clulab at MEDIQA-Chat 2023: Summarization and classification of medical dialogues

Kadir Bulut Ozler and **Steven Bethard**
The University of Arizona
Tucson, AZ, USA
{ozler,bethard}@arizona.edu

Abstract

Clinical Natural Language Processing has been an increasingly popular research area in the NLP community. With the rise of large language models (LLMs) and their impressive abilities in NLP tasks, it is crucial to pay attention to their clinical applications. Sequence to sequence generative approaches with LLMs have been widely used in recent years. To be a part of the research in clinical NLP with recent advances in the field, we participated in task A of MEDIQA-Chat at ACL-ClinicalNLP Workshop 2023. In this paper, we explain our methods and findings as well as our comments on our results and limitations.

1 Introduction

With the increase in accumulated digital medical records in the healthcare field, it is vital to recognize the need of automation in processing medical documents. The automation of medical document processing has been enhancing the efficiency of clinical documentation while enabling healthcare professionals to increase their quality of service. The advancements of medical imaging with machine learning has been integrated into medical decision making systems for the last decades (Erickson et al., 2017; Wernick et al., 2010; Latif et al., 2019), whereas NLP techniques have recently been proven useful for the field (Kreimeyer et al., 2017; Wu et al., 2020). The interest in clinical NLP applications has been growing each year. Especially with the emergence of large language models, there has been an increasing number of research work in exploring their potential applications in the clinical domain.

Use of transformer based large language models has been proven to give impressive performance increases on variety of benchmarks in NLP (Devlin et al., 2018). We have seen dramatic growth on LLM applications across many NLP tasks (Min et al., 2021). LLMs have also shown significant potential on clinical NLP tasks (Kalyan et al., 2022; Lee et al., 2020). Prompt/instruct based language models (Ouyang et al., 2022; Chowdhery et al., 2022) have recently gained attention and already shown promising results in the clinical domain (Singhal et al., 2022). These large language models hold promise especially for generative tasks like summarization (Xie et al., 2023).

MEDIQA-Chat Tasks (Ben Abacha et al., 2023) at ACL-ClinicalNLP Workshop is a shared task that focuses on summarization and generation of patient-doctor conversations. The shared task has 3 subtasks. In the task A, participants aim to generate an artificial section summary from a short patient-doctor dialogue and its associated section header out of 20 possible headers. In the task B, participants aim to generate an artificial clinical note from a full patient-doctor dialogue. In the task C, participants aim to generate an artificial doctor-patient dialogue from a clinical note. We officially participated in task A and reporting results for both task A and task B in this paper. The submission scripts can be found here[1].

2 Dataset

In our experiments we only used the official dataset of the shared task. Table 1 shows the number of samples in each task and split. Task A has 20 different section headers. The label distribution of section headers can be found in Table A1 in Appendix A. As the nature of the medical dialogues, some section headers have very few occurrences in the dataset.

3 Methods

In this section we explain the methods we applied to approach task A and task B. In all our experiments we used transformer (Vaswani et al., 2017)

[1] https://github.com/kbulutozler/MEDIQA-Chat-2023-clulab

Proceedings of the 5th Clinical Natural Language Processing Workshop, pages 144–149
July 14, 2023 ©2023 Association for Computational Linguistics

Task	Training Set	Validation Set	Test Set
A (Ben Abacha et al., 2023)	1201	100	200
B and C (Yim et al., 2023)	67	20	40

Table 1: Official released dataset statistics. Task B and task C are using the same samples. Task A has pairs of section summaries and section headers. Task B and C have pairs of long patient-doctor dialogue and full clinical note.

Model	Description
Clinical-T5-Base	Further MLM pre training of T5-Base with mimic data
Clinical-T5-Sci	Further MLM pre training of SciFive (Phan et al., 2021) model with mimic data
Clinical-T5-Scratch	MLM pre training of randomly initialized T5-Base on only mimic data

Table 2: Clinical-T5 models and their short descriptions.

based large language models to benefit from their transferable knowledge to our domain.

