Automatic Radiology Report Generation: Evaluating the Alignment of NLP Metrics with Radiologist Assessment

Andrea Santomauro[1,2,*,†], Ivan Gallesio[3], Xhorxhi Kaci[4], Giorgio Leonardi[1,2] and Luigi Portinale[1,2]

[1]Computer Science Institute, DiSIT, Università del Piemonte Orientale, Alessandria (Italy)
[2]Integrated Lab. of AI and Medical Informatics, DAIRI , AOUAL, Alessandria (Italy)
[3]SOC Radiologia, AOUAL, Alessandria, (Italy)
[4]Department of Surgical Sciences, Università di Torino, Torino (Italy)

Abstract

Automatic radiology report generation (ARRG) represents a critical advancement in healthcare, addressing challenges such as professional shortages and diagnostic errors. Evaluating the quality of narrative reports still remains a significant hurdle and challenge. Traditional NLP metrics primarily rely either on surface-level n-gram overlaps or synonym matching, which is often not enough for the domain-specific demands of radiology. This paper evaluates the alignment of traditional NLP metrics with expert radiologist assessments in the context of ARGG. Narrative reports were generated using a multimodal decoder-only transformer architecture, and their quality was evaluated using BERTScore, BLEU-4, and ROUGE. Independent board-certified radiologists assessed the reports based on clinically significant error categories, providing a benchmark for comparison. Despite the general positive assessment by experts, results reveal a limited correlation between automatic scores and radiologist judgments, highlighting the inadequacy of conventional metrics in capturing critical clinical aspects. This highlights the need for domain-specific evaluation frameworks that align with clinical standards to ensure the reliability of ARRG systems.

Keywords

Automated Radiology Report Generation, NLP metrics, Clinical Evaluation

1. Introduction

Automatic Radiology Report Generation (ARRG) has recently gained a lot of attention in the healthcare community [1] triggered by several factors such as the restricted number of professionals [2], the risk of important errors in the common practice [3] and a shortage in the skills needed to perform precise and correct diagnostic reports, as well as the adoption of policies aimed at minimizing the risk for malpractice liability [4].

Since the use of structured reports is currently limited to some specific diagnostic investigations such as mammography and bone densitometry (DXA) [5], a challenging task in this context is the construction of narrative reports [6, 7, 8], involving the generation of a structure-free set of sentences, providing the descriptions of the findings captured by the radiologist, as well as his/her diagnostic impressions. Our claim is that evaluating the performance of narrative reports using traditional NLP metrics, may not be enough when specific criteria concerning the findings present in a report must be met. These metrics primarily rely on n-gram overlaps or word embedding, while radiology reports usually demand a deep semantic understanding and domain-specific reasoning, as they often involve complex relationships between findings, impressions, and recommendations. Furthermore, variations in phrasing can carry significant diagnostic implications (as an example, “no evidence of malignancy” is not equivalent to “malignancy is absent,” despite sharing similar n-grams or synonyms). Additionally,

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy
*Corresponding author.
†Research carried out during the National PhD on AI in Healthcare and Life Sciences, Università Campus Biomedico, Roma (Italy)
andrea.santomauro@uniupo.it (A. Santomauro); igallesio@ospedale.al.it (I. Gallesio); xkaci@aslcn2.it (X. Kaci); giorgio.leonardi@uniupo.it (G. Leonardi); luigi.portinale@uniupo.it (L. Portinale)

errors in radiology reports can result in severe clinical consequences, yet conventional NLP metrics fail to account for this, treating all mismatches uniformly without distinguishing critical errors from minor discrepancies. The primary objective of this work is to assess the degree of alignment between

