Medical Dialogue Summarization System using BART-base

Minh Pham[1]
Alex Golding
Ka Cheung Chan
Sean Tong

[1]bp00551@surrey.ac.uk

aq02174@surrey.ac.uk

kc01381@surrey.ac.uk

st01330@surrey,ac,uk

***Abstract*– This paper presents a natural language processing system designed to automatically generate structured clinical summaries from doctor-patient conversations. This system leverages a BART-base pre-trained model fine-tuned on the MTS-Dialog dataset to transform multi-turn medical dialogues into coherent clinical narratives. The model employs a sequence-to-sequence transformer architecture with a bidirectional encoder and autoregressive decoder, specifically optimized for medical terminology and conversational patterns. Training was conducted using PyTorch and Hugging Face Transformers on an NVIDIA A100 GPU, with evaluation metrics including BERTScore F1 and ROUGE-L. This approach offers a promising solution to reduce clinician documentation workload while maintaining the accuracy and completeness necessary in medical documentation.**

I. Introduction

Medical dialogues between healthcare providers and patients contain rich information that needs to be efficiently documented in clinical notes. However, it is also a repetitive task that takes away time from the clinician. An article made by Jo Makosinski remarked that on average, a clinician takes 13 hours a week to document notes, which is ⅓ of their work week[1]. An automated conversion of this process can reduce a clinician's workload, making direct patient care easier. By integrating already existing transcription technology, a generated summary will take less than a second to be outputted.

Current approaches to dialogue text summarization typically employ extractive or abstractive methods. Extractive methods identify and extract key sentences from source text, while abstractive methods generate new text that captures essential information. Recent advances in transformer-based language models (e.g., BART, T5, and their biomedical variants) have significantly improved abstractive summarization capabilities.

There are also several challenges that make medical dialogue summarization complicated. Currently, there's an abundance of publicly available transformers base language models such as T5, BART, GPT and their variants that do a great job at summarizing texts and or dialogues. But, in a field that needs high precision, along with a specific domain of terminology. Furthermore, timing is also crucial in medical care, so the model would need to be able to document this with great accuracy with regards to this.

In this work, we present a natural language processing system designed to automatically generate concise and structured summaries from doctor-patient conversations. This system leverages the BART base model fine-tuned on the MTS-Dialog dataset to transform multi-turn medical dialogues into coherent clinical narratives.

II. Relevant works

BioBart, a generative language model that was pretrained on PubMed abstracts was developed and released by Yuan et al. (2022)[2]. Their work demonstrated significant performance improvements across various biomedical NLG tasks including dialogue systems, summarization, entity linking, and named entity recognition. Importantly, they found that pre-training without the sentence permutation task led to better

performance on biomedical NLG tasks, suggesting domain-specific considerations for model architecture.

This research directly relates to our medical dialogue summarization project as it shows that domain-specific pre-training of transformer models can significantly improve performance on medical text tasks, including summarization of clinical content.

Lu et al. (2022) introduced ClinicalT5, a T5-based generative language model specifically pre-trained on clinical text[3]. Building upon the SciFive model (which was pre-trained on biomedical literature), they further trained ClinicalT5 on approximately 2 million clinical notes from the MIMIC-III database using a span-mask denoising objective. Their extensive evaluation showed that ClinicalT5 significantly outperformed general-domain T5 models and compared favorably with other domain-specific models across various clinical NLP tasks, including document classification, named entity recognition, and natural language inference. They also demonstrated ClinicalT5's effectiveness on real-world clinical applications such as predicting 30-day readmission risk and patient mortality. Their work highlights the importance of domain-specific language models for clinical text, which has unique linguistic characteristics compared to general text or even non-clinical biomedical content.

In our model, we opted for fine-tuning BART-base directly on medical dialogues. This choice eliminates potential biases introduced by BioBART's pretraining on PubMed abstracts, which primarily contain formal scientific writing rather than conversational medical language. This work will also specifically focus on doctor-patient dialogue summarization, allowing for more targeted optimizations for conversational medical language.

