's input specifications. The 'max_length' parameter was also properly configured to guarantee that dialogue and summary lengths stayed within the predefined ranges. The end result was a meticulously organised dataset that had been carefully calibrated and formatted to act as the foundation for our Seq2Seq model's (BART Model's) subsequent fine-tuning.

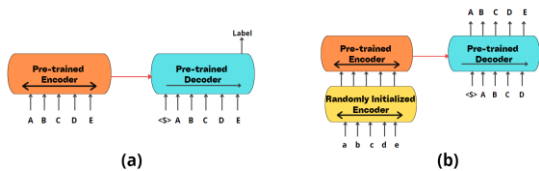


Fig.3 Fine tuning BART for (a) Classification; (b) Translation

IV. EXPERIMENTS AND RESULT

A. Dataset

A two-fold approach was employed in data acquisition strategy to augment the dataset for training the BART model. Initially, a comprehensive dataset comprising 300 audio files capturing patient-doctor conversations and their

corresponding transcripts was acquired from Figshare[11], a recognized repository for scholarly content.

Recognizing the need for a more extensive training corpus, the Chat-GPT model was utilized to synthetically generate an additional 900 conversations along with their concise summaries. This augmented dataset harmonizes authentic dialogues with simulated conversations, fostering a diversified and enriched training substrate for the refinement of BART-large summarization model.

B. Model Selection

In this paper, the BART-large, Pegasus, and T5-large models were each fine-tuned separately. Every model contributed its distinct architecture and abilities to the project, enhancing the variety of experimentation and validation.

Using the "facebook/bart-large-cnn" checkpoint, the experiment was started with the BART-large model. Due to its superior performance in sequence-to-sequence tasks, BART, which is based on the Transformer architecture, is a great choice for dialogue summarization. Due to the model's use of bidirectional attention, the entire dialogue context could be considered during summary generation. The "AutoModelForSeq2SeqLM" class was utilised, and its parameters were tuned to the dialogue summarization task during the fine-tuning process.

Pegasus was also utilised to determine how model size affected performance. This purpose was served by the "Pegasus" checkpoint. In an effort to balance performance and efficiency, Pegasus keeps the essential elements of BART while showcasing a more condensed architecture. Pegasus underwent fine-tuning using the same methods, revealing how model size affected the results of dialogue summarization.

The realm of T5-large was explored as a further extension of the exploration. The text-to-text transfer Transformer architecture serves as the foundation for T5, which, in contrast to BART, frames a variety of NLP tasks as translation tasks. It was modified for dialogue summarization by taking advantage of T5's generality. This model's potential for the task at hand was revealed by the "t5-large" checkpoint, which put it through the fine-tuning pipeline.

C. Model-Specific Configuration for Seq2Seq Fine-Tuning

After the data had been prepared for processing and the foundation for model integration had been established, attention was turned to the configuration necessary for honing our Seq2Seq models. Notably, the models—namely, facebook/bart-large-cnn, Pegasus, and t5-large—were expertly integrated into the Seq2Seq framework, where they produced abstractive summaries for patient call interactions. The establishment of the training parameters that governed the fine-tuning procedure was a crucial step.

The following crucial hyperparameters were established for each model: "batch_size," which was carefully selected to enhance training. Different models used different values; for example, facebook/bart-large-cnn and Pegasus used batch sizes of 4, while t5-large used a batch size of 1 due to GPU restrictions. It was wise to set the "evaluation_strategy" to "epoch," which punctuated the conclusion of each training epoch with a model evaluation and gave timely insights into the model's developing competence. The scale of parameter updates was determined by the "learning_rate," an important agent, balancing rapid convergence with stability.

To carefully balance computational resources and effective training, the parameters "per_device_train_batch_size" and "per_device_eval_batch_size" controlled the granularity of data chunks processed per device. The invention of "gradient_accumulation_steps" served as a clever workaround for memory restrictions, allowing for larger batch sizes while taming memory consumption. With "Weight_decay" set to 0.01, regularisation vigour was enhanced, overfitting tendencies were reduced, and generalisation skills were strengthened.

