

Semantic Textual Similarity Assessment in Chest X-ray Reports Using a Domain-Specific Cosine-Based Metric

Sayeh Gholipour Picha^a, Dawood Al Chanti^b and Alice Caplier^c

Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

Keywords: Semantic Similarity, Medical Language Processing, Biomedical Metric.

Abstract: Medical language processing and deep learning techniques have emerged as critical tools for improving healthcare, particularly in the analysis of medical imaging and medical text data. These multimodal data fusion techniques help to improve the interpretation of medical imaging and lead to increased diagnostic accuracy, informed clinical decisions, and improved patient outcomes. The success of these models relies on the ability to extract and consolidate semantic information from clinical text. This paper addresses the need for more robust methods to evaluate the semantic content of medical reports. Conventional natural language processing approaches and metrics are initially designed for considering the semantic context in the natural language domain and machine translation, often failing to capture the complex semantic meanings inherent in medical content. In this study, we introduce a novel approach designed specifically for assessing the semantic similarity between generated medical reports and the ground truth. Our approach is validated, demonstrating its efficiency in assessing domain-specific semantic similarity within medical contexts. By applying our metric to state-of-the-art Chest X-ray report generation models, we obtain results that not only align with conventional metrics but also provide more contextually meaningful scores in the considered medical domain.


1 INTRODUCTION


Advancements in deep learning for medical language processing have significantly improved healthcare clinical analysis, particularly in the domain of medical imaging applications. Notably, there has been substantial progress in generating chest X-ray reports comparable to those written by radiologists. However, a critical challenge persists in the chest X-ray application—assessing the semantic similarity between generated reports and the ground truth.


Identifying semantic similarities in medical texts is a difficult task within the language processing domain (Alam et al., 2020). This task necessitates a comprehensive grasp of the entire medical text corpus, the ability to recognize key content, and a profound understanding of the semantic relationships between these critical keywords at an expert level. While existing metrics and approaches for capturing semantic similarity in natural language are effective, they are not designed for the complexities of medical

content. The need for a robust metric to assess semantic similarity in medical texts has become increasingly evident, particularly in applications like chest X-ray report generation, and continues to be an active area of research (Endo et al., 2021), (Miura et al., 2021), (Yu et al., 2022).

State-of-the-art chest X-ray report generation models (Chen et al., 2020), (Miura et al., 2021), (Endo et al., 2021) still rely on conventional Natural Language Processing (NLP) methods like BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), and ROUGE (Lin, 2004) to evaluate the generated reports against ground truth references. However, these metrics produce unreliable results due to their inability to comprehend and compare the semantic similarity of key medical terms. A medical semantic similarity metric would not only provide more significant evaluation scores but could also be incorporated into the training process to improve model performance, potentially leading to enhanced diagnostic accuracy and decision-making. Additionally, as part of our ongoing research, our goal is to focus on providing visual interpretations of chest X-ray reports using text-to-image localization. As a consequence, a robust semantic similarity evaluation metric suitable

^a  <https://orcid.org/0000-0003-2675-5463>

^b  <https://orcid.org/0000-0002-6258-6970>

^c  <https://orcid.org/0000-0002-5937-4627>

for medical content will ensure the reliability of generated reports and will enable us to achieve more accurate localization and interpretation of image content.

In this context, we propose a new metric designed to assess and assign scores about the semantic similarity of medical texts. Our metric consists of two sequential steps: first, we identify the primary clinical entities, and subsequently, we evaluate the similarity between these entities using the domain-specific Cosine similarity score. Notably, our approach considers the presence of negations and detailed descriptions associated with medical entities during the evaluation process. To this end, our contributions include:

- Introduction of a novel system for clinical entity extraction from medical texts.
- Proposition of a new scoring system for the evaluation of semantic similarity that suits medical and natural texts.
- Presentation of a validation method for scoring verification.

This paper is structured as follows: Section 2 discusses related works; Section 3 presents the theoretical and mathematical part of the novel metric; Section 4 validates the metric; Section 5 discusses the results; Finally, Section 6 concludes the paper.

2 RELATED WORKS

Recent studies have addressed the challenge of similarity evaluation between generated medical reports and the ground truth through various approaches other than conventional NLP metrics. Researchers have often introduced innovative metrics in the process.