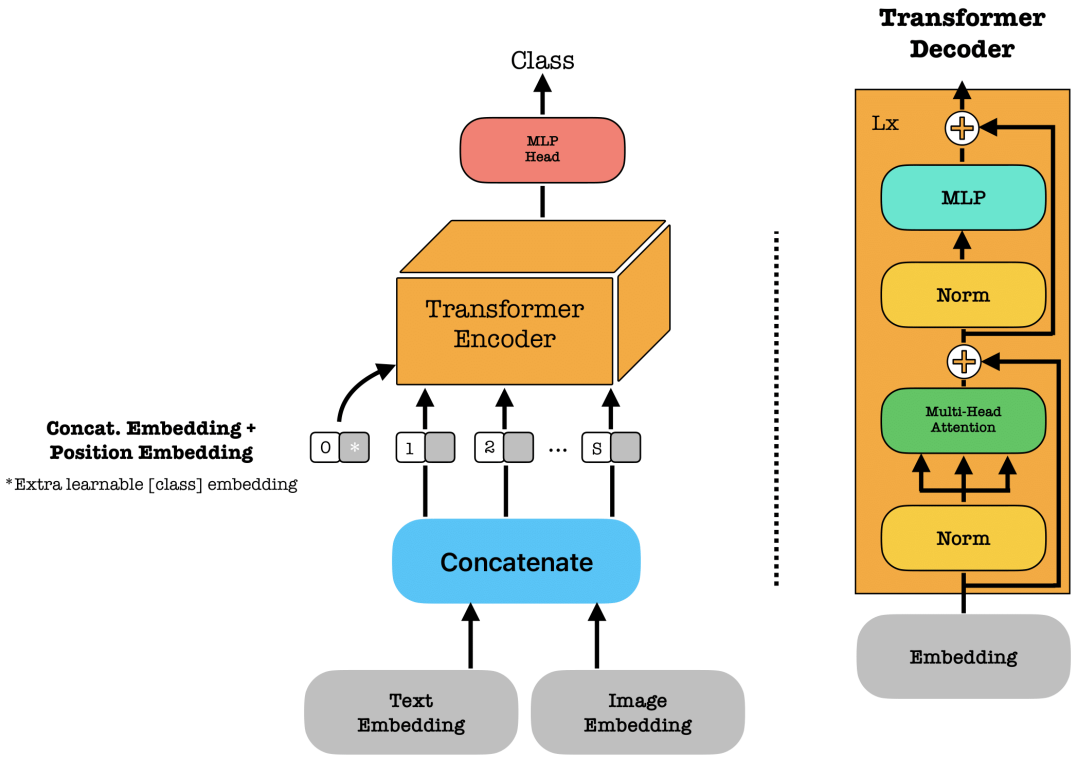

Figure 1: Multimodal decoder-only transformer

the aforementioned metrics and expert radiologists' assessments. To achieve this, we: (1) generate narrative reports using the multimodal decoder-only transformer architecture depicted in Figure 1; (2) compute the BERTScore [9], BLEU-4, and ROUGE-L [10] metrics for the generated reports; (3) assess the alignment of these metrics with expert radiologists' evaluations.

The multimodal decoder-only Transformer architecture processes concatenated input embeddings, where the first K tokens correspond to the image embeddings and the subsequent $N - K$ tokens correspond to the text embeddings, where N is the total number of tokens. After concatenation, a positional encoding is applied to the combined sequence, which is then fed into a GPT-2-like decoder-only Transformer, with the model trained to autoregressively predict only the text tokens as output.

2. Evaluation Methodology

Two of the authors of the present paper are independent board-certified radiologists, and have been asked to assess the quality of the generated reports; they have been provided with both the generated and the corresponding ground-truth reports, along with the associated X-ray images. Specifically, they were asked to count the errors found in the generated report across the following categories: (1) *False prediction of finding*; (2) *Omission of finding*; (3) *Incorrect location of finding*; (4) *Incorrect severity of finding*. This evaluation method is quite standard in similar contexts as described in [11]. They have been finally provided with 50 generated reports and their corresponding images and ground truth reports. Table 1 shows the distribution of the number of errors (n) found by the expert radiologists