In order to manage storage pragmatism without jeopardising training checkpoints, the "save_total_limit," modestly set at 2, imposed restraint on checkpoint accumulation. The cyclic exposure of the entire dataset to the model's scrutiny was captured by the "num_train_epochs" parameter, promoting convergence without veering into overfitting. Last but not least, the "predict_with_generate" parameter opened the world of forecasts during evaluation, allowing the model to produce summaries and ushering in a thorough performance evaluation.

The orchestration of these parameters delineated a calibrated training environment, where each parameter choice harmonized with others to propel the model toward an optimal dialogue summarization process.

D. Data Collation and Training

The "DataCollatorForSeq2Seq" class is instrumental in orchestrating the preparation of input-output pairs for training. This pivotal class ensures that tokenized dialogues and summaries are appropriately aligned, padded, and masked for efficient model training. By providing a seamless interface between tokenized data and the model, the data collator streamlines the training process and enhances the model's understanding of the relationship between dialogues and corresponding summaries.

The "Seq2SeqTrainer" class integrates all these aspects into a coherent training process. It orchestrates the iterative fine-tuning of the model using batches of tokenized data, enabling the model to learn how to distill patient call interactions into accurate summaries. The trainer's holistic approach encapsulates the synthesis of model architecture, data collation, and training configurations.

E. Evaluation and Performance Metrics

The effectiveness of the Seq2Seq model in dialogue summarization requires rigorous evaluation using ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics, which are widely recognized benchmarks in natural language processing. These metrics assess the alignment between machine-generated and reference summaries and include ROUGE-1, ROUGE-2, and ROUGE-L scores.

The ROUGE-1 metric measures unigram overlap, indicating that all models capture important dialogue content. The ROUGE-2 metric emphasizes the models' ability to preserve meaningful phrases in summaries. The ROUGE-L metric emphasizes long sequences, highlighting coherence and fluency. The ROUGE-Lsum metric is also calculated, which provides a comprehensive view of the models' overall performance.

The evaluation process uses a fine-tuned model to predict summaries for the validation dataset. The compute_metrics function quantifies model performance through token decoding, sentence tokenization, and ROUGE. It computes essential metrics like precision, recall, and F1-score, and considers linguistic nuances using stemmers for improved comparability.

F. Result and Analysis

The evaluation results of the fine-tuned Seq2Seq models for dialogue summarization are presented in Table I and Table II. The tables showcase the performance and ROGUE scores of three distinct models: facebook/bart-large-cnn, Pegasus, and t5-large.

Model	Rouge-1	Rogue-2	Rogue-L
facebook/bart-large-cnn	41.99	21.23	32.56
Pegasus	47.93	25.05	40.86
t5-large	18.19	12.0	15.95

Table I. Training Results of Samsun Dataset

In Table I, the Samsun dataset [12] serves as the training data for three prominent models: Facebook/bart-large-cnn, Pegasus, and t5-large. The table presents a clear demonstration of Pegasus superior performance, exhibiting a noteworthy Rogue-L score difference of 8.3 when compared to Facebook/bart-large-cnn and a remarkable 24.91 in comparison to t5-large.

Conversely, in Table II, we employ GPT-3-synthetic data to train three distinct models: Facebook/bart-large-cnn, Pegasus, and t5-large. The table unmistakably illustrates that Facebook/bart-large-cnn outperforms the other models, showcasing a significant Rogue-L score difference of 12.83 with respect to Pegasus and a substantial 22.47 when compared to t5-large.

Model	Rouge-1	Rogue-2	Rogue-L
facebook/bart-large-cnn	55.32	37.28	46.25
Pegasus	42.12	20.54	33.42
t5-large	20.74	10.34	19.78

Table II. Training Results of GPT-3 Synthetic Dataset

While the inclination may naturally lean toward selecting the Pegasus model due to its notably high Rouge metrics, it is imperative to consider the specific context of the data. In the Samsun dataset, dialogues are typically limited to a maximum of two exchanges. This circumstance differs markedly from scenarios such as doctor-patient calls, which often involve more extensive conversations. Consequently, we have elected to prioritize the Facebook/bart-large-cnn model, as it emerged as the top performer on the GPT-3-synthetic dataset. This decision is guided by the fact that the GPT-synthetic

dataset comprises an average of 20 pairs of conversations, rendering it a more suitable choice for real-time applications in contexts where more extended dialogues are prevalent.