In the CXR-RePaiR model by Endo et al. (Endo et al., 2021) a unique approach for automatically evaluating chest X-ray report generation is proposed by introducing the CheXbert vector similarity metric, using the CheXbert labeler (Smit et al., 2020) — a specialized tool for chest X-ray report labeling. The process involves extracting labels from generated reports, comparing them with ground truth labels, and presenting the final score using cosine similarity. While this approach outperforms the BLEU metric, its applicability is limited to the specific context of chest X-ray reports and does not readily extend to other medical applications. The limitations arise from Chexbert being exclusively trained for chest X-ray reports. Moreover, the Chexpert labels (Irvin et al., 2019) (Atelectasis, Cardiomegaly, Consolidation, Edema, Enlarged Cardiomeastinum, Fracture,

Lung Lesion, Lung Opacity, No Finding, Pleural Effusion, Pleural Other, Pneumonia, Pneumothorax) are specific to the chest X-ray dataset, further limiting the generalizability of the approach to other medical contexts.

In a separate study, Yu et al. (Yu et al., 2022) introduced a novel metric targeting the quantification of overlap of clinical entities between ground truth and generated reports in chest X-ray report generation. They use the RadGraph model (Jain et al., 2021), a language model trained on a limited subset of reports from the MIMIC-CXR dataset (Johnson et al., 2019). The MIMIC-CXR dataset consists of chest X-ray images with corresponding reports, and the RadGraph dataset includes medical entities from chest X-ray reports annotated by radiologists. The approach by Yu et al. is similar to the BLEU score, exclusively considering the exact matches among the primary entities in generated and ground truth reports, overlooking the semantic similarity of these entities. Furthermore, the generalizability of this approach to other medical applications is constrained by the RadGraph model's specialization in extracting only chest X-ray related entities. Nonetheless, while the RadGraph model acknowledges negations in the texts, they are treated merely as labels to the entities, and the details of entity descriptions are not factored into the evaluation process.

In a recent study, Patricoski et al. (Patricoski et al., 2022) conducted an evaluation of seven BERT models to assess semantic similarity in clinical trial texts. Notably, the pre-trained BERT model known as SciBERT (Beltagy et al., 2019) demonstrated better performance compared to the other BERT models, even outperforming the standard BERT model, which secured the second position in this evaluation. This study underlines the promising potential of BERT models in semantic similarity evaluation. However, it has a drawback associated with using BERT models without preprocessing. BERT models operate at a token-by-token level, evaluating semantic similarity by comparing all tokens with each other, a computationally intensive process that gives relatively low scores. Despite this computational challenge, it is important to consider the significant potential in SciBERT, particularly due to its huge clinical dictionary. This finding underscores the need for careful consideration of preprocessing strategies to maximize the effectiveness of BERT models in semantic similarity evaluations.

Notably, the absence of a comprehensive, general semantic similarity evaluation metric for medical content persists. Consequently, we introduce a novel metric for Medical Corpus Similarity Evaluation (MCSE)

to comprehensively address and resolve these challenges.

3 METHODOLOGY

We developed a novel metric for Medical Corpus Similarity Evaluation (MCSE), by exclusively extracting key medical entities and employing a pre-trained BERT model to assess the semantic similarity of these entities within chest X-ray reports. This targeted approach allows BERT to concentrate solely on important information and reduces the computational load during comparison. Importantly, our methodology goes beyond extracting main entities, we also consider the negations and detailed descriptions associated with the primary medical entities in chest X-ray reports. Our MCSE metric consists of two essential steps:

1. Clinical Entity Extraction.
2. Domain Similarity Evaluation.

3.1 Clinical Entity Extraction

The most important part of comprehending semantic similarity evaluation in text relies on identifying the key elements, often referred to as clinical entities, within medical texts. These entities typically fall into categories related to anatomical body parts, symptoms, laboratory equipment, and diagnoses. Each category is typically signaled by certain words within a sentence. However, there are additional words that precede or follow these main entities, offering descriptions.

To address these complexities, we employ the Scispacy model (Neumann et al., 2019) for extracting primary clinical entities from medical text using the embedded clinical dictionary in this model (BC5CDR: a corpus comprising 1500 PubMed articles with 4409 annotated chemicals, 5818 diseases, and 3116 chemical-disease interactions (Li et al., 2016)). Subsequently, we automatically process the entire text to identify associated negations and adjectives related to these key entities. These elements are then integrated to provide a comprehensive representation of the considered text. In the context of this research, the category of laboratory equipment is deliberately excluded, aligning with the specific focus of our application. Table 1 presents an example of medical text and the extracted entities using our method and the Scispacy method without any cleaning process. While we employ the Scispacy model for initial entity extraction, it is evident that this model

alone may not suffice. An additional automated post-processing step is needed to refine and integrate related entities. The post-processing steps involve eliminating a single adjective or non-medical entities, excluding entities categorized as lab equipment, identifying and adding the relevant adjective to the remaining medical entities, including the existing negation into these primary entities, and screening out any reported diagnostic procedures terms. These processes are essential to ensure that the final output is presented as a cohesive set of primary medical entities, ready for practical use.

