SINAI at RadSum23: Radiology Report Summarization Based on Domain-Specific Sequence-To-Sequence Transformer Model

Mariia Chizhikova, Manuel C. Díaz Galiano
L. Alfonso Ureña López and **M. Teresa Martín Valdivia**
Department of Computer Science, University of Jaén
Campus Las Lagunillas, s/n, 23071, Jaén, Spain
mchizhike@ujaen.es

Abstract

This paper covers participation of the SINAI team in the shared task 1B: Radiology Report Summarization at the BioNLP workshop held on ACL 2023. Our proposal follows a "sequence-to-sequence" approach which leverages pre-trained multilingual general domain and monolingual biomedical domain pre-trained language models. The best performing system based on domain-specific model reached 33.96 F1RadGraph score which is the fourth best result among the challenge participants. This model was made publicly available on HuggingFace. We also describe an attempt of Proximal Policy Optimization Reinforcement Learning that was made in order to improve the factual correctness measured with F1RadGraph but did not lead to satisfactory results.

1 Introduction

Radiology explorations constitute an essential pillar in diagnosis and treatment of many diseases nowadays. The radiology report documents communicate crucial findings in a radiology study to the referring physician and/or patient. A standard radiology report ends with an Impression section that summarizes the most relevant findings. This section, due to the communicative purpose of the radiology reports, may be considered its most significant part (Zhang et al., 2018). Despite the fact of communication being crucial in assuring quality of healthcare, studies have shown that relevant findings may often be eluded (Gershanik et al., 2011). In the recent years, this problem drew the attention of the research community, allowing the clearness and conciseness of the communication to become of greater focus in radiology.

The Task 1B at the BioNLP workshop held on ACL 2023 draws the community effort to designing systems capable of automating the generation of radiology impressions. The work carried out in this growing field in the recent years mostly focused on Chest X-ray due to availability of large curated datasets of chest radiography exploration reports, namely MIMIC-CXR (Johnson et al., 2019) and Open-i Chest X-ray (Demner-Fushman et al., 2016). The goal of the task 1B at the BioNLP workshop is to foster development of automatic radiology report summarization system and expanding their applicability by incorporating seven different modalities and anatomies in the provided data.

This paper covers participation of the SINAI team from the University of Jaén in the first challenge of the shared task, which required the development of a mono-modal radiology report summarization system that would maximize the factual correctness of its output. We propose to automate the generation of radiology impressions with "sequence-to-sequence" learning that leverages the power of publicly available pre-trained models, both general domain and biomedical domain-specific. More specifically, we opted for fine-tuning `flan-t5-base`, a general domain text-to-text transfer transformer fine-tuned with instructions (Chung et al., 2022), and `SciFive-base-Pubmed_PMC`, a model with the same architecture trained on large-scale biomedical corpora (Phan et al., 2021). We also cover a Proximal Policy Optimization experiment we carried out with the objective of maximizing factual correctness of generated impressions. Our best performing system based on domain-specific model reached 33.96 F1RadGraph score which is the fourth best result among the challenge participants.

The remainder of the paper is organized as follows: Section 2 offers a brief description of the data provided for this task. Section 3 discusses the metrics used for system evaluation and Section 4 describes the systems presented by our team for the official evaluation with the results that are disclosed in Section 5. Finally we draw the conclusions of our work in Section 6.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 530–534

2 Data

The data provided for the mono-modal challenge contained 73,255 radiology reports divided into three subsets: training (5,9317 reports), development (7,413) and test (6,526 reports) (Delbrouck et al., 2023). The official evaluation was carried out on a hidden test dataset comprised of 6,531 samples. Training and hidden test sets included reports describing 2 different types of radiology explorations, namely CT and MRI performed on 7 different anatomic regions: head, abdomen, chest, spine, neck, sinus and pelvis. Whereas the development and test datasets lacked samples of spine, abdomen, pelvis and neck MRI as well as sinus CT reports. Table 1 summarizes the text length statistics across train, validation and test subsets. As for the hidden test set, the provided findings are generally shorter: maximum 512 tokens per entry and minumum 8, being the mean length 69.299 with the standard deviation of 42.317.

	Train	Val	Test
Findings			
Max	1203	750	794
Min	2	4	2
Mean	112.969	113.719	123.209
SD	65.764	67.007	70.963
Impression			
Max	353	277	284
Min	0	2	0
Mean	47.272	47.675	46.312
SD	35.298	35.752	32.743

Table 1: Text length statistics

2.1 Pre-processing

Given that the data contained some empty impressions, the first step of text pre-processing consisted in the elimination of 3 empty impressions from training set and 1 from the validation set. Next, we eliminated all tokens that followed the final of the last sentence both in findings and impressions. This step was implemented due to presence of some unidentified character strings at the end of some sequences. Then, we substituted all the triple underscores that indicated presence of sensitive information eliminated during the anonymization process with a new special token `<SI>` to prevent model from generating repetitive underscores. Finally, a task-specific pre-fix `summarize:` was added at the beginning of each finding section and the dataset was shuffled. The rest of text pre-processing relied on model's tokenizer.