III. Methods

This study focuses on developing an extractive summarization model for medical dialogue texts, employing a fine-tuned sequence-to-sequence (Seq2Seq) transformer architecture. We selected the BART-base pre-trained model (Lewis et al., 2020) as our foundation due to its proven effectiveness in text generation and summarization tasks[4]. BART combines a bidirectional encoder with an autoregressive decoder, making it particularly suitable for our task of transforming medical dialogues into concise, informative summaries.

The model architecture consists of:

1. **Encoder**: A transformer-based encoder that processes the input dialogue text. The encoder captures the contextual information within the medical conversation, including important medical details, patient history, and diagnostic information.
2. **Decoder**: A transformer-based decoder that generates the summary text. The decoder leverages the encoded representations to produce summaries that retain essential medical information while removing redundancies.

The input to our model is a tokenized representation of doctor-patient dialogues, with a maximum sequence length of 512 tokens to accommodate lengthy medical conversations. The output is a generated summary with a maximum length of 128 tokens, designed to capture the key medical information present in the dialogue.

We employed a fine-tuning strategy that adapts the pre-trained BART model to the specific domain of medical dialogues. The process involves:

1. **Task-specific Training**: Optimizing the model's ability to identify the type of summarization best needed from the dialogue, ie. patient history, family history, etc.
2. **Synonym training using back translation**: Optimizing the model’s ability to recognize synonyms by augmenting the dataset.
3. **Pre-fine tuning model before training**: By training BART under a large text summarization dataset, the model will be better at extracting key information.

More details on these strategies in the implementation section.

]IV. Implementation

A) Experimental Setup

The experiments were conducted using the PyTorch framework and the Hugging Face Transformers library (version 4.18.0) in a Google Colab environment with a NVIDIA A100 GPU. The implementation leveraged NVIDIA CUDA for acceleration, with 40GB of available GPU memory.

B) Dataset and Preprocessing

We utilized the MTS Dialog medical conversation dataset comprising doctor-patient dialogues paired with expert-written summaries. The dataset is pre split into 4 sets, one training set, one validation set and two test sets. The training set contained 1201 dialogue-summary pairs, while the other sets included 100 pairs.

We also utilized the augmented dataset that MTS Dialogue provided. By translating the original data into another language and then translating it back to English, the model will be introduced to new synonyms and be more adapted to real world applications.

Preprocessing steps included:

1. Tokenization using BART's native tokenizer
2. Truncation of input dialogues to 512 tokens
3. Truncation of target summaries to 128 tokens

To guide the model to do certain tasks, we also utilized the section header of the MTS-Dialogue dataset. The header contains relevant first level information that can help guide the model to extract the correct information from dialogues. It is processed along with the input to guide the output to be of a certain format.

C) Training Configuration

The model and tokenizer is first loaded from the facebook/bart-base repository. It is then pre finetuned with datasets specialized in summarization tasks. The datasets chosen were the XSUM dataset. It was trained with the following hyperparameters:

- Learning rate: 3e-5
- Batch size: 4
- Training epochs: 3
- Optimizer: AdamW with weight decay of 16
- Evaluation strategy: Per epoch

The model was then trained with the following hyperparameters:

- Learning rate: 5e-5
- Batch size: 4 (effective batch size of 8 with gradient accumulation)
- Training epochs: 6
- Optimizer: AdamW with weight decay of 0.01
- Gradient accumulation steps: 2
- Evaluation strategy: Per epoch

D) Generation Parameters

During inference, we employed beam search with the following configuration:

- Beam size: 12
- Length penalty: 0.6
- Maximum generation length: 128 tokens

V. Results

A) Evaluation Metrics

We evaluated our fine-tuned BART model using several metrics to provide an accurate assessment of the summarization quality:

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measures the overlap of n-grams between the generated summary and reference summary. We report:

- **ROUGE-1**: Measures unigram overlap.
- **ROUGE-2**: Measures bigram overlap.
- **ROUGE-L**: Based on the Longest Common Subsequence (LCS), capturing sentence-level structure similarities.