3.2 Domain Similarity Evaluation

Having successfully extracted and shifted our focus to the primary entities within the medical corpus, the next step involves assessing their semantic similarities by assigning corresponding scores.

After processing entity extraction, we calculate a similarity score for the sequences of entities. Let $T = (t_1, \dots, t_N)$ represent the reference text entities and $\hat{T} = (\hat{t}_1, \dots, \hat{t}_M)$ represent the generated text or candidate text entities. Initially, we identify the exact same medical entities in both sequences and determine the total count ($|C^{(i)}|$). For the remaining entities, we construct a similarity matrix, where each element represents the similarity score between entities, as illustrated in table 2.

$$S_i = \frac{\max y_{i,j}}{\max y_{i,j} + \bar{y}_{i,j}} \quad i = (0, 1, \dots, M) \quad j = (0, 1, \dots, N) \quad (1)$$

$$y_{i,j} = \text{Similarity}(r_i, \hat{r}_j) \quad (2)$$

$$\begin{cases} C^{(i)} = t_i, & \text{if } t_i = \hat{t}_j \\ r_i = t_i \ \& \ \hat{r}_j = \hat{t}_j & \text{if } t_i \neq \hat{t}_j \end{cases} \quad (3)$$

Where M is the number of total candidate entities, r_i and \hat{r}_j are the sequence between no matched entities as in equation (3), and S_i is a normalized similarity score between r_i and \hat{r}_j . The similarity score $\text{Similarity}(r_i, \hat{r}_j)$ in equation (2) is derived from spaCy (Honnibal et al., 2020), a BERT model trained on word2vec, to evaluate similarities using domain cosine similarity.

To evaluate the similarity of candidate entities with the reference entities, we compute the maximum score for each column and normalized it with the column's average (S_j). We then sum these scores for each column, adding them to $|C^{(i)}|$. To obtain the final similarity score between the two corpora, we divide this sum by the total number of candidate entities. This process is explained in Equation (4).

Table 1: In the right column there is an example of medical text. In the left column, there are clinical entities extracted using the Scispacy model without any cleaning process, and In the middle column, there are clinical entities extracted using our method.

Medical Text	Extracted Entities using our method	Extracted Entities using Scispacy (Neumann et al., 2019)
1. Interval clearance of left basilar consolidation. 2. Patchy right basilar opacities, which could be seen with minor atelectasis, but given the context clinical correlation is suggested regarding any possibility for recurrent or new aspiration pneumonitis at the right lung base. 3. Increased new interstitial abnormality, suggesting recurrence of fluid overload or mild-to-moderate pulmonary edema; aspiration could also be considered. Inflammation associated with atypical infectious process is probably less likely given the waxing and waning presentation.	fluid overload, inflammation, aspiration pneumonitis, minor atelectasis, mild to moderate pulmonary edema, left basilar consolidation, patchy right basilar opacities, interstitial abnormality	Interval, clearance, left basilar, consolidation, Patchy, right basilar, opacities, minor, atelectasis, clinical, recurrent, aspiration, pneumonitis, right lung base, Increased, interstitial abnormality, recurrence, fluid, overload, mild-to-moderate pulmonary edema, aspiration, Inflammation, associated with, atypical, infectious process, waxing, waning, presentation

$$MCSE := \frac{|C^{(i)}| + \sum_{i=1}^M S_i}{M} \quad (4)$$

Where $|C^{(i)}|$ is the number of exactly matched entities between the two corpora of T and \hat{T} .

For instance, Table 2 provides an example of the probable similarity score that two sets of entities can receive. These entities have been extracted using our medical entity extraction procedure.

In the table, the two corpora received a score of 0.55 according to our MCSE metric. However, the calculated BLEU score for them is approximately zero. Upon analyzing the two medical texts, it becomes evident that although the candidate text does refer to the same side of the chest as in the reference text and that both texts indicate the presence of pulmonary edema and pulmonary masses, their overall similarity is relatively limited. The score of 0.55 carries a more meaningful value in this context compared to the nearly zero score generated by BLEU.

