Overview of the RadSum23 Shared Task on Multi-modal and Multi-anatomical Radiology Report Summarization

Jean-Benoit Delbrouck and **Maya Varma** and **Pierre Chambon** and **Curtis P. Langlotz**
Stanford University
{jbdel,mvarma2,pchambon,langlotz}@stanford.edu

Abstract

Radiology report summarization is a growing area of research. Given the Findings and/or Background sections of a radiology report, the goal is to generate a summary (called an Impression section) that highlights the key observations and conclusions of the radiology study. Recent efforts have released systems that achieve promising performance as measured by widely used summarization metrics such as BLEU and ROUGE. However, the research area of radiology report summarization currently faces two important limitations. First, most of the results are reported on private datasets. This limitation prevents the ability to reproduce results and fairly compare different systems and solutions. Secondly, to the best of our knowledge, most research is carried out on chest X-rays. To palliate these two limitations, we propose a radiology report summarization (RadSum) challenge on i) a new dataset of eleven different modalities and anatomies pairs based on the MIMIC-III database ii) a multimodal report summarization dataset based on MIMIC-CXR enhanced with a brand-new test-set from Stanford Hospital. In total, we received 112 submissions across 11 teams.

1 Introduction

The radiology report documents and communicates crucial findings in a radiology study. A standard radiology report usually consists of a Background section that describes the exam and patient information, a Findings section, and an Impression section (Kahn Jr et al., 2009). In a typical workflow, a radiologist first dictates the detailed findings into the report and then summarizes the salient findings into the more concise Impression section based also on the condition of the patient. Automating this summarization task is critical because the Impression section is the most important part of a radiology report, and manual summarization can be time-consuming and error-prone.

Despite its importance, recent studies (Zhang et al., 2018, 2020; Hu et al., 2022) or challenges (Abacha et al., 2019) on new automated radiology report summarization systems solely focus on chest X-ray or sometimes, omit the modality and anatomy concerned in the used radiology reports (Karn et al., 2022). In addition, while existing models are optimized to generate summaries achieve high performance on the ROUGE metric (Lin, 2004), this does not guarantee factually correct summaries (Zhang et al., 2020).

To palliate these two limitations, we propose a challenge with two brand new pre-processed datasets of new modalities (MR and CT), anatomies (chest, head, neck, sinus, spine, abdomen, pelvis) and institutions (Stanford Hospital). We further use a new metric, called F_1RadGraph (Delbrouck et al., 2022a) to evaluate the factual completeness and correctness of generated radiology impressions.

CT Abd/pelv	CT Chest	CT Head
15,989	12,786	31,402
CT Spine	MR Head	CT Neck
5,517	7,313	1,140
CT Sinus	Mr Spine	MR Abdomen
1,267	2,821	1,061
MR Neck	MR Pelvis	
230	253	

Table 1: Number of reports (findings/impression pairs) for each new modality/anatomy in the MIMIC-III summarization dataset, totaling 79,779 samples.

The challenge took place on ViLMedic (Delbrouck et al., 2022b), a modular framework for vision and language multimodal research in the medical field. This library contains reference implementations of state-of-the-art medical vision and language architectures but also hosts AI challenges.

The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks, pages 478–482
July 13, 2023 ©2023 Association for Computational Linguistics

2 Datasets

Two summarization datasets were proposed for the shared task. The MIMIC-III summarization dataset contains 11 different anatomy-modality pairs (Chen et al., 2023) and the MIMIC-CXR summarization dataset contains findings and impression sections from chest X-ray studies paired with chest X-rays images. In addition to the official MIMIC-CXR test-set, we are also releasing a brand new out-of-institution test-set from the Stanford Hospital.

