Universität Zürich[UZH]

Bachelorarbeit
zur Erlangung des akademischen Grades
Bachelor of Science in Informatik
der Wirtschaftswissenschaftlichen Fakultät der Universität Zürich

RadEval

A radiology-aware model-based evaluation metric for report generation

Verfasser / Author: Amos Rubin Calamida
Matrikel-Nr / Student ID: 19-734-797

Referent / Supervisor: Prof. Dr. Martin Volk

Betreuer / Advisor: Dr. Farhad Nooralahzadeh, Prof. Dr. Michael Krauthammer

Institut für Informatik / Department of Informatics
Institut für Quantitative Biomedizin / Department of Quantitative Biomedicine

Abgabedatum / Submission Date: 01.04.2023

Zusammenfassung

In unserer Arbeit präsentieren wir eine neuartige, automatisierte Radiologie-spezifische Bewertungsmetrik, die zur Evaluation von maschinell generierten Radiologieberichten verwendet werden kann. Wir nutzen die bestehende, erfolgreiche COMET Metrik-Architektur, welche wir für die Anwendung in der Radiologie anpassen und optimieren. Mit dieser Architektur trainieren und veröffentlichen wir vier medizinisch ausgerichtete Modell-Checkpoints, welche unter Verwendung verschiedener Kombinationen von Encodern und Korpora von Radiologieberichten erstellt werden. Einer der Modell-Checkpoints wird mithilfe von RadGraph, einem Radiologie Knowledge-Graph, trainiert, und die von RadGraph abgeleiteten RadGraph F1- und RadCliQ-Scores werden in unsere Parallel-Korpora integriert, um deren Qualität zu verbessern. Unsere Auswertung der Ergebnisse zeigen, dass die entwickelte Metrik eine mittlere bis hohe Korrelation mit bereits etablierten Metriken wie BERTscore, BLEU und S_emb score aufweist, was auf ihre potenzielle Wirksamkeit als radiologiespezifische Bewertungsmetrik hinweist.

Abstract

In our work, we propose a novel automated radiology-specific evaluation metric that can be used for evaluating the performance of machine-generated radiology reports. We utilize the existing successful COMET metric architecture, which we adapt and optimize for use in the radiology domain. Using this architecture, we train and publish four medically-oriented model checkpoints using various combinations of encoders and corpora of radiology reports. One of the model checkpoints is trained using RadGraph, a radiology knowledge graph, and the thereof-derived RadGraph F1 and RadCliQ scores are integrated into our contributed parallel corpora to enhance their quality. Our results show that the developed metric exhibits a moderate to high correlation with established metrics such as BERTscore, BLEU, and S_emb score, indicating its potential effectiveness as a radiology-specific evaluation metric.

Acknowledgement

I want to express my gratitude to Prof. Dr. Martin Volk, whose support and supervision enabled me to undertake this thesis and to Prof. Dr. Michael Krauthammer for providing me with the opportunity to work on my thesis within the framework of his esteemed Lab.

Moreover, I would like to extend my sincere appreciation to Dr. Farhad Nooralahzadeh for his invaluable guidance, suggestions and assistance throughout the writing, ideation and creation process of my thesis.

Lastly, I would like to thank Yves Lüthi, Tarek Alakmeh and Nino Büchi, very much for their diligent proofreading and constructive feedback.

Contents

List of Figures

List of Tables

List of Acronyms

AAAI	Association for the Advancement of Artificial Intelligence
BLEU	Bilingual Evaluation Understudy
BLEURT	Bilingual Evaluation Understudy with Representations for Transformers
COMET	Cross-lingual Optimized Metric for Evaluation of Translation
CXR	Chest X-ray
ICCV	International Conference on Computer Vision
LLM	Large Language Model
METEOR	Metric for Evaluation of Translation with Explicit ORdering
MeSH	Medical Subject Headings
NIST	National Institute of Standards and Technology
NLG	Natural Language Generation
NLP	Natural Language Processing
PCA	Principle component analysis
QE	Quality Estimation
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
RadCliQ	Radiology Report Clinical Quality
SPICE	Semantic Propositional Image Caption Evaluation
SVD	Singular Value Decomposition
TF-IDF	Term Frequency - Inverse Document Frequency

1 Introduction

1.1 Motivation

I am writing this thesis at the intersection between the Department of Informatics, the Department of Computational Linguistics and the Department of Quantitative Biomedicine at the University of Zurich. My thesis is part of a larger project called ≪Multimodal Multilingual Clinical NLP≫ conducted within the Krauthammer Lab working together with the University Hospital of Zurich. I am extremely thankful that I am able to contribute to an active research project and to have the possibility to propose a new metric to aid the successful automatic generation of reports in the radiology domain's NLG and NLU applications.