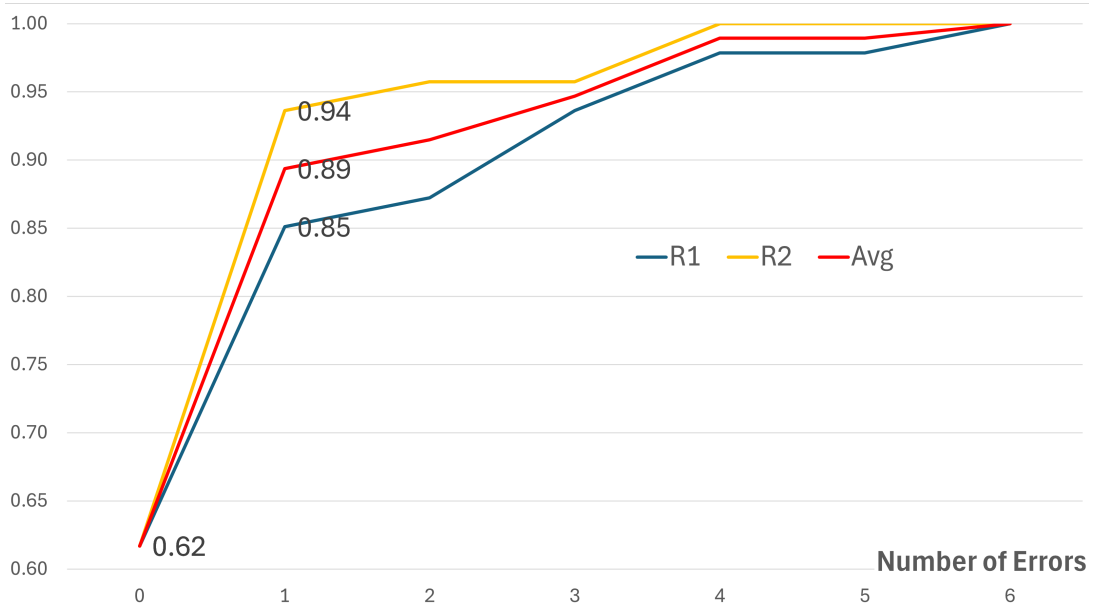

Figure 2: Cumulative distribution (`cdf`) of errors reported by radiologists

(indicated as R1 and R2), where the counts spans over the four categories of errors described above[1]. From this error distribution we can estimate the probability mass function (`pmf`) as shown in Table 2.

	n=0	n=1	n=2	n=3	n=4	n=5	n=6	$WAVG$
R1	29	11	1	3	2	0	1	**1.71**
R2	29	15	1	0	2	0	0	**1.19**
AVG	29	13	1	1.5	2	0	0.5	**1.45**

Table 1
Distribution of Errors Determined by Radiologists (AVG is the average between the two experts and WAVG the average weighted by the number of errors)

The last column of Table 2 reports the expected numbers of errors $E[n]$ given by the estimated `pmf`. We can notice that less than one error for report is found, both by each radiologist and on average.

	n=0	n=1	n=2	n=3	n=4	n=5	n=6	$E[n]$
R1	0.62	0.23	0.02	0.06	0.04	0	0.02	**0.77**
R2	0.62	0.32	0.02	0.06	0	0.04	0	**0.53**
AVG	0.62	0.22	0.02	0.03	0.04	0	0.01	**0.65**

Table 2
Distribution of errors (estimated `pmf`)

Figure 2 shows the cumulative distribution (`cdf`) of the errors. In particular the `cdf` for $n = 0$ is the probability of getting no error in the report and for $n = 1$ is the probability of getting *at most* one error. We can conclude that radiologists have considered the overall quality of the generated reports to be really satisfactory.

Given that, we have considered to measure the alignment of the expert evaluations with respect to the automatic scores. To this end, we have produced the scatter plots in Figure 3, Figure 4 and Figure 5, plotting the number of errors (averaged between the two radiologists) versus F_{BERT} (F-score of BERTScore), BLEU-4 and $F_{ROUGE-L}$ (F-score of ROUGE-L) respectively. As expected, regression lines show a clear negative linear correlation between the score and the number of reported errors.

[1]We refers to a total number of 47 reports, since 3 cases have not been considered as evaluable by the physicians.