4 VALIDATION

While the underlying logic of this metric is reasonable, it is imperative that we validate the results robustly. Given the use of chest X-ray reports for this particular application, we have conducted an extensive search within existing datasets to identify an appropriate validation method. After a comprehensive review of various datasets, we concluded that it would

be more effective to conduct separate validations for the different steps of the proposed metric.

4.1 Clinical Entity Extraction Process

In order to rigorously validate our clinical entity extraction process, we employ the RadGraph dataset (Jain et al., 2021). This dataset is a valuable resource in which radiologists thoroughly annotated the primary clinical entities in chest X-ray reports as either "definitely present" within the report or "definitely absent". Importantly, in cases where a negation is associated with a particular entity, it is annotated as "definitely absent."

To achieve our validation objectives, we executed our entity extraction process on the reports within this dataset. Subsequently, we compare the number of similar entities extracted through our method with the annotations provided by radiologists, particularly focusing on the two categories of "definitely present" and "definitely absent". This systematic comparison allows us to assess the accuracy and effectiveness of our clinical entity extraction methodology in the context of chest X-ray reports, aligning with radiological standards. Throughout the validation process, covering all reports in our study, our method consistently achieves a high level of accuracy. On average, it accurately recognizes 75% of entities marked as "definitely present" and successfully identifies 76% of entities labeled as "definitely absent". In our entity extraction process, we deliberately omit anatomical entities like "chest" or "lung," as they are redundant to

Table 3: A sample table featuring Chexpert labels (1. Atelectasis, 2. Cardiomegaly, 3. Consolidation, 4. Edema, 5. Enlarged Cardiomeastinum, 6. Fracture, 7. Lung Lescion, 8. Lung Opacity, 9. No Finding, 10. Pleural Effusion, 11. Pleural Other, 12. Pneumonia, 13. Pneumothorax, 14. Support Devices) extracted from chest X-ray reports of five patients (Subject ##) from the MIMIC-CXR database (Johnson et al., 2019).

Subject ##	Atelectasis	Cardiomegaly	Consolidation	Edema	Enlarged Cardiomeastinum	Fracture	Lung Lescion	Lung Opacity	No Finding	Pleural Effusion	Pleural Other	Pneumonia	Pneumothorax	Support Devices
01								0		1		1	-1	
02							1					1		
03									1			0		
04		1	0					-1					0	1
05	1									1				

Table 4: Reports corresponding to the subjects listed in Table 3 from the MIMIC-CXR dataset (Johnson et al., 2019).

Subject.##	Report
01	Lung volumes remain low. There are innumerable bilateral scattered small pulmonary nodules which are better demonstrated on recent CT. Mild pulmonary vascular congestion is stable. The cardio mediastinal silhouette and hilar contours are unchanged. Small pleural effusion in the right middle fissure is new. There is no new focal opacity to suggest pneumonia. There is no pneumothorax.
02	A triangular opacity in the right lung apex is new from prior examination. There is also fullness of the right hilum which is new. The remainder of the lungs are clear. Blunting of bilateral costophrenic angles, right greater than left, may be secondary to small effusions. The heart size is top normal.
03	Mild to moderate enlargement of the cardiac silhouette is unchanged. The aorta is calcified and diffusely tortuous. The mediastinal and hilar contours are otherwise similar in appearance. There is minimal upper zone vascular redistribution without overt pulmonary edema. No focal consolidation, pleural effusion or pneumothorax is present. The osseous structures are diffusely demineralized.
04	The endotracheal tube tip is 6 cm above the carina. Nasogastric tube tip is beyond the GE junction and off the edge of the film. A left central line is present in the tip is in the mid SVC. A pacemaker is noted on the right in the lead projects over the right ventricle. There is probable scarring in both lung apices. There are no new areas of consolidation. There is upper zone redistribution and cardiomegaly suggesting pulmonary venous hypertension. There is no pneumothorax.
05	A moderate left pleural effusion is new. Associated left basilar opacity likely reflect compressive atelectasis. There is no pneumothorax. There are no new abnormal cardiac or mediastinal contour. Median sternotomy wires and mediastinal clips are in expected positions.

contrasting labels. The red horizontal line within the figure serves as the dividing line distinguishing between similar and opposite evaluations. Upon reviewing these results, it becomes evident that a distinct boundary exists between reports sharing the same clinical diagnoses and those with entirely dissimilar diagnoses. Notably, there are no blue dots below a 70% similarity threshold, whereas six orange dots have scores above 70% across 70 label sequences, which is certainly not very high. Nevertheless, de-