To visually depict the performance disparities, the Fig.4 below presents a comparison of the average ROUGE scores across different models for both datasets:

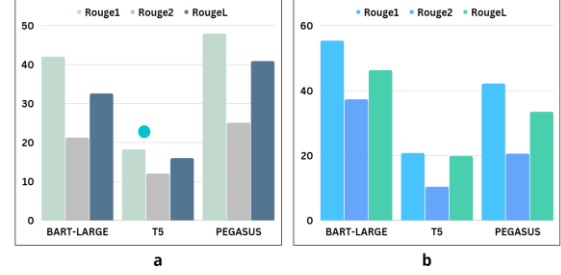


Fig.4 Average Rouge Scores of Models; a) Samsun Dataset; b) GPT-3 Synthetic Dataset

The facebook/bart-large-cnn model exhibits superior performance across all relevant metrics, making it a compelling choice for our dialogue summarization project. Its ability to generate summaries that mirror the content, structure, and coherence of the reference summaries aligns well with our objective of creating accurate and informative dialogue summaries.

V. CONCLUSION

In conclusion, the proposed BART-large model demonstrated exceptional performance in dialogue summarization, as evidenced by high ROUGE scores and comparisons with benchmark datasets such as Samsun. Its bidirectional attention, fine-tuned parameters, and robust training process enabled it to distill patient call interactions into accurate and concise summaries. As a result, the BART-large model was selected for integration into the application, ensuring the delivery of high-quality summaries and enhancing the accessibility and digestibility of recorded content from both pre-recorded calls and metaverse platform recordings. Furthermore, considerations for future enhancements include the incorporation of advanced features such as real-time collaboration, sentiment analysis, and user-specific content recommendations, aimed at further improving the model's accuracy and enriching the user experience and insights derived from the application.

REFERENCES

- [1] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and

Luke Zettlemoyer. 2018. Deep contextualized word representations. In NAACL.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for 106 Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[3] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342–8360, Online. Association for Computational Linguistics.

[4] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36:1234 – 1240.

[5] Yuxian Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2022. Domainspecific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23.

[6] Hsiao-Tuan Chao, Lucy Liu, and Hugo J Bellen. 2017. Building dialogues between clinical and biomedical research through cross-species collaborations. In *Seminars in cell & developmental biology*, volume 70, pages 49–57. Elsevier.

[7] Qiao Jin, Zheng Yuan, Guangzhi Xiong, Qianlan Yu, Huaiyuan Ying, Chuanqi Tan, Mosha Chen, Songfang Huang, Xiaozhong Liu, and Sheng Yu. 2022. Biomedical question answering: A survey of approaches and challenges. *ACM Comput. Surv.*, 55(2).

[8] Tianxiang Sun, Xiangyang Liu, Xipeng Qiu, and Xuan-jing Huang. 2021. Paradigm shift in natural language processing. arXiv preprint arXiv:2109.12575.

[9] Hang Yan, Tao Gui, Junqi Dai, Qipeng Guo, Zheng Zhang, and Xipeng Qiu. 2021. A unified generative framework for various NER subtasks. In

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 5808–5822, Online. Association for Computational Linguistics.

[10] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive entity retrieval. In *International Conference on Learning Representations*.

[11] Fareez, F., Parikh, T., Wavell, C. et al. A dataset of simulated patient-physician medical interviews with a focus on respiratory cases. *Sci Data* 9, 313 (2022). <https://doi.org/10.1038/s41597-022-01423-1>

[12] P. Yi and R. Liu, "A Relation Enhanced Model For Abstractive Dialogue Summarization," 2022 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC), Suzhou, China, 2022, pp. 240-246, doi: 10.1109/CyberC55534.2022.00047.