KnowLab at RadSum23: comparing pre-trained language models in radiology report summarization

Jinge Wu
University College London
jinge.wu.20@ucl.ac.uk

Daqian Shi
University of Trento
daqian.shi@unitn.it

Abul Hasan
University College London
a.kalam@ucl.ac.uk

Honghan Wu
University College London
honghan.wu@ucl.ac.uk

Abstract

This paper presents our contribution to the RadSum23 shared task organized as part of the BioNLP 2023. We compared state-of-the-art generative language models in generating high-quality summaries from radiology reports. A two-stage fine-tuning approach was introduced for utilizing knowledge learnt from different datasets. We evaluated the performance of our method using a variety of metrics, including BLEU, ROUGE, Bertscore, CheXbert, and RadGraph. Our results revealed the potentials of different models in summarizing radiology reports and demonstrated the effectiveness of the two-stage fine-tuning approach. We also discussed the limitations and future directions of our work, highlighting the need for better understanding the architecture design's effect and optimal way of fine-tuning accordingly in automatic clinical summarizations.

1 Introduction

Summarization of radiology reports is a useful tool for both doctors and patients. It allows doctors to quickly prioritize and extract essential information from long documents, saving time and improving patient outcomes. This is especially beneficial in medical settings, where time is often limited and there may be a large volume of documents to review.

The rapid development of natural language processing (NLP) techniques, particularly large generative language models such as GPT (Radford et al., 2018), have significantly advanced the field of automatic text summarization. However, there is still much to explore when it comes to generating summaries for specific domains or tasks. For example, generating radiology reports is a challenging task that requires specialized knowledge and language expertise. Radiology reports are typically written in a specific format and include complex medical terminology, which makes it difficult for traditional summarization techniques to produce accurate and comprehensive summaries.

In the past, several methods have been proposed for radiology report summarization. Chen et al. (2018) developed a method for generating summaries of radiology reports using an attention-based neural network. Zhang et al. (2018) explored the problem using an augmented pointer-generator model resulted in high overlap with human-generated references. Based on this work, MacAvaney et al. (2019) then built an ontology-aware pointer-generator, which led to improved summarization quality. Furthermore, Delbrouck et al. (2022b) proposed ViLMedic, which is a replicable pipeline that can reproduce the latest results in various medical tasks using multimodal data resources (images and texts) for generating radiology reports.

This paper investigates transfer learning methodologies for adapting large generative language models to the specific domain of radiology report generation. Our main contribution involves a two-stage fine-tuning procedure on a large corpus of radiology reports to handle the specialized language and structure of these reports. We conducted experiments using two clinical datasets: MIMIC-III, which includes various kinds of radiology reports, and MIMIC-CXR, which includes only chest x-ray radiology reports. The work and results are presented as part of the RadSum23 shared task at BioNLP 2023 (Delbrouck et al., 2023).

2 Task Description and Dataset

The goal of this research is to generate a summary (*"Impression"* section in Table 1) that highlights the key observations and conclusions of the radiology study. They are generally written by a radiologist after analyzing medical images such as X-rays or CT scans (*"Findings"* section in Table 1). Our group focused on the textual data and work only on the text-based radiology reports.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 535–540
July 13, 2023 ©2023 Association for Computational Linguistics

Findings: the patient is status post intubation. note is made of degenerative disc disease involving multiple levels of c-spine, however, there is no evidence of fracture of the cervical spine. again note is made of multiple skull base fractures, as noted on the prior head ct. note is made of multiple hemorrhagic contusions in posterior fossa, as noted in the prior head ct. the lung apices are unremarkable. note is made of small amount of deep tissue emphysema posterior to the clavicles.
Impression: multiple skull fractures and hemorrhagic contusions as noted on the prior head ct. no evidence of c-spine fractures. djd of the c-spine. small amount of distal air posterior to the clavicles. the information has been communicated with ed physicians in person.

Table 1: A radiology report sampled from MIMIC-III.

Two datasets are used in this challenge: the MIMIC-III and MIMIC-CXR Radiology Report Summarization datasets (Johnson et al., 2016, 2019). MIMIC-III radiology reports contains free-text radiology reports from the Beth Israel Deaconess Medical Center. It contains radiology reports from various modality-anatomy including head, abdomen chest, etc. MIMIC-CXR radiology reports, which is a subset of MIMIC-IV dataset, that only contains chest radiology reports.

In terms of the evaluation, the test-sets has been split into two by the organizers. The details of data description can be found in Table 2.

