Care4Lang at MEDIQA-Chat 2023: Fine-tuning Language Models for Classifying and Summarizing Clinical Dialogues *

Amal Alqahtani[1,2], Rana Salama[1], Mona Diab[1,3], Abdou Youssef[1]
[1]The George Washington University, DC, USA
[2]King Saud University, Riyadh, KSA
[3]Meta AI, USA
{amalqahtani,raref,mtdiab,ayoussef}@gwu.edu

Abstract

Summarizing medical conversations is one of the tasks proposed by MEDIQA-Chat to promote research on automatic clinical note generation from doctor-patient conversations. In this paper, we present our submission to this task using fine-tuned language models, including T5, BART and BioGPT models. The fine-tuned models are evaluated using ensemble metrics including ROUGE, BERTScore and BLEURT. Among the fine-tuned models, Flan-T5 achieved the highest aggregated score for dialogue summarization.

1 Introduction

Clinical dialogue summarization has emerged as a crucial task in clinical natural language processing (NLP). In a clinical NLP dialogue between a doctor and a patient, relevant information about the patient's medical history, visit summary, health condition, and other details are discussed. Summarizing these dialogues can significantly benefit doctors by enabling them to quickly review key points from past conversations and extract relevant information from clinical notes without having to sift through an extended transcript. Moreover, it can assist doctors in making better decisions by providing them with a concise and accurate conversation record. Therefore, developing effective clinical dialogue summarization systems is of great importance in improving the quality of healthcare delivery. However, clinical dialogue summarization presents unique challenges and goals that differ from summarization in other domains. Clinical summaries need to capture relevant information based on the context of the text, like medical histories, follow-ups, or current diagnoses.

In this paper, we describe our submission to the MEDIQA-Chat shared task (Ben Abacha et al., 2023) the Dialogue2Note Summarization task, task-A. We observe that from the conversation it is important to: (1) capture all the medical conditions and terminology described in the dialogue (eg. cough, fever, shortness of breath etc.); (2) discern all the affirmatives and negatives on medical conditions correctly (no allergies, having a cough for 2 days); and, (3) bias towards copying from the source text while not being completely extractive. Our approach involves studying the effectiveness of fine-tuning pre-trained language models, including T5, GPT, and BART models. We compare the effectiveness of pre-trained models on dialogues, clinical data, and general models.

Section Header	Train	Validation
ALLERGY	60	4
ASSESSMENT	34	4
CHIEF COMPLAINT	77	4
DIAGNOSIS	19	1
DISPOSITION	15	2
EMERGENCY DEPARTMENT COURSE	8	3
EXAM	23	1
FAMILY HISTORY/SOCIAL HISTORY	351	22
GYNECOLOGIC HISTORY	5	1
HISTORY of PRESENT ILLNESS	282	20
IMAGING	6	1
IMMUNIZATIONS	8	1
LABS	2	1
MEDICATIONS	54	7
OTHER HISTORY	2	1
PAST MEDICAL HISTORY	118	4
PAST SURGICAL HISTORY	63	8
PLAN	11	3
PROCEDURES	3	1
REVIEW OF SYSTEMS	60	11
Total	1201	100

Table 1: Overview of Task A Section Headers used for dialogue classification.

2 Shared Task and Dataset

The MEDIQA-Chat 2023 proposed two shared tasks that are related to clinical note summarization and generation (Ben Abacha et al., 2023):

1. **Dialogue2Note Summarization Task:** Given a conversation between a doctor and patient, the task is to generate a clinical note summarizing the conversation with one or multiple note sections (e.g. Assessment, Past Medical History, Past Surgical History). This task

*The first two authors contributed equally to this work.

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 524–528

Doctor: Have you had any surgeries in the past?
Patient: Nope I have not.
Doctor: Anything?
Patient: No.

Section Header: Past Surgical History
Note/Summary: He has not had any previous surgery.

Figure 1: An example of a doctor-patient dialogue, section header and summary.

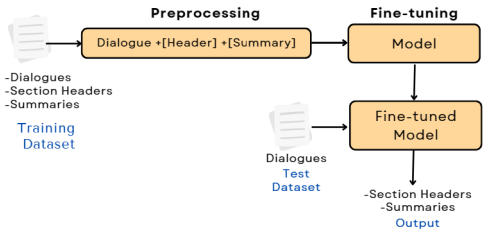

Figure 2: The proposed approach for Task A

includes two subtasks on the generation of specific sections (subtask A) and full notes (subtask B) from doctor-patient conversations.