3 Evaluation metrics

The submissions for the mono-modal challenge were evaluated using BLEU4, ROUGE-L, BertScore and F1RadGraph. BLEU4 is a metric traditionally used for evaluation of machine translation systems that assesses the proximity of translations to their corresponding labels. It does not quantify the clarity or grammatical accuracy of the model's outputs, but instead employs statistical measures to guarantee that all the words in the generated outputs are also present in the targets (Papineni et al., 2002). ROUGE (Recall-Oriented Understudy for Gisting Evaluation) is one of the most commonly used metrics for automatic text summarization evaluation by comparing generated summaries to other (ideal) summaries created by human. The ROUGE-L measures the longest matching sequences of words by looking for the longest common substrings in the generated and reference summaries (Lin, 2004). BERTScore computes a similarity score for each token in the candidate sentence with each token in the reference sentence (Tianyi et al., 2020). Finally, the F1RadGraph leverages a system trained on the RadGraph dataset that extracts relevant entities and relations between them. The proposed F1-style reward focuses on entities and whether or not there is a connection between them in the reference summary compared to the prediction (Delbrouck et al., 2022).

4 Methods

With the objective of measuring the impact of domain-specific pre-training on the T5 model's performance on radiology report summarization we carried-out fine-tuning of two models of the same architecture for this task. We also attempted to optimize factual correctness of the impressions generated by the general domain model through Proximal Policy Optimization.

4.1 Text-to-text Transfer Transformer

The first system is based on `flan-t5-base`[1], which is a general domain T5 model subjected to instruction fine-tuning on 1.8K tasks (Chung et al., 2022). Authors report significant improvements of performance across a wide range of prompting setups

[1] https://huggingface.co/google/flan-t5-base

and evaluation tasks compared to prior version of T5 (Raffel et al., 2020).

4.2 Sci-Five

The second proposed system is based on `SciFive-base-Pubmed_PMC` [2], a model that was initialized with pre-trained weights from the base T5 model and trained for extra 200k steps to optimize it in the context of biomedical literature (Phan et al., 2021). Authors reported that this model achieved SOTA or near-SOTA results in such tasks as Named Entity Recognition (NER), Relation Extraction (RI) and Natural Language Inference (NLI) and state the need of evaluation of this system on tasks that require a longer sequence of text as an output, such as document summarization.

4.3 Fine-tuning for summarization task

The models were fine-tuned for the summarization task on 2 NVIDIA A100 40GB GPUs with the ROUGE-L as target metric. We also experimented with changing the objective metric to F1RadGraph (Delbrouck et al., 2022), but that resulted in a lower performance in terms of text quality (1 decimal point drop in BLEU4 and ROUGE-L during official evaluation on test set).

In order to maximize the resulting performance of out system we carried out a hyperparameter optimization reliant on the Optuna framework that provides efficient parameter sampling and trial pruning algorithms (Akiba et al., 2019). For each model we searched for the optimal values of the learning rate and weight decay that could be float numbers in range between 1e-5 and 5e-5 or 1e-12 and 1e-1 respectively. The batch size per device was set on 16 and an early stopping strategy was implemented to stop each trial and restore best model weights if the target metric does not improve for 5 epochs. Table 2 summarizes the optimal hyperparameters selected for each model after 5 trials.

	Flan-T5	**Sci-Five**
Learning rate	0.000013	0.00003
Weight decay	0.0000049	0.0000029
Epochs	22	46

Table 2: Hyperparameters selected during the optimization

[2] `https://huggingface.co/razent/SciFive-base-Pubmed_PMC`

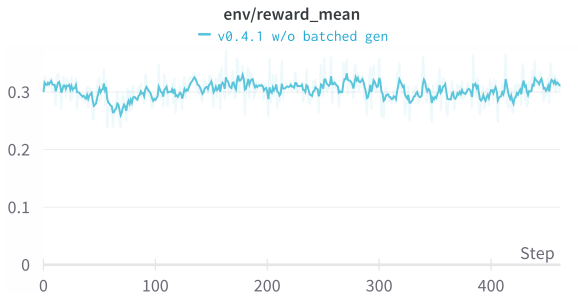

Figure 1: Reward mean dynamic during the PPO learning

4.4 Proximal Policy Optimization

With an intention of maximizing factual correctness of the generates summaries, we subjected a fine-tuned general domain T5 model to a reinforcement learning (RL) procedure using the Proximal Policy Optimization (PPO) method that used F1RadGraph as the metric to maximize. In contrast to conventional policy gradient techniques that update the model's parameters once per data sample, PPO implements a new objective function that permits the execution of multiple rounds of minibatch updates (Schulman et al., 2017).