BERTScore addresses limitations of n-gram based metrics by leveraging contextual embeddings from pre-trained language models. It computes:

- **Precision**: Similarity of each token in the candidate summary with the most similar token in the reference.
- **Recall**: Similarity of each token in the reference with the most similar token in the candidate.
- **F1**: Mean of precision and recall.

B) Summary of Results

Table 1 presents the performance of our fine-tuned BART model on the MTS-Dialog Test set 1.

Metric	Accuracy(%)
Bertscore Precision	36.74
Bertscore Recall	30.43
Bertscore F1	33.36
Rouge-1	37.07
Rouge-2	17.05
Rouge-L	31.11

Table 1: Performance metrics for BART model on MTS-Dialog validation set

C) Analysis and Discussion

Our fine-tuned BART model demonstrates moderate effectiveness at summarizing multi-turn dialogues in the specialized domain of the MTS-Dialog dataset. The BERTScore Precision of 36.74% indicates that tokens in the generated summaries align reasonably well with tokens in the reference summaries. The BERTScore Recall of 30.43%, while lower than precision, shows decent performance in capturing information present in the reference summaries. This results in a BERTScore F1 of 33.36%, suggesting a relatively balanced performance between precision and recall.

The ROUGE metrics provide additional insights about n-gram overlap. The ROUGE-1 score of 37.07% indicates good unigram overlap between generated summaries and ground truth, suggesting the model captures key terminology from the dialogues. However, the considerably lower ROUGE-2 score of 17.05% suggests challenges in maintaining exact bigram structures and phrase-level coherence - which is expected, given only one ground truth summary in the dataset. The resulting ROUGE-L of 31.11% represents the model's ability to capture longest common subsequences between generated and reference summaries.

Strengths:

1. **Semantic alignment**: The solid BERTScore F1 of 33.36% demonstrates the model's ability to generate summaries that are semantically similar to the references, even when exact wording differs.
2. **Precision-focused**: The higher BERTScore Precision compared to Recall indicates that the content included in generated summaries is generally accurate and relevant.
3. **Key terminology capture**: The strong ROUGE-1 performance suggests effective identification of important individual terms and concepts from dialogues.

Shortcomings:

1. **Phrasal coherence challenges**: The relatively low ROUGE-2 score indicates difficulty maintaining exact bigram structures from reference summaries.
2. **Information coverage gaps**: The BERTScore Recall being lower than Precision suggests the model sometimes fails to include all relevant information from the reference summaries.
3. **Structural preservation**: The moderate ROUGE-L suggests room for improvement in maintaining longer sequential patterns important for the narrative flow of dialogue summaries..

D) Suggested improvement methods

Our current approach could be enhanced in several ways. While we utilize both ROUGE-L and BERTScore for evaluation, there is room for improvement in extracting the unique aspects of medical dialogue summarization. One key limitation is that the dataset provides only a single reference summary per dialogue, which may not fully reflect the range of valid outputs. For instance, if a patient describes symptoms of "head pain," the model might reasonably summarize this as a "neurologic issue" or "migraine," yet receive a low ROUGE/BERTScore due to lack of n-gram or token overlap with the ground truth "headache." This highlights a limitation in relying solely on automatic metrics in such specialized domains.

One direction would be implementing a hierarchical encoding mechanism specifically designed for medical dialogues. Such architecture could better model the turn-taking dynamics between doctors and patients, with separate

encoders for each utterance and dialogue flow. This approach could improve the model's ability to track information across long conversations and distinguish between critical diagnostic exchanges and general conversational elements.

Another potential enhancement would be incorporating medical knowledge integration through specialized adapters or domain-specific pre training objectives. By leveraging external medical knowledge bases like UMLS (Unified Medical Language System), the model could better recognize and properly contextualize specialized terminology that appears infrequently in the training data.