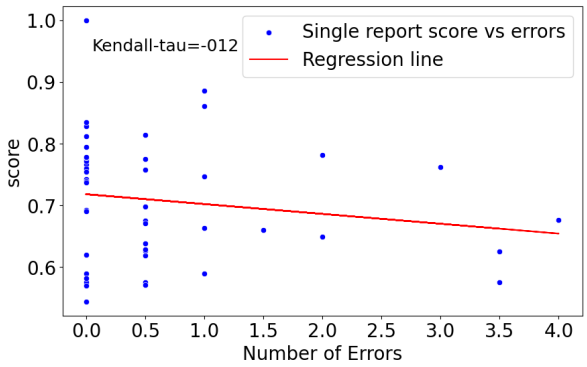

Figure 3: Expert alignment with F_{BERT}

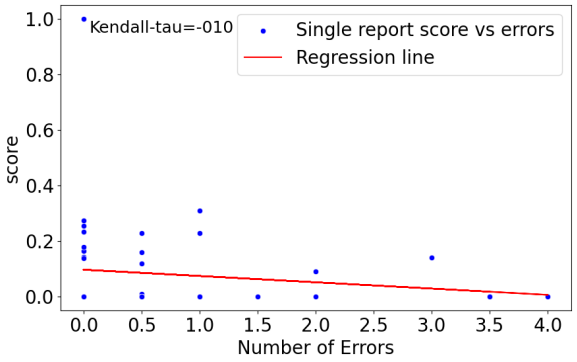

Figure 4: Expert alignment with BLEU-4

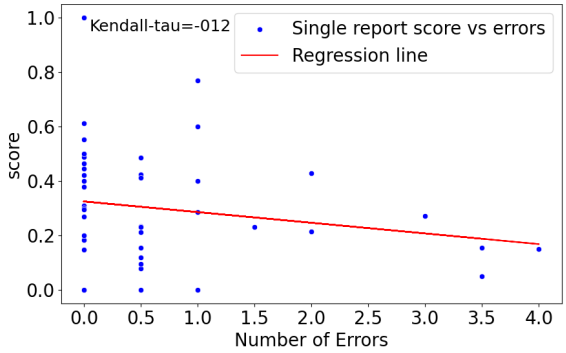

Figure 5: Expert alignment with $F_{ROUGE-L}$

However, by looking at the Kendall τ coefficient (a standard measure of the correlation strenght in this situation [11]), we can notice that the degree of correlation is not very high, suggesting that, in this experiment, automatic scoring does not capture in a strong way the expert judgement. This is quite evident by looking at the very different scores that reports with few errors get, suggesting that reporting the correct findings may be done in a quite different number of ways that automatic scores do not easily capture.

3. Conclusions

In conclusion, the proposed evaluation shows that:

- the generative architecture in Figure 1 is able to produce reports rated as good quality by experts
- automatic NLP scores may have trouble in capturing good reports with a few errors
- NLP scores do not align very well with expert judgments if we consider the error categories described in the paper.

This highlights the need for domain-specific evaluation frameworks that align with clinical standards to ensure the reliability of ARRG systems.

Declaration of Generative AI usage

The authors declare that no generative AI tools were used in the preparation of this manuscript.

References

[1] Y. Liao, H. Liu, I. Spasic, Deep learning approaches to automatic radiology report generation: A systematic review, Informatics in Medicine Unlocked 39 (????). doi:https://doi.org/10.1016/j.imu/2023.101273.

[2] Royal College of Radiologists, Clinical radiology census report, http://tinyurl.com/ukcrr2021, 2021.

[3] L. Berlin, Defending the "missed" radiographic diagnosis, American Journal of Roentgenology 176 (2001) 317–322.

[4] A. Babu, M. Brooks, The malpractice liability of radiology reports: minimizing the risk., Radiographics 35 (2015) 547–554.

[5] J. Nobel, K. van Geel, S. Robben, Structured reporting in radiology: a systematic review to explore its potential, European Radiology 32 (2022) 2837–2854.

[6] X. Zeng, L. Wen, Y. Xu, C. Ji, Generating diagnostic report for medical image by high-middle-level visual information incorporation on double deep learning models, Comput Methods and Programs in Biomedicine 197 (2020).

[7] Z. Chen, Y.Song, T.-H. Chang, X. Wan, Generating radiology reports via memorydriven transformer, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020), 2020.

[8] F. Nooralahzadeh, N. Gonzalez, T. Frauenfelder, K. Fujimoto, M. Krauthammer, Progressive transformer-based generation of radiology reports, arXiv:2102.09777, 2021.

[9] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with bert, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[10] S. Lee, J. Lee, H. Moon, C. Park, J. Seo, S. Eo, S. Koo, . Lim, A survey on evaluation metrics for machine translation, Mathematics 11 (2023). URL: https://www.mdpi.com/2227-7390/11/4/1006.

[11] F. Yu, M. Endo, R. Krishnan, I. Pan, A. Tsai, E. P. Reis, E. K. U. N. Fonseca, H. M. H. Lee, Z. S. H. Abad, A. Y. Ng, C. P. Langlotz, V. K. Venugopal, P. Rajpurkar, Evaluating progress in automatic chest x-ray radiology report generation, Patterns (N. Y.) 4 (2023) 100802.