3.1 Task A

In task A, we aim to obtain a section summary of a short doctor-patient dialogue and its corresponding section header. The input is the dialogue and the expected output is section summary and header for the dialogue. Our first approach to this task was to obtain section summary and the header with the same model, however the generative models we used were not able to accomplish this approach accurately. We realized the models were able to summarize the dialogue to some extent, but predicting section headers were usually missing or in wrong grammar. Therefore, we decided to use different models for section summary and section header.

In our hyperparameter search on validation set, we explored several models for both classification and summarization tasks of task A. For the summarization task, we fine-tuned T5-Small and T5-Base (Raffel et al., 2020) along with Clinical-T5 models (Lehman and Johnson, 2023; Goldberger et al., 2000). For the classification task, we fine-tuned roberta-base (Liu et al., 2019) and longformer (Beltagy et al., 2020). In Table 2, we gave short descriptions of Clinical-T5 models that were trained on mimic-iii (Johnson et al., 2016, 2023b; Goldberger et al., 2000) and mimic-iv (Johnson et al., 2023c,a; Goldberger et al., 2000) datasets.

In order to predict section header, we used dialogue-section header pairs as input to our models. In other words, we trained several classification models to predict section headers from given dialogue. We call this input format Dialogue-Header in Table 5. With this approach, we did not obtain reasonable accuracy scores. We considered the possibilities that data size is not enough and dialogue is too long to be informative.

As our final approach to section header prediction, we decided to use section summary and section header pairs as input to the classification models. We call this input format Summary-Header in Table 5. Our hypothesis was that summaries are shorter than dialogues and presumably still contain information about corresponding section header. In order to expand the dataset size to get a better performance, we employed our summarization models that were capable of outputting reasonable section summaries to do data augmentation. For each sample in the dataset, we obtained n+1 section summaries where n is number of summarization models we used and 1 is the original section summary. With this simple method we increased data size n times for the classification model.

In the development stage, our best model for section header prediction was Roberta-base with 100 epochs, 16 batch size and 1e-4 learning rate. Our best model for section summary was Clinical-T5-Sci with 500 epochs, 8 batch size and 1e-4 learning rate.

3.2 Task B

In task B, we aim to obtain full clinical note summary with main section headers from a long doctor-patient dialogue. The input is the dialogue and the expected output is full clinical note summary that includes main section headers. As first approach, we used generative models explained in subsection "Task A" to produce full clinical note from the dialogues. We realized a very weak performance on generating full clinical note summary with accurate section headers. We decided to fine-tune a single

Hyperparameter	Range
learning rate	1e-4, 5e-5, 2e-5
batch size	8, 16, 32
epochs	100, 250, 500
weight decay	0.01
gradient accumulation steps	8

Table 3: Hyperparameter space explored on all experiments.

Model	rogue1	rogue2	rogueL	rogueLsum
T5-Small	0.267	0.086	0.229	0.232
T5-Base	0.313	0.123	0.273	0.272
Clinical-T5-Scratch	0.238	0.085	0.189	0.192
Clinical-T5-Base	0.263	0.110	0.224	0.218
Clinical-T5-Sci	0.329	0.125	0.288	0.289

Table 4: Section summarization results on validation set of Task A. For each model, best combination of hyperparameters have been selected.

generative model for each main section with the hypothesis that more specialized models would lead to better performance.

For each sample in the task B dataset, we extracted 4 main sections from the long clinical notes. The main section headers are "HISTORY OF PRESENT ILLNESS", "PHYSICAL EXAM", "RESULTS", "ASSESSMENT AND PLAN". Therefore we trained 4 single models for each main section header. In each case, the input is the dialogue and the output is summary of a given section header. We then combined them to obtain the full clinical note.