Notably, we also tried to create specialized tokens of medical terms, representing them as a single token to preserve their meaning. Upon further testing, this method seems to yield diminishing returns likely because of training data size. However, this method should give an improvement if the training data is substantial.

From a patient safety perspective, another module has to be implemented at the end of the generation pipeline. This will be an extraction to double check any critical information such as medication dosages, allergies or negated symptoms. A confidence score should also be included in every generated output and it should flag any output below a certain threshold.

E)Sample Outputs

Table 3 (Included at the end of this paper) presents representative examples of dialogue inputs and their corresponding generated summaries, illustrating both successful cases and typical error patterns.

F) Analysis on generated summaries and its metrics.

The model showed promising results. Upon further inspection of the generated text, we can see that it struggles at identifying terminologies or flows that it has not seen before. This is a challenge when fine-tuning for a field that needs precision with an extensive vocabulary.

Furthermore, the training data includes non-contributory sections, which can be hard to evaluate because the models will just output a general summarization for this. Rouge L score is also a limited representation of the results as it penalizes the use of synonyms and also has very limited semantic understanding.

Another weakness that both Rouge and Bertscore shares is that it relies on the ground truth of the summary, and since MTS-Dialog and many other datasets only have one ground truth per each dialogue, it can unfairly penalize summaries that include extra data, especially non-contributory summaries.

G) Manual evaluation of the generated summary

As mentioned in the previous section, the computational metrics are not reliable in assessing the summarization of this task. Some examples of this will be given below.

Dialogue	Ground truth	Summary
Doctor: Hello, how are you today? Patient: I am doing well. Doctor: Great. What would you like to bring up today? Patient: I have some questions about my liver. Doctor: Alright. Let's start with the basics. Do you drink? Excessive drinking can cause issues with the liver. Patient: No, I do not. I take a lot of Tylenol for pain and I am worried it is effecting my liver. Doctor: Okay, that is a common concern. We can address that today. Do you happen to smoke? Patient: No, I do not smoke.	Negative for use of alcohol or tobacco.	The patient is a nonsmoker and nondrinker and does not smoke.
Doctor: Hello, I will ask you a few questions. Patient: Hm.	Not otherwise pertinent.	NEUROLOGIC: Normal; Negative for

Doctor: Any headaches or breathlessness? Patient: Nope. Doctor: Any skin problems? Patient: Nope, nothing new. Doctor: Any pain in the chest or anywhere else? Patient: Nope. Doctor: Any other problem that I should know. Patient: I can't think of any.		headaches, syncope, chest pain, or pain.
Doctor: Hello, how are you doing today? Before we get started, I am going to do a basic review of your medical and social history. Patient: Sounds good. Nothing has changed from last visit. Doctor: On your chart we have that you are a non smoker, non drinker, and do not use drugs. We also have written that you do not have any past surgeries or medical conditions. Does that check out? Patient: Yes. That's all the same as last time.	Reviewed and unchanged.	CONSTITUTIONAL: No smoking, alcohol, or drug use.
Doctor: Do you have any family medical history? Patient: No. I don't know of any family health problems.	Noncontributory.	noncontributory. There is no family history.

Table 2: Some examples of low scores generated outputs

The first 3 examples on table 2 all scored around 0.0 on both Rouge and Bertscore. These examples show that when a summary is generated, it normally will output information relevant to the dialogue, while this could also be included in a normal summarization. The nature of the dataset having only one summary means that the computational metrics used will always be low in these cases.

H) Scalability

Scalability is not a major concern for this system, as medical note summarization is typically performed after a consultation rather than in real-time. The generation latency of a few seconds is well within acceptable limits for clinical workflows. However, the system should still be able to handle daily usage volume efficiently. Assuming a typical clinician sees 20–30 patients per day, the model must consistently process multiple dialogues with varying lengths and complexities. While the current implementation was trained and tested on high-end hardware (e.g., NVIDIA A100 GPU), deployment in resource-constrained settings may require optimization strategies such as model distillation or quantization to reduce inference costs. The model currently truncates input at 512 tokens, which was sufficient for the dataset used, but this can be easily adjusted to accommodate longer consultations, provided memory resources are managed appropriately. Overall, the system is scalable for routine clinical usage, but broader adoption would benefit from computational efficiency improvements.