2.1 MIMIC-III summarization dataset

MIMIC-III (Johnson et al., 2016) is a large, freely-available database comprising de-identified health-related data associated with over forty thousand patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012. This data comprises radiology reports from a wide range of modality (medical imaging techniques) and anatomy (body parts). To create a new radiology report summarization dataset, we first chose the 5 of the most frequent modality/anatomy pairs in the pool of MIMIC-III reports, namely CT head, CT spine, CT chest, CT abdomen-pelvis and MR head. We discard chest X-rays as they are included in the MIMIC-CXR dataset (Johnson et al., 2019). The number of samples per pair is available and constitutes enough data to train a deep learning model. We also picked 5 less represented modality/anatomy pairs that act as out-of-domain (OOD) test-sets, namely MR Spine, CT sinus, MR pelvis, MR abdomen, MR Neck.

For each report, we extract the findings and impression section. However, the finding section is not always stated as such. With the help of one board-certified radiologist, and for each modality/anatomy pair, we create a mapping of the section header that acts as "findings". As an example, for CT head, findings could be referred as "non-contrast head ct", "ct head", "ct head without contrast", "ct head without iv contrast", "head ct", "head ct without iv contrast" or "cta head". This "findings" mapping contains up to 537 candidate sections for our whole dataset. We also discarded reports where multiple studies are pooled in the same radiology reports, leading to multiple intricate observations in the impression section. We release our mapping as well as the code to recreate the dataset from scratch).

The final dataset consists of a train, validation and test splits of respectively 59,320, 7,413 and 13,057 findings-impression pairs. In the scope of this challenge, the test split has been split in two: one public test-set of 6,526 samples and one hidden test-set (participants don't have access to the ground-truth impressions) of 6,531 samples.

2.2 MIMIC-CXR summarization dataset

The MIMIC-CXR (Johnson et al., 2019) is a multimodal summarization dataset that contains chest X-ray findings and impression sections paired with chest X-rays images. It consists of 125,417 training samples, 991 validation samples and 1624 test samples. Exactly 237,564 images are associated to those studies. In the scope of this challenge, we also released one hidden test-set (participants don't have access to the ground-truth impressions) of 1000 samples with images from Stanford Hospital. This test-set has been de-identified using the "hide in plain sight" method (Chambon et al., 2022).

3 Metrics

We proceed to evaluate the submitted systems using the BLEU (Papineni et al., 2002) and ROUGEL metrics (Lin, 2004). A few other metrics were used to score the factual correctness of the generated impressions:

F_1CheXbert (Zhang et al., 2020) This score uses cheXbert (Smit et al., 2020), a Transformer-based model trained to output abnormalities of chest X-rays given a radiology report as input. F_1CheXbert is the f1-score between the prediction of cheXbert over the generated report $\hat{y}$ and the corresponding reference y. The f1 score is calculated over the 5 main observations to be consistent with Zhang et al. (2020). This metric is suitable for the MIMIC-CXR summarization dataset exclusively.

BERTScore (Zhang et al., 2019) An automatic evaluation metric used for testing the goodness of text generation systems. Unlike existing popular methods that compute token level syntactical similarity, BERTScore focuses on computing semantic similarity between tokens of reference and hypothesis.

F_1RadGraph (Delbrouck et al., 2022a) Metric

Dataset	Team	BLEU4	ROUGEL	Bertscore	F1-cheXbert	F1-RadGraph
MIMIC-III hidden test-set (6531 samples)	shs-nlp	18.36	35.32	57.26	N/A	36.94
	utsa-nlp	16.05	34.41	57.08	N/A	36.31
	aimi	16.61	33.43	55.54	N/A	35.12
	sinai	17.38	32.32	55.04	N/A	33.96
	knowlab	13.23	32.02	55.64	N/A	33.39
	nav-nlp	15.13	32.39	55.34	N/A	33.37
	elirf	18.06	30.19	53.94	N/A	32.58
MIMIC-CXR hidden test-set (Stanford Hospital, 1000 samples)	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
MIMIC-III test-set 6526 samples	utsa-nlp	15.99	34.07	56.30	N/A	35.25
	shs-nlp	17.33	33.93	55.49	N/A	34.93
	nav-nlp	15.31	32.33	54.49	N/A	32.68
	sinai	17.12	31.62	54.33	N/A	32.65
	knowlab	13.86	32.22	54.91	N/A	32.49
	elirf	17.41	29.57	52.24	N/A	31.40
	aimi	1.25	24.45	45.54	N/A	21.24
MIMIC-CXR test-set 1624 samples	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 2: Leaderboard results for MIMIC-III and MIMIC-CXR datasets.