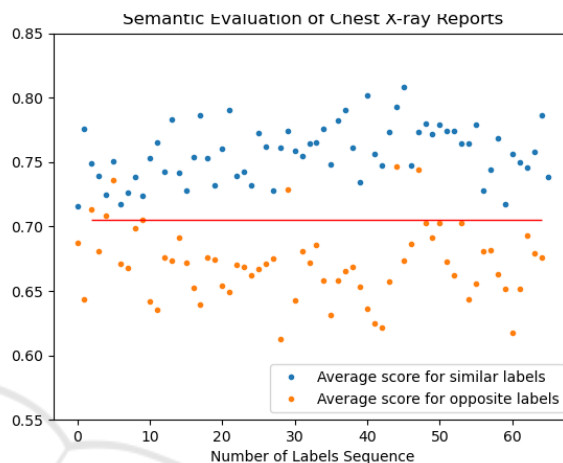


Figure 1: Semantic Evaluation of Chest X-ray reports. Each blue dot represents the mean score of semantic evaluation for reports with similar label sequences, while each orange dot signifies the mean score of semantic evaluation for reports with opposing labels. The red horizontal line represents the classification boundary.

spite this differentiation between similar and opposite evaluations, some level of similarity, exceeding 50%, persists within the opposing category. This can be attributed to the implemented cosine similarity within the medical domain, which introduces a certain bias towards tokens in the same medical domain. Unfortunately, this bias cannot be entirely eliminated, as it plays a substantial role in the evaluation process. However, a clear boundary remains between similar and contrasting reports.

5 RESULTS AND DISCUSSION

In our original application of chest X-ray report generation, we incorporate our metric to assess the outputs of various models. We compare our results with the BLEU scores evaluated by these models, specifically, the CXR-RePaiR (Endo et al., 2021) and R2Gen (Chen et al., 2020) models, both being state-of-the-art models for generating chest X-ray reports. Our evaluation focuses on measuring the semantic similarity between the generated reports and the ground truth. Table 5 presents the BLEU scores obtained from these

models and our metric’s semantic evaluation. As anticipated, the BLEU scores are relatively low, signifying a substantial dissimilarity between the generated results and the ground truth for both the CXR-RePair and R2Gen models despite being regarded as state-of-the-art models for chest X-ray report generation. These models still employ the BLEU metric for evaluation, primarily due to the scarcity of more suitable metrics and the need for a standardized evaluation process for comparative purposes. Conversely, our metric produces more promising results for both of these models. While our metric’s scores align with the BLEU scores, indicating higher scores for both BLEU and our MCSE metric in the case of the R2Gen model compared to the CXR-RePair, our metric provides a deeper evaluation. It suggests a degree of similarity to the ground truth rather than outright dissimilarity in BLEU, thus making the generated reports more reliable and trustworthy, which is a crucial advancement in the field.

Table 5: The result of BLEU score of 2-gram for state-of-the-art models and the result of our novel metric on these models outcomes.

Models	BLEU	Our MCSE
R2Gen (Chen et al., 2020)	0.212	0.71
CXR-RePair (Endo et al., 2021)	0.069	0.64

Table 6 provides an example of medical text generated and evaluated using both a BLEU score and our MCSE metric. It’s evident that, according to the BLEU score, these two texts appear vastly different, even though they share the same primary medical entities. However, when we delve into the context, we can notice that “moderately severe” serves as a description for the main entity, “pulmonary edema”, in the generated text. Similarly, in the second part of the text, the main medical entity is “pleural effusions”, and terms like “likely” and “no large” are used to describe this entity, which may not be identical but share semantic similarities. This subtle context evaluation is precisely what our metric considers, yielding a similarity score of 0.64 for these texts, which we argue is a more accurate reflection compared to the BLEU score.

Lastly, the significant benefit of employing this metric lies in its capacity for comparative analysis alongside other evaluation measures. For instance, when examining the outcomes of the BLEU score, with its word-by-word analysis, situations may arise where the results are totally inaccurate, casting doubt on their reliability, despite the models performing well overall. Integrating the results of our novel

Table 6: A comparative example of using the BLEU score and our adapted metric with medical reference and generated text.

	BLEU	MCSE
Reference Sentence: “Pulmonary edema, cardiomegaly, likely pleural effusions.”	0.047	0.64
Generated Sentence: “Moderately severe bilateral pulmonary edema with no large pleural effusion.”		

MCSE metric into the evaluation process allows us to semantically analyze and ascertain the dependability of the models’ textual outputs within the context of medical content.