VI. Conclusion

This paper presented a dialogue summarization approach using a fine-tuned BART model on the MTS-Dialog dataset. Our evaluation using ROUGE metrics and BERTScore demonstrated moderate effectiveness with a ROUGE-1 score of 37.07%, ROUGE-2 score of 17.05%, and BERTScore F1 of 36.74%. The model successfully captured key content points but showed limitations in preserving sequential information and fully understanding conversational context.

Improvements that could be explored:

1. **Dialogue-specific architectures**: Developing models that better account for turn-taking dynamics and conversational structure through hierarchical encoding or specialized attention mechanisms.
2. **Role-aware modeling**: Incorporating explicit representation of speaker roles

(Patient vs. Doctor) to improve attribution and contextual understanding.

3. **Domain knowledge integration**: Incorporating medical terminology to enhance summary relevance.

Our work demonstrates both the potential and current limitations of sequence-to-sequence models for dialogue summarization in specialized domains, providing a foundation for more sophisticated approaches to capturing the complexity of multi-turn conversations and medical terminology.

VII. Acknowledgements

Individual contribution report available along with all other deliverables in the submission folders on Surrey Learn

REFERENCES

[1] J. Makosinski, "Clinicians spend a third of their time on clinical documentation," BBH, 2015.

[2] H. Yuan, Z. Yuan, R. Gan, J. Zhang, Y. Xie, and S. Yu, "BioBART: Pretraining and evaluation of a biomedical generative language model," arXiv preprint arXiv:2204.03905, Apr. 2022.

[3] Q. Lu, D. Dou, and T. H. Nguyen, "ClinicalT5: A Generative Language Model for Clinical Text," in Findings of the Association for Computational Linguistics: EMNLP 2022, Abu Dhabi, United Arab Emirates, Dec. 2022, pp. 5436-5444.. [Online]. Available: https://aclanthology.org/2022.findings-emnlp.398.pdf

[4] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension," arXiv preprint arXiv:1910.13461, Oct. 2019.

Original Summary	Generated Summary
This is a return visit to the renal clinic for this patient. She is an 85-year-old woman with history as noted above. Her last visit was approximately four months ago. Since that time, the patient has been considerably more compliant with her antihypertensive medications and actually had a better blood pressure reading today than she had had for many visits previously. She is not reporting any untoward side effect. She is not having weakness, dizziness, lightheadedness, nausea, vomiting, constipation, diarrhea, abdominal pain, chest pain, shortness of breath or difficulty breathing. She has no orthopnea. Her exercise capacity is about the same. The only problem she has is musculoskeletal and that pain in the right buttock, she thinks originating from her spine. No history of extremity pain.	The patient is an 85-year-old female who comes in today for evaluation of rightbuttocks region that she thinks is radiating from her lower back or spine area. The patient states that she has not seen her blood pressure this good in the past. Her last visit to renal clinic was 4 months ago. She states she has done much better than before in terms of managing her hypertension with her current medication. Since that time, she has had no side effects. At the time of her last visit, the patient states she is doing much better on her current medications. No nausea, vomiting
The patient's father also suffered from bilateral hip osteoarthritis.	His father also has arthritis in both of his hips.

Two view x-rays were taken of the shoulder. There are no osseous abnormalities or significant degenerative changes.	X-rays are two-sided. No bone abnormalities or degenerative changes.
The patient has NG tube in place for decompression. She says she is feeling a bit better.	Nasogastric tube placed for decompression.
No known drug allergies.	There are no known drug allergies in the past.

Table 3: Ground truth summary vs. Summary generated by our model