Dataset	**Train**	**Val**	**Test**
MIMIC-III	59,320	7,413	6,526
MIMIC-III(hidden)			6,531
MIMIC-CXR	125,417	991	1,624
MIMIC-CXR(hidden)			1,000

Table 2: Data description of RadSum23 shared task.

3 Methods

We applied pre-trained language models that are specific for generative tasks. The models are fine-tuned for the task of radiology report summarization. Specifically, we fine-tuned two different types of pre-trained models - BART (Lewis et al., 2019) and T5 (Raffel et al., 2020). By fine-tuning these models on a large corpus of radiology reports, we aim to generate high-quality summaries that capture the most important information in the reports.

The BART model, developed by Facebook (Lewis et al., 2019), is a denoising auto-encoder designed for pre-training sequence-to-sequence (Seq2Seq) models. It combines the state-of-the-art (SOTA) performance of both the BERT (Devlin et al., 2018) and GPT (Radford et al., 2018) models, inheriting the bidirectional encoder and left-to-right decoder models' benefits. Because it is constructed on the Seq2Seq Transformers architecture, BART is an excellent choice for abstractive summarization, as it can generate novel text and paraphrase the input text. BART's original pre-training involved masked language modeling (MLM) with six noises, an enhancement of the BERT model's single noise masking strategy, making it less likely to learn biased information.

T5, short for Text-To-Text Transfer Transformer, was proposed by Google and utilizes the standard transformer architecture pre-trained on denoising, where spans of text are replaced with a drop token. T5 was trained using a "text-to-text" approach, which means it can perform a wide range of tasks by converting input text into output text, such as language translation, summarization, question answering, and even programming tasks. T5 is also known for its ability to perform well with few-shot learning, meaning it can quickly adapt to new tasks with just a few examples.

BART and T5 are both transformer-based architectures with subtle differences in their layer configurations. T5 is trained as a causal language model (i.e., predicting next tokens), while BART is trained on a masked language modeling objective. Their performances vary depending on the specific tasks.

To optimize the performance and generalizability of our approach, we employed a two-stage fine-tuning process using two different datasets - MIMIC-III and MIMIC-CXR. In the first stage, we fully fine-tuned the pre-trained model on the MIMIC-III dataset using the typical fine-tuning approach. This involved training the model on the dataset to adapt it to the specific task of radiology report summarization. In the second stage, we performed another round of fine-tuning on the MIMIC-CXR dataset by freezing the last two layers in the encoder and decoder. It is assumed that freezing the last two layers can help prevent the model from overfitting the training data. Lower layers tend to learn more generalizable feature representations, while higher layers may be more prone to overfitting on the specific patterns of the training data. By freezing the last two layers, we can limit the degree of fitting to the training data, enabling the model to generalize better on new, unseen data.

Overall, the two-stage fine-tuning process is a key component of our approach, allowing us to leverage the strengths of both datasets and optimize the performance of the model on the task of radiology report summarization.

3.1 Experiment Setup

For comparison, we fine-tuned BART base model[1] and BioBART model [2](Yuan et al., 2022), T5 base model[3] and SciFive model [4](Phan et al., 2021). BioBART is a generative language model that adapts BART to the biomedical domain by pretraining on large PubMed corpora. SciFive is a domain-specific T5 model that pretrained on large biomedical corpora.

We tuned the hyper-parameters during the training phase to optimize the performance of our approach. The initial learning rate is 2e-5. The maximum epochs used for training is set to 20 with batch size of 16. The maximum length for input data is 1024. The maximum length for output is 128. The beam size is set to 5, and no_repeat_ngram_size is set to 2. Other hyper-parameters are set as their default values.

3.2 Evaluation Metrics

According to the instructions, we consider five evaluation metrics for this work, including BLEU, Rouge, BERT score, CheXbert, and RadGraph (Papineni et al., 2002; Lin, 2004; Zhang et al., 2019; Smit et al., 2020; Delbrouck et al., 2022a).

BLEU and Rouge scores measures the overlap between the generated summary and references based on n-grams (Papineni et al., 2002; Lin, 2004). The main difference between Rouge and BLEU is that Rouge emphasizes recall and BLEU focuses on precision. This means that Rouge is more focused on capturing the important content of the summary, while BLEU is more focused on ensuring that the summary is grammatically correct and fluent. BERT score measures the similarity between the embeddings of the machine-generated text and the references based on the contextual embeddings generated by the BERT model (Zhang et al., 2019). CheXbert revised the BERT score by adding expert annotations on Chest X-rays (Smit et al., 2020). RadGraph is a metric used for evaluating radiology report generation (Delbrouck et al., 2022a). It provides better domain-adjusted evaluation based on a novel Information Retrieval(IE) method (i.e. entities and relationships) from MIMIC-CXR dataset.