6 CONCLUSION

In our research, we tackle the challenge of semantic similarity scoring in medical corpora, driven by the inadequacy of existing metrics that, while suitable for machine translation evaluation, fall short in the field of medical semantic assessment. Our innovative metric draws inspiration from how humans comprehend text, centering on the extraction of key terms and their relational context. It introduces a novel approach for extracting clinical entities from medical text, considering not only the entities themselves but also the associated descriptions and negations. Additionally, we created a new method for scoring the semantic relationships between these entities by using the domain cosine similarity. The validation process allowed us to analyze and validate each of these steps individually, unraveling a clear distinction between reports sharing the same diagnosis and those diverging in this regard.

For our research, we focused on the application of chest X-rays, a critical domain where a robust semantic evaluation metric is highly valuable. We applied our metric to some of the latest state-of-the-art models, and the results harmonized with other evaluation metrics, affirming their reliability.

While our validation process and implementation yielded successful outcomes, we encountered the challenge of an inherent bias in domain cosine similarity. This challenge has illuminated a promising direction for our future research, as we explore ways to mitigate this bias and advance the field of medical semantic evaluation.

Material, Codes, and Acknowledgement: Results can be reproduced using the code available in the GitHub repository <https://github.com/sayeh19>

94/Medical-Corpus-Semantic-Similarity-Evaluation.git. All the computations presented in this paper were performed using the (Gricad,) infrastructure (<https://gricad.univ-grenoble-alpes.fr>), which is supported by Grenoble research communities.

REFERENCES

- Alam, F., Afzal, M., and Malik, K. M. (2020). Comparative analysis of semantic similarity techniques for medical text. In *2020 International Conference on Information Networking (ICOIN)*, pages 106–109.
- Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.
- Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China. Association for Computational Linguistics.
- Chen, Z., Song, Y., Chang, T.-H., and Wan, X. (2020). Generating radiology reports via memory-driven transformer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1439–1449, Online. Association for Computational Linguistics.
- Endo, M., Krishnan, R., Krishna, V., Ng, A. Y., and Rajpurkar, P. (2021). Retrieval-based chest x-ray report generation using a pre-trained contrastive language-image model. In *Proceedings of Machine Learning for Health*, volume 158 of *Proceedings of Machine Learning Research*, pages 209–219.
- Gricad. infrastructure supported by grenoble research communities.
- Honnibal, M., Montani, I., Van Landeghem, S., and Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python.
- Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., and Ng, A. Y. (2019). Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):590–597.
- Jain, S., Agrawal, A., Saporta, A., Truong, S., Duong, D. N., Bui, T., Chambon, P., Zhang, Y., Lungren, M. P., Ng, A. Y., Langlotz, C., and Rajpurkar, P. (2021). Radgraph: Extracting clinical entities and relations from radiology reports. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.
- Johnson, A. E. W., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. (2019). Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317.
- Li, J., Sun, Y., Johnson, R. J., Sciaky, D., Wei, C.-H., Leaman, R., Davis, A. P., Mattingly, C. J., Wieggers, T. C., and Lu, Z. (2016). BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068.
- Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Miura, Y., Zhang, Y., Tsai, E., Langlotz, C., and Jurafsky, D. (2021). Improving factual completeness and consistency of image-to-text radiology report generation. In Toutanova, K., Rumshisky, A., Zettlemoyer, L., Hakkani-Tur, D., Beltagy, I., Bethard, S., Cotterell, R., Chakraborty, T., and Zhou, Y., editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5288–5304, Online. Association for Computational Linguistics.
- Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.
- Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (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.
- Patricoski, J., Kreimeyer, K., Balan, A., Hardart, K., Tao, J., Anagnostou, V., Botsis, T., Investigators, J. H. M. T. B., et al. (2022). An evaluation of pretrained bert models for comparing semantic similarity across unstructured clinical trial texts. *Stud Health Technol Inform*, 289:18–21.
- Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., and Lungren, M. P. (2020). Chexpert: Combining automatic labelers and expert annotations for accurate radiology report labeling using bert. In *Conference on Empirical Methods in Natural Language Processing*.
- Yu, F., Endo, M., Krishnan, R., Pan, I., Tsai, A., Reis, E. P., Fonseca, E. K. U. N., Ho Lee, H. M., Abad, Z. S. H., Ng, A. Y., Langlotz, C. P., Venugopal, V. K., and Rajpurkar, P. (2022). Evaluating Progress in Automatic Chest X-Ray Radiology Report Generation. preprint, Radiology and Imaging.