Our experiment relied on Transformer Reinforcement Learning (TRL) python library (von Werra et al., 2020) and consisted in executing 2 epochs of RL model optimization with 4 PPO epochs per each batch of data. That is, for each batch of 128 training samples model generated summaries that were evaluated against the reference summaries using the F1RadGraph metric. The predictions, the references and the reward were passed alongside with the reference model to the PPO algorithm that carried out 4 epochs of model optimization with minibatch size of 16 samples.

The experiment was executed on 4 NVIDIA A100 40GB GPUs and in its progress no significant F1RadGraph improvement was observed, as can be seen on Figure 1.

The absence of any positive outcome from the PPO training could be caused by several reasons. Firstly, all layers were shared between the reference model and the policy, which made possible running the experiment on the GPU setup described earlier, but could prevent the policy from improving. Secondly, despite the fact that F1RadGraph was validated as an appropriate reward for RL with self-critical sequence training method (Delbrouck et al., 2022), PPO might not be suitable for maxi-

mizing it. Finally, the issue might be related to the "sequence-to-sequence" nature of our approach or the T5 model itself.

5 Results

Our team made a total of three submission on the MIMIC-III hidden test set that corresponded to the three systems we have described in 4. A fine-tuned biomedical domain-specific Sci-Five model resulted to be the best performing one in terms of F1RadGraph, which is the reference metric of this challenge reaching 32.48 points. Table 3 summarizes the results obtained by all the presented systems.

While the performance of all three models is relatively consistent in terms of BERTscore, it can be concluded that the RL experiment resulted in a slight decrease of the quality of model's generations. The results also suggests that domain-specific training is beneficial for such complex tasks as radiology report summarization.

6 Conclusions

This article describes the contribution of the SINAI team from the University of Jaén to the ACL2023-BioNLP shared task. More specifically, we introduce an impression section generation systems that follow a "sequence-to-sequence" approach to the problem and leverage pre-trained publicly available models. In our study we compare the performance of general-domain T5 model with Sci-Five, the result of a domain adaptation process though additional training on biomedical datasets. After fine-tuning on the datasets provided for this challenge, a system based on Sci-Five resulted the best performing one, achieving 33.96 F1RadGraph score. We made is model available for the research community on HugginFace[3].

We also describe an experiment of subjecting the fine-tuned general-domain model to a RL learning with PPO method that, unfortunately did not resulted in a performance improvement. Thus, the future work will focus on improvement of factual correctness of the output though RL techniques in order to overcome the presented problem.

Limitations

The limitations of the presented system derive directly from the pre-trained models they are based on. Flan-T5 was pre-trained on a large-scale text dataset and was not assessed for existing biases. As a result the model itself is potentially vulnerable to generating equivalently inappropriate content or replicating inherent biases in the underlying data (Raffel et al., 2020). Thus, it is not recommended to utilize such systems in any application without first conducting a thorough evaluation of the safety and fairness concerns that are specific to the particular application in question.

[3] https://huggingface.co/chizhikchi/sci-five-radsum23

Acknowledgements

This work has been partially supported by WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government, and projects CONSENSO (PID2021-122263OB-C21), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) funded by Plan Nacional I+D+i from the Spanish Government, and project PRECOM (SUBV-00016) funded by the Ministry of Consumer Affairs of the Spanish Government.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

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*.

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

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.

Dina Demner-Fushman, Marc D Kohli, Marc B Rosenman, Sonya E Shooshan, Laritza Rodriguez, Sameer Antani, George R Thoma, and Clement J McDonald. 2016. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2):304–310.

System	BLEU4	ROUGEL	BERTscore	F1RadGraph
Flan-T5	17,1	31,71	55,22	33,71
Sci-five	**17,38**	**32,32**	**55.04**	**33.96**
Flan-T5 PPO	13,5	31,21	54.53	32.48

Table 3: Official evaluation results on the hidden test set

Esteban F Gershanik, Ronilda Lacson, and Ramin Khorasani. 2011. Critical finding capture in the impression section of radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 465. American Medical Informatics Association.

Alistair EW Johnson, Tom J Pollard, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Yifan Peng, Zhiyong Lu, Roger G Mark, Seth J Berkowitz, and Steven Horng. 2019. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*.

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

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*.

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.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Zhang Tianyi, Kishore Varsha, Felix Wu, Q. Weinberger Kilian, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert. 2020. Trl: Transformer reinforcement learning. `https://github.com/lvwerra/trl`.

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*.