4 Experiments and Results

Table 3 and Table 4 present the experimental results obtained from the MIMIC-III and MIMIC-CXR datasets. In the MIMIC-III experiments, we performed fine-tuning on four models and evaluated their performance on the two test sets: MIMIC-III test set and the MIMIC-III hidden test set. The BART model achieved the highest overall performance on both test sets, while the BioBART model achieved the highest Bertscore on the MIMIC-III test set and the highest RadGraph score on the MIMIC-III hidden test set.

For the MIMIC-CXR experiments, we performed a second round of fine-tuning on the models that were previously fine-tuned on MIMIC-III. The baseline model in Table 4 refers to the model that underwent only one round of fine-tuning using the MIMIC-CXR data. To evaluate the performance, we included the CheXbert F1 score as it is commonly used for evaluating chest X-ray radiology reports. Overall, the T5 model achieved the best performance on the MIMIC-CXR test set, while the BART and BART_freeze models had higher RougeL and Bertscore. In terms of the MIMIC-CXR hidden test set, T5 achieved the best performance, while T5_freeze had the highest F1 CheXbert score. The results showed that our two-stage fine-tuning approach yielded better performance than the baseline model in both test sets.

In summary, our experimental results demonstrated that BART and T5 models have different strengths and may perform better on different datasets or evaluation metrics. This may due to various factors, such as the size and complexity of the dataset, the quality of the training data, etc.

There is considerable potential for future improvements to our approach. Firstly, our results suggest that freezing the last two layers of the models may yield some advantages, but this finding requires further investigation to fully understand the effects of this approach. Secondly, we acknowl-

[1] https://huggingface.co/facebook/bart-base

[2] https://huggingface.co/GanjinZero/biobart-base

[3] https://huggingface.co/t5-base

[4] https://huggingface.co/razent/SciFive-base-Pubmed

Models	BLEU4	RougeL	Bertscore	F1-RadGraph
BART	**13.86**	**32.22**	54.61	**32.49**
BioBART	13.24	32.14	**54.76**	32.45
T5	13.40	31.48	54.27	31.74
scifive	12.59	31.99	54.49	32.42
BART(hidden)	**13.69**	**32.12**	**55.65**	33.22
BioBART(hidden)	13.23	32.02	55.64	**33.39**
T5(hidden)	13.27	31.49	55.03	32.06
scifive(hidden)	12.37	31.79	55.24	33.03

Table 3: MIMIC-III test results.

Models	BLEU4	RougeL	Bertscore	F1-CheXbert	F1-RadGraph
BART	22.92	**46.53**	63.85	74.56	47.98
BART_freeze	21.40	46.34	**66.63**	73.48	46.89
T5	**22.97**	46.15	63.43	**75.14**	**48.04**
T5_freeze	22.54	46.27	63.24	74.00	47.82
baseline	14.12	27.67	47.44	66.58	29.51
BART(hidden)	11.92	32.00	53.30	66.01	38.21
BART_freeze(hidden)	9.43	30.22	51.76	64.91	35.38
T5(hidden)	14.41	**33.63**	**54.72**	67.20	**39.98**
T5_freeze(hidden)	13.69	33.08	54.47	**68.45**	39.69
baseline	**17.07**	32.28	54.45	67.77	37.38

Table 4: MIMIC-CXR test results.

edge that there are limitations to our current training process, and we believe that improvements can be made to further optimize the performance of our models. Specifically, we plan to explore alternative training strategies and hyper-parameter tuning methods to improve the performance of our models on the MIMIC-III and MIMIC-CXR datasets. Additionally, we recognize the importance of testing our models on larger datasets and expanding the scope of our evaluations to include other natural language processing tasks. Finally, we will continue to monitor developments in the field and adapt our methods accordingly to ensure that we remain at the forefront of natural language processing research.

5 Conclusion

This paper presented our investigation into transfer learning for the radiology report summarization task. Specifically, we conducted a comprehensive evaluation of different pre-trained language models which were built by deploying encoder-decoder architectures. We introduced a two-stage fine tuning methodology which involved BART, T5, and their variants in the biomedical domain, and compared their performances on the summarization task.