2. **Note2Dialogue Generation Task:** Given a clinical note, the task is to generate a synthetic doctor-patient conversation related to the information described in the full clinical note.

We participated on **Dialogue2Note (subtask A)**. In this task, given a conversation between a doctor and a patient, the goal is to produce:

1. A section header which is one of twenty normalized section labels, shown in Table 1 to classify the type of conversation.

2. A summarization for the conversation or dialogue into concise and condensed notes. The generated summaries should be tailored to the type of information required based on the section header.

2.1 Dataset

For this task, a doctor-patient conversations dataset is shared by (Ben Abacha et al., 2023). The dataset consists of transcripts of conversational dialogues between doctors and patients. Each dialogue is annotated with associated section headers and corresponding summary notes. The dataset is split into three subsets: a training set, a validation set, and a test set. The training set contains 1,201 pairs of conversations and their associated section headers and . The validation set contains 100 pairs of conversations and their summaries, while the test set contains 200 conversations. Table 1 shows the section headers distributions over the dataset. Figure 1 shows a snippet of the dataset for a doctor-patient conversation along with the section header and the summary.

2.2 Evaluation Metric

For task evaluation, an ensemble of metrics are used to ensure more comprehensive and accurate measures for the quality of generated summaries and headers. ROUGE (Lin, 2004) is a concrete evaluation metric for summarization that conventionally adopts as the standard metric for evaluating summarization tasks. ROUGE involves the F1 scores for ROUGE-1, ROUGE-2, and ROUGE-L. Additionally, BLEU scores(Papineni et al., 2002), used in conjunction with ROUGE score to calculate the semantic correlation of reference and predicted summaries by utilizing token-level matching functions. Furthermore, BERTScore (Zhang et al., 2020) are calculated to capture semantic similarities between summaries and their corresponding reference text at the sentence level. Each of these metrics has its own strengths and weaknesses, and combining them can help mitigate some of these limitations and allow for a more holistic view of the quality of the generated summaries. The ensemble metric can provide a more robust and reliable evaluation that takes into account both the lexical and semantic similarity between summaries and references, as well as the human judgments of quality.

3 Approach

For this submission, we fine-tuned a number of pre-trained language models for implicit classification of headers and note summarization. Since the expected summaries differ in accordance with the associated section header, we fine-tuned the models using supervised training to jointly classify and learn corresponding summaries using the provided training dataset. All models were fine-tuned using Hugging Face Transformers (Wolf et al., 2019). Figure 2 shows a general flow of our approach.

3.1 Data Preprocessing

A key challenge in this task is to generate summaries based on the associated section header, this involves first classifying the dialogue into one of the 20 given headers and accordingly generating a summary. To tackle this challenge, we initially

	Model	ROUGE-1	ROUGE-2	ROUGE-L	Percision	Recall	F1	BLEU	Agg.
Validation Dataset	Flan-T5 Base	0.338	0.147	**0.266**	0.667	0.685	0.670	0.511	0.50
	Flan-T5 Large	0.305	0.120	0.255	**0.691**	0.621	0.645	0.510	0.480
	Flan-T5 SAMSum	**0.348**	**0.149**	0.264	0.660	**0.696**	**0.672**	0.52	**0.510**
	Clinical T5	0.261	0.087	0.226	0.601	0.610	0.596	0.467	0.440
	BioGPT	0.170	0.061	0.125	0.481	0.589	0.519	0.359	0.349
	BART-Large	0.248	0.106	0.168	0.511	0.698	0.580	**0.561**	0.463
	BioBART	0.250	0.107	0.169	0.518	0.689	0.581	0.550	0.460
Test Dataset	Flan-T5 Base	0.344	0.155	0.280	0.671	**0.685**	0.672	0.508	0.508
	Flan-T5 Large	0.332	0.140	0.283	**0.689**	0.644	0.485	0.492	0.508
	Flan-T5 SAMSum	**0.3581**	**0.165**	**0.289**	0.6701	0.70	**0.678**	**0.514**	**0.517**

Table 2: Results of different models fine-tuned for Task A on Validation and Testing Dataset as generated by MediQA shared Task. The precision, recall and F1 scores are based on BERTScore. Agg. represents aggregated results. Best results per dataset are in Bold.