leveraging RadGraph (Jain et al., 2021) annotation scheme and model to design F-score style score that measures the consistency and completeness of generated radiology reports compared to reference reports based on observation and anatomy entities.

4 Results

Table 2 provides an overview of the performance of all teams across the four test-sets, with rankings based on the F_1RadGraph metric. We congratulate shs-nlp, utsa-nlp, and aimi for their outstanding performance on the MIMI-III hidden test-set, as well as ku-dmis-msra, utsa-nlp, and knowlab for their impressive results on the Stanford Hospital hidden test-set.

It's worth noting that the F_1RadGraph metric doesn't necessarily correspond with the rankings based on F_1CheXbert on the MIMIC-CXR dataset. This indicates that each metric offers a unique perspective on the factual correctness of the generated summaries.

5 System descriptions

KnowLab THis solution is a comparison of 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. In particular, authors first compared the fine-tuning results using MIMIC-III with SOTA generative pre-trained language models including BART and T5 models as well as their biomedical variants BioBART and SciFive. Then they did a second round of fine-tuning with MIMIC-CXR using the BART and T5 models fine-tuned from MIMIC-III, by freezing the last two layers during training. They 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.

SINAI The system proposed by the SINAI team follows a "sequence-to-sequence" approach, utilizing pre-trained language models that are tailored for both general and biomedical domains. Through fine-tuning, a significant improvement in performance was achieved, with the domain-specific model reaching a F1RadGraph score of 33.96 - ranking fourth among all challenge participants. The team also attempted to utilize Proximal Policy Optimization Reinforcement Learning to further improve factual correctness, but unfortunately, this did not yield satisfactory results.

nav-nlp They took part in Task 1B: Radiology Report Summarization. Multiple runs were submitted for evaluation from solutions utilizing transfer learning from pre-trained transformer models, which were then fine-tuned on MIMIC-III dataset, for abstractive report summarization. The task was evaluated using different evaluation metrics such as ROUGEL, Bertscore, F1-RadGraph and the corresponding scores of our best performing system are 32.33, 54.49, 32.68 respectively.

UTSA-NLP The system for the MIMIC-CXR task uses a two-stage approach to generate an impression from chest X-ray images and text reports. The first stage involves a multimodal image-text retrieval model that retrieves the most similar radiology reports from a medical corpus based on joint embeddings. The second stage uses a text-only model trained on modified inputs from the first stage to generate the final impression. An ensemble model is used for robustness. Authors fine-tune our pre-trained MIMIC-CXR model on the MIMIC-III corpus using the text-only model and synthetic data augmentation. Multiple models are trained and ensemble for both tasks.

shs-nlp Authors propose RadBloomz, an extension of the domain adaptation paradigm beyond the typical method of *pretrain-and-finetune* or instruction-tuned LLM for domain-specific tasks. They refer to our approach as *general-pretrain-prompt-tune-and-special-pretrain*. With this approach, the model is trained using the same initial LM objective in each of the three training stages (i.e., general pretraining, prompt-tuning and domain specialized pretraining), which is a significant advantage. They continued the pretraining of the instruction-tuned Bloomz 7 billion parameter model on large-scale medical text data from MIMIC IV notes radiology reports dataset to form RadBloomz, and evaluated this adaptation paradigm on the radiology report summarization task. The proposed system in a zero-shot setting exhibits better performance than *pretrain-and-finetune* methods on fact-based scoring metrics for impression generation.