All the experiments on task B has been conducted after the official results. In the development stage, our best models for full clinical note summary were Clinical-T5-Sci with 500 epochs, 16 batch size and 5e-5 learning rate for all main section headers models.

3.3 Post-processing

We applied a simple post-processing method on summaries after analyzing initial summarization results. This method takes a generated summary and removes sentences that have been repeated in the summary already. We aimed to increase text quality with this post-processing operation.

4 Experiments and Results

In all our experiments, we used 4 32GB Nvidia V100 GPUs. We used Huggingface's transformers library (Wolf et al., 2019) as the basis of our experiment scripts. For our efforts to obtain the best models based on validation sets, we explored a hyperparameter space that can be seen in Table 3.

4.1 Task A

For task A, we report our results on validation set and test set. The results of generating section summaries on validation set can be found in Table 4. The metrics we measured are rogue1, rogue2, rogueL and rogueLsum (Lin, 2004). Other metrics that were officially used in the task were excluded due to their computational cost during the extensive experimenting process. As seen from the table, it is interesting to see Clinical-T5-Scratch and Clinical-T5-Base models to underperform in comparison to T5-Small and T5-Base models. Only Clinical-T5-Sci model overperformed T5-Small and T5-Base. Intuitively, we were expecting extra or from scratch training of T5 models on medical domain would perform better on summairizing doctor-patient dialogues. For our official submission, we selected Clinical-T5-Sci model.

The results of predicting section headers on validation sets can be found in Table 5. The metric we measured is accuracy as it is the only official metric for section header prediction. As seen from the table, we can see our data augmentation method that is explained in methods section improves the performance regardless of model choice. On the other hand, even without data augmentation, our approach of Summary-Header input pair in comparison to Dialogue-Header input pair improves the performance regardless of model choice as well. For our official submission, we selected Roberta-

Model	Input Format	Augmentation	Accuracy
Longformer	Dialogue-Header	no	0.243
Longformer	Summary-Header	no	0.258
Longformer	Summary-Header	yes	0.433
Roberta-base	Dialogue-Header	no	0.534
Roberta-base	Summary-Header	no	0.577
Roberta-base	Summary-Header	yes	0.723

Table 5: Section header classification results on validation set of Task A. For each model, best combination of hyperparameters have been selected.

Model	rogue1	rogue2	rogueL	rogueLsum
T5-Small	0.224	0.081	0.193	0.199
T5-Base	0.263	0.104	0.241	0.247
Clinical-T5-Scratch	0.238	0.095	0.187	0.209
Clinical-T5-Base	0.245	0.093	0.201	0.204
Clinical-T5-Sci	0.286	0.112	0.254	0.262

Table 6: Full clinical note summarization results on validation set of Task B. For each model, best combination of hyperparameters have been selected.

base model to be used with Summary-Header input format and with data augmentation.

In the official test set results, we obtained 54% accuracy for predicting section headers that put us at 27th rank out of 31 submissions. For the section summaries, we obtained an aggragate score of 0.4953 that put us at 20th out of 31 submissions. Our post-processing method neither improved nor reduced the summarization score. For all our submissions our code runs and exactly reproduces according to task organizers.

4.2 Task B

For task B, we report our results on validation set. We do not have official test set results for task B as we did not complete the experiments before the submission deadline. The results of generating full clinical notes can be found in Table 6. We used the same rouge metrics as task A to measure our performance. We expected our approach to not be competitive as we used specialized models for each of the 4 main sections whereas full clinical notes have other sections as well. As you can see from the table, we see a similar trend to task A, where T5-Base model outperforms Clinical-T5 models except Clinical-T5-Sci. Since we do not have access to annotated version of the test set, we cannot measure our performance other than validation set results.

5 Ethics Statement

Certain ethical considerations should be taken into account while creating automated systems for processing doctor-patient conversations. The common faults of the proposed systems should be disclosed to system users. Users should be trained to properly use and identify common mistakes of the systems. Since the data to be processed is medical records, it is essential that both data and background models should be stored within strong security measures. Lastly, patients and doctors should be informed that their conversations are recorded and may be used by the automated systems.