We also evaluated the effectiveness of our approach using various metrics, such as BLEU, ROUGE, Bertscore, CheXbert, and RadGraph, and demonstrated that our method achieved high performance on the summarization task. In fact, our approach was ranked third on the leaderboard of the RadSum23 MIMIC-CXR hidden test set, highlighting the competitiveness of our approach compared to other methods.

Acknowledgements

We would like to thank the BioNLP 2023 RadSum23 Shared Task organizers who provide the valuable resources including the baseline pipeline ViLMedic (Delbrouck et al., 2022b) and this exciting tasks (Delbrouck et al., 2023).

References

Ninghui Chen, Hongfang Liu, and Xiaojun Wan. 2018. Radiology report summarization using attention-based neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 26(11):2123–2134.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis P Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. *arXiv preprint arXiv:2210.12186*.

Jean-Benoit Delbrouck, Khaled Saab, Maya Varma, Sabri Eyuboglu, Pierre Chambon, Jared Dunnmon, Juan Zambrano, Akshay Chaudhari, and Curtis Langlotz. 2022b. Vilmedic: a framework for research at the intersection of vision and language in medical ai. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 23–34.

Jean-Benoit Delbrouck, Maya Varma, Pierre Chambon, and Curtis Langlotz. 2023. Overview of the radsum23 shared task on multi-modal and multi-anatomical radiology report summarization. In *Proceedings of the 22st Workshop on Biomedical Language Processing*, Toronto, Canada. Association for Computational Linguistics.

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*.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317.

Alistair EW 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):1–9.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

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

Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, Ish Talati, and Ross W Filice. 2019. Ontology-aware clinical abstractive summarization. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 735–744. ACM.

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.

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*.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. 2020. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. In *EMNLP 2018 Workshop on Health Text Mining and Information Analysis*.

Models	**BLEU4**	**RougeL**	**Bertscore**	**F1-RadGraph**
shs-nlp	18.36	35.32	57.26	36.94
utsa-nlp	16.05	34.41	57.08	36.31
aimi	16.61	33.43	55.54	35.12
sinai	17.38	32.32	55.04	33.96
knowlab	13.23	32.02	55.64	33.39
nav-nlp	15.13	32.39	55.34	33.37
elirf	18.06	30.19	53.94	32.58

Table 5: Leaderboards for MIMIC-III hidden test-set (6531 sample)

Models	**BLEU4**	**RougeL**	**Bertscore**	**F1-RadGraph**
utsa-nlp	15.99	34.07	56.30	35.25
shs-nlp	17.33	33.93	55.49	34.93
nav-nlp	15.31	32.33	54.49	32.68
sinai	17.12	31.62	54.33	32.65
knowlab	13.86	32.22	54.91	32.49
elirf	17.41	29.57	52.24	31.40
aimi	1.25	24.45	45.54	21.24

Table 6: Leaderboards for MIMIC-III test-set (6526 sample)

Models	**BLEU4**	**RougeL**	**Bertscore**	**F1-cheXbert**	**F1-RadGraph**
ku-dmis-msra	18.62	34.57	55.90	72.36	43.20
utsa-nlp	16.33	34.97	55.54	69.41	42.86
knowlab	14.41	33.63	54.72	67.20	39.98
shs-nlp	14.59	32.43	53.99	68.99	38.40
aimi	5.15	31.84	47.83	64.18	32.05
iuteam1	1.99	26.08	46.75	40.28	27.35
e-health csiro	4.12	21.58	43.86	53.46	23.86
nlpaueb	5.03	19.87	41.84	50.69	23.26

Table 7: Leaderboards for MIMIC-CXR hidden test-set (1000 sample)

Models	**BLEU4**	**RougeL**	**Bertscore**	**F1-cheXbert**	**F1-RadGraph**
utsa-nlp	25.87	47.86	64.74	77.93	51.84
ku-dmis-msra	25.58	47.75	64.80	76.29	50.96
shs-nlp	25.32	47.48	63.61	74.34	49.00
knowlab	22.97	46.15	63.43	75.14	48.04
e-health csiro	17.97	44.14	61.47	71.67	44.95
iuteam1	10.10	40.44	56.44	58.01	39.48
nlpaueb	11.69	36.80	55.50	59.53	36.92

Table 8: Leaderboards for MIMIC-CXR test-set (1624 sample)