KU-DMIS-MSRA In this paper, authors introduce CheXOFA, a new pre-trained vision-language model (VLM) for the chest X-ray domain. Our model is initially pre-trained on various multimodal datasets within the general domain before being transferred to the chest X-ray domain. Following a prominent VLM, OFA, they unify various domain-specific tasks into a simple sequence-to-sequence schema. It enables the model to effectively learn the required knowledge and skills from limited resources in the domain.

ELiRF The authors pre-trained a general domain BART model with a focus on two aspects. Firstly, adapting the model to the biomedical domain using data from MIMIC dataset, and secondly, using the News Abstractive Summarization pre-training methodology to increase the abstractivity of the summaries generated by injecting linguistic knowledge. They then fine-tuned the resulting pre-trained models with various amounts of data from the shared task to create multiple models.

6 Conclusion

We proposed a Radiology Report Summarization (RadSum23) challenge with two brand new pre-processed datasets of new modalities (MR and CT), anatomies (chest, head, neck, sinus, spine, abdomen, pelvis) and institutions (Stanford Hospital). We further use a new metric, called F_1RadGraph (Delbrouck et al., 2022a) to evaluate the factual completeness and correctness of generated radiology impressions. Our competition gathered 112 submissions across 11 teams. The final leaderboard can be accessed at the following address: https://vilmedic.app/misc/bionlp23/leaderboard/[1]. The complete data of this competition (datasets and submissions from participants) can be downloaded at https://vilmedic.app/misc/bionlp23/sharedtask/.

[1] In case this link doesn't work anymore, please visit https://github.com/jbdel/vilmedic

References

Asma Ben Abacha, Chaitanya Shivade, and Dina Demner-Fushman. 2019. Overview of the mediqa 2019 shared task on textual inference, question entailment and question answering. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 370–379.

Pierre J Chambon, Christopher Wu, Jackson M Steinkamp, Jason Adleberg, Tessa S Cook, and Curtis P Langlotz. 2022. Automated deidentification of radiology reports combining transformer and "hide in plain sight" rule-based methods. *Journal of the American Medical Informatics Association*. Ocac219.

Zhihong Chen, Maya Varma, Xiang Wan, Curtis Langlotz, and Jean-Benoit Delbrouck. 2023. Toward expanding the scope of radiology report summarization to multiple anatomies and modalities. In *Proceedings of the 61th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, Toronto, Canada. Association for Computational Linguistics.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. 2022a. Improving the factual correctness of radiology report generation with semantic rewards. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4348–4360, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Jinpeng Hu, Zhuo Li, Zhihong Chen, Zhen Li, Xiang Wan, and Tsung-Hui Chang. 2022. Graph enhanced contrastive learning for radiology findings summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4677–4688.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. 2021. Radgraph: Extracting clinical entities and relations from radiology reports. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

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):1–8.

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.

Charles E Kahn Jr, Curtis P Langlotz, Elizabeth S Burnside, John A Carrino, David S Channin, David M Hovsepian, and Daniel L Rubin. 2009. Toward best practices in radiology reporting. *Radiology*, 252(3):852–856.

Sanjeev Kumar Karn, Ning Liu, Hinrich Schütze, and Oladimeji Farri. 2022. Differentiable multi-agent actor-critic for multi-step radiology report summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1542–1553.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew Lungren. 2020. Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1500–1519.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yuhao Zhang, Daisy Yi Ding, Tianpei Qian, Christopher D Manning, and Curtis P Langlotz. 2018. Learning to summarize radiology findings. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 204–213.

Yuhao Zhang, Derek Merck, Emily Tsai, Christopher D Manning, and Curtis Langlotz. 2020. Optimizing the factual correctness of a summary: A study of summarizing radiology reports. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5108–5120.