6 Discussion and Future Scope

In this paper, we explored the capabilities of LLMs on summarization and classification of doctor-patient dialogues. We experimented for task A and task B but managed to have an official submission on task A. We documented our thought processes and approaches and stated our results. We obtained results that both supported and contradicted our hypothesis. Due to hardware and budget limitations we did not have the chance to explore latest large models. The obvious future work would be on applying public instruct based models if the hardware capacity is enough or private instruct based models if the budget allows. More future work could be on preprocessing of the dialogues. Intuitive postprocessing approaches could also be explored.

7 Limitations

In this shared task, our main limitation has been lack of access to advanced GPUs that can fit massive language models. Given the limited time, we explored a small range of models and hyperparameter space. Considering their proven generative capabilities, these models would be better starting point for producing summaries and dialogues which would allow researchers to focus more on pre/post processing and error analysis. Another limitation has been lack of free access to massive language models that offer paid API. However, using private/commercial models for research purposes is open to debate in NLP community and isn't in the scope of this work.

References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

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.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bradley J Erickson, Panagiotis Korfiatis, Zeynettin Akkus, and Timothy L Kline. 2017. Machine learning for medical imaging. *Radiographics*, 37(2):505–515.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23).

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. Mimic-iv.

Alistair Johnson, Tom Pollard, and Roger Mark. 2023b. Mimic-iii clinical database.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Li wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023c. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1).

Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1).

Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. Ammu: a survey of transformer-based biomedical pretrained language models. *Journal of biomedical informatics*, 126:103982.

Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of biomedical informatics*, 73:14–29.

Jahanzaib Latif, Chuangbai Xiao, Azhar Imran, and Shanshan Tu. 2019. Medical imaging using machine learning and deep learning algorithms: a review. In *2019 2nd International conference on computing, mathematics and engineering technologies (iCoMET)*, pages 1–5. IEEE.

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(4):1234–1240.

Eric Lehman and Alistair Johnson. 2023. Clinical-t5: Large language models built using mimic clinical text.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. *arXiv preprint arXiv:2111.01243*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Long N Phan, James T Anibal, Hieu Tran, Shaurya Chanana, Erol Bahadroglu, Alec Peltekian, and Grégoire Altan-Bonnet. 2021. Scifive: a text-to-text transformer model for biomedical literature. *arXiv preprint arXiv:2106.03598*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2022. Large language models encode clinical knowledge. *arXiv preprint arXiv:2212.13138*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Miles N Wernick, Yongyi Yang, Jovan G Brankov, Grigori Yourganov, and Stephen C Strother. 2010. Machine learning in medical imaging. *IEEE signal processing magazine*, 27(4):25–38.

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*.

Stephen Wu, Kirk Roberts, Surabhi Datta, Jingcheng Du, Zongcheng Ji, Yuqi Si, Sarvesh Soni, Qiong Wang, Qiang Wei, Yang Xiang, et al. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association*, 27(3):457–470.

Qianqian Xie, Zheheng Luo, Benyou Wang, and Sophia Ananiadou. 2023. A survey on biomedical text summarization with pre-trained language model. *arXiv preprint arXiv:2304.08763*.

Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. Acibench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation. *Submitted to Nature Scientific Data*.

A Appendix

Section Header	**Number**
ALLERGY	64
ASSESSMENT	38
CC	81
DIAGNOSIS	20
DISPOSITION	17
EDCOURSE	11
EXAM	24
FAM/SOCHX	373
GENHX	302
GYNHX	6
IMAGING	7
IMMUNIZATIONS	9
LABS	3
MEDICATIONS	61
OTHER HISTORY	3
PASTMEDICALHX	122
PASTSURGICAL	71
PLAN	14
PROCEDURES	4
ROS	71

Table A1: Section header label space and its statistics in training and validation data of task A.