Model	%
Flan-T5 Base	28
Flan-T5 Large	49
Flan-T5 SAMSum	30
Clinical T5	43
BioGPT	23
BART-Large	63
BioBART	69

Table 3: Results of Section Header Classification as a percentage of correctly classified headers.

Model	Accuracy
Flan-T5 Large	0.565
Flan-T5 SAMSum	0.375
Flan-T5 Base	0.345

Table 4: Results of Section Header Classification for the Shared Task A from published results

prepare the data to incorporate both the header and corresponding summary in the input data before fine-tuning. We append labels to each dialogue to tag headers and summaries as follows: "<Dialogue> *Doctor: .. Patient:...* <Header> *header* <Summary> *reference summary*".

3.2 Model Variants

We used a variant of different Sequence-to-Sequence models for our experiments including:
T5 (Raffel et al., 2020) a unified text-to-text language model. We used Flan-T5(Chung et al., 2022) that was further pre-trained on more tasks and languages. Different versions of this model includes, FLan-T5-base[1], FLan-T5-large [2] and FLan-T5-SamSum [3], a Flan-T5 model that is further pre-trained on the SAMSum dataset [4] containing about 16k messenger-like conversations with summaries. In addition to Clinical-T5 (Lu et al., 2022) which is a T5 model pre-trained on clinical text [5]
Bio-GPT (Luo et al., 2022) is a domain-specific generation pretrained model based on the Transformer language model architecture. BioGPT is trained on 15 million PubMed abstracts and is used for processing biomedical text data.
BART (Lewis et al., 2019) for summarization [6]. We also used BioBART (Yuan et al., 2022) which is a BART model pretained on biomedical data [7].

4 Evaluation and Results

Evaluation is performed using the metrics described in (Ben Abacha et al., 2023) and mentioned in Section 2.2. The script provided in the shared task[8] was used for evaluating the fine-tuned models. Evaluation was performed on the validation dataset only as the test dataset references are not available. Table 4 shows the results of our fine-tuned models used for note summarization on the validation dataset. We list the ROUGE-1, ROUGE-2 and ROUGE-L scores, in addition to BERTScore (precision, recall and F1) and BLEU scores. We also include the aggregated score. The Table also includes the final runs scores published by MEDIQA-Chat on the Test dataset. As shown in the table, Flan-T5-SAMSum out-performed all models ex-

[1] https://huggingface.co/google/flan-t5-base
[2] https://huggingface.co/google/flan-t5-large
[3] https://huggingface.co/google/flan-t5-base
[4] https://huggingface.co/datasets/samsum
[5] https://huggingface.co/luqh/ClinicalT5-base
[6] https://huggingface.co/facebook/bart-large-xsum
[7] https://huggingface.co/GanjinZero/biobart-large
[8] https://github.com/abachaa/MEDIQA-Chat-2023/blob/main/scripts/evaluate_summarization.py

cept on BLEU score. On average, Flan-T5 models outperformed other models in header based summarization, they achieved higher scores in ROUGE and BERTScores. Although they didn't perform as well on the number of matching headers, results in Table 3. BART models achieved the highest scores in BLEU scores with more than 4% using BART-Large model. However, there aggregated score was significantly less than Flan-T5 SAMSum. BioGPT achieved the least scores across all metrics and header classification. Given the best models from validation dataset evaluation, we submitted the 3 Flan-T5 models that achieved the best scores; Flan-T5 SAMSum, Flan-T5 Large and Flan-T5 Base. Table 2 shows the accuracy results achieved on the test dataset for the submitted runs. The best submitted models are available on HuggingFace [9] for results replication.

5 Related Work

Automated note generation from doctor-patient conversations has been the subject of several recent studies in natural language processing and healthcare. One line of research has focused on developing machine learning models to automatically generate clinical notes from speech or text data, using deep learning and natural language generation techniques (Zhang et al., 2018; Enarvi et al., 2020; Joshi et al., 2020; Knoll et al., 2022). Other studies have explored the use of voice recognition and speech-to-text technologies to transcribe doctor-patient conversations and generate notes in real time (Zuchowski and Göller, 2022). Additionally, some researchers have investigated using pre-trained language models, such as BERT and GPT, to improve the accuracy and efficiency of note generation (Chintagunta et al., 2021). Overall, these efforts aim to reduce the burden on healthcare providers by automating the tedious task of note-taking and ultimately improving the quality and accessibility of patient records.

6 Conclusion

We utilize several pre-trained models for Task A in MEDIQA-Chat shared task. The main objective of this task is to develop clinical dialogue summarization in accordance with a classified section header for every dialogue. We fine-tuned different models for our experiments. Among the models we used, we found that Flan-T5, originally trained on dialogue datasets, outperformed other models that were trained on clinical data or summarization tasks. Specifically, Flan-T5 SAMSum outperformed all models except for summarization scores. It can also be concluded that summarization models trained on summarizing text, not dialogues, as in BioGPT, performed poorly on summarization tasks. In contrast, BART models performed better than the BioGPT model. Empirically, we found BioGPT to generate text that was not originally in the text, which is considered critical in the context of health records. Finally, since Flan-T5 SAMSum achieved the best results, we anticipate that further unsupervised training for the Flan-T5 language model with clinical dialogues would improve the results.

Limitations

Generating clinical notes or summaries of clinical conversations using NLP technology is a rapidly developing field with great potential. However, there are several limitations to this technology that must be considered. Firstly, NLP models rely on high-quality data to achieve accurate results. In the medical field, obtaining such data can be challenging due to privacy concerns and regulations. Secondly, the complex and technical nature of medical language poses a challenge to NLP models, which may struggle to understand and interpret medical terminology and abbreviations accurately. Additionally, clinical conversations often involve sensitive information that requires careful handling, making it important to ensure the security and privacy of generated clinical notes. This field is considered a safety critical area, where high precision is expected, therefore, the use of NLP models in such clinical settings must be performed with caution and under medical professionals' supervision to ensure the generated notes' accuracy and reliability.

Ethics Statement

When developing an automated system for clinical note generation from doctor-patient conversations, it is crucial to consider various ethical considerations. One such consideration is the privacy and confidentiality of patient information. The system must be designed to comply with regulations and guidelines for protecting patient data. Additionally, there must be explicit consent processes, ensuring that patients understand how their data will be used

[9] https://huggingface.co/Amalq/flan-t5-base-samsum-taskA

and allowing them to opt-out if desired. The system must also be developed fairly and transparent, ensuring it does not perpetuate biases or contribute to health disparities. Moreover, the system must be accurate and reliable, as errors or inaccuracies could lead to incorrect diagnoses or treatments. Overall, it is essential to approach the development of an automated system for clinical note generation with a solid ethical framework to ensure that it aligns with the highest standards of patient care and ethical conduct.

References

Asma Ben Abacha, Wen-wai Yim, Griffin Adams, Neal Snider, and Meliha Yetisgen. 2023. Overview of the mediqa-chat 2023 shared tasks on the summarization and generation of doctor-patient conversations. In *ACL-ClinicalNLP 2023*.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. 2023. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, Dubrovnik, Croatia. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware gpt-3 as a data generator for medical dialogue summarization. In *Machine Learning for Healthcare Conference*, pages 354–372. PMLR.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Seppo Enarvi, Marilisa Amoia, Miguel Del-Agua Teba, Brian Delaney, Frank Diehl, Stefan Hahn, Kristina Harris, Liam McGrath, Yue Pan, Joel Pinto, et al. 2020. Generating medical reports from patient-doctor conversations using sequence-to-sequence models. In *Proceedings of the first workshop on natural language processing for medical conversations*, pages 22–30.

Anirudh Joshi, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2020. Dr. summarize: Global summarization of medical dialogue by exploiting local structures. *arXiv preprint arXiv:2009.08666*.

Tom Knoll, Francesco Moramarco, Alex Papadopoulos Korfiatis, Rachel Young, Claudia Ruffini, Mark Perera, Christian Perstl, Ehud Reiter, Anya Belz, and Aleksandar Savkov. 2022. User-driven research of medical note generation software. *arXiv preprint arXiv:2205.02549*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *CoRR*, abs/1910.13461.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. page 10.

Qiuhao Lu, Dejing Dou, and Thien Nguyen. 2022. ClinicalT5: A generative language model for clinical text. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5436–5443, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. 2022. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6). Bbac409.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. BioBART: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109, Dublin, Ireland. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. *arXiv preprint arXiv:1809.04698*.

Matthias Zuchowski and Aydan Göller. 2022. Speech recognition for medical documentation: an analysis of time, cost efficiency and acceptance in a clinical setting. *British Journal of Healthcare Management*, 28(1):30–36.