1.2 Research Questions

This thesis has the intention to propose a new automatic metric for the evaluation of automatically generated radiology annotation reports by improving upon existing metrics built for other domains (both automatic like *COMET (Crosslingual Optimized Metric for Evaluation of Translation)* [Rei et al., 2020] and classic like *SPIDEr (Semantic Propositional Image Description Evaluation)* [Liu et al., 2017] or *BLEU* [Papineni et al., 2002]) adding a radiology-specific knowledge graph called *RadGraph* [Jain et al., 2021].
Our concrete research questions are as follows:

- RQ 1: Can an existing successful metric model architecture be adapted and optimized to develop a novel radiology-specific metric for evaluating the quality and accuracy of automatically generated radiology reports?
- RQ 2: To what extent does the integration of RadGraph, a radiology-specific knowledge graph, impact the precision and dependability of the assessment metric in evaluating the efficacy and accuracy of automatically generated radiology reports?

1.3 Thesis Structure

The opening chapter of this thesis serves as an introduction to the research topic and the context in which it was conducted.

Chapter 2 aims to provide readers with the necessary background information regarding the technologies and terminologies used throughout the thesis. This chapter reviews relevant literature and identifies key concepts and definitions. Additionally, it discusses the latest trends and developments in the field to contextualize the research and provide a comprehensive understanding of the research problem.

Chapter 3 describes the methodology used to develop the new metric. This chapter presents the two approaches considered for developing the proposed metric and elaborates on the selected approach, including the process of creating the data foundation, selecting the appropriate model architecture, and performing the training.

The penultimate chapter 4 presents the results of the thesis, including a comparison of the proposed metric with other commonly used metrics. This chapter provides an analysis of the performance of the proposed metric, as well as its advantages and limitations.

The final chapter 5 serves as the conclusion of this thesis, providing a reflection on the research outcome and its implications. This chapter summarizes the main findings and contributions of the research, and discusses its limitations and future directions for further research.

2 Background

2.1 Natural Language Generation

Natural Language Generation (NLG) is a rapidly growing subfield of Natural Language Processing (NLP) research that focuses on generating coherent and grammatically correct text in natural language. NLG technology has numerous applications, including the automated generation of reports, letters, and poems, as well as summarizing long articles, paraphrasing sentences, and performing automatic translation [Gatt and Krahmer, 2018].

Even though it is difficult to get a complete and accurate definition of NLG [Evans et al., 2002], there is usually a classification of tasks within NLG. There are two broad categories: text-to-text generation and data-to-text generation [Gatt and Krahmer, 2018]. Text-to-text generation involves taking existing text and transforming it into a well-formed new text, while data-to-text generation involves extracting information from large volumes of structured or non-structured data (which doesn't have to be text but also images [Dong et al., 2022] or other media) gathered from various sources such as the internet or human interactions, and then generating new text based on that information [Gatt and Krahmer, 2018]. Modern NLG systems such as GPT-3 [Brown et al., 2020] and ChatGPT [OpenAI, 2022] have the ability to deduce answers to complicated questions, which may have never been answered in that exact way by a human before, making them good examples of data-to-text systems (albeit without the multi modality, as these systems as for now do not allow other inputs than text).

These NLG systems are powered by advanced Large Language Models (LLMs), which are Deep Neural Network models that work by predicting the probability of a given word or sequence of words appearing next in a given context [Bender et al., 2021; Qiu et al., 2020]. LLMs have been instrumental in advancing the field of NLG, allowing for more sophisticated and accurate text generation

As NLG technology continues to improve, it is expected to become even more prevalent in various industries, including critical applications such as in medicine.

2.2 Evaluation in NLG tasks

Evaluation metrics are crucial for assessing the performance of NLU and NLG systems. There are several types of metrics available for this purpose, with traditional metrics being the most widely used due to their ease of calculation. These metrics were originally developed for summarization and machine translation applications but have since been extended to other NLU and NLG tasks. However, traditional metrics have shown limitations in their correlation with human judgments [Blagec et al., 2022; Sai et al., 2022]. Still, more than half of the recent publications in NLP have been relying upon automated, mostly traditional metrics [Novikova et al., 2017]. Because of that, Novikova et al. [2017] argued that there is a need for newer evaluation metrics.

In the meantime, several new metrics appeared, but haven't been able to rival the traditional metrics. According to Leiter et al. [2022] those newer metrics haven't made their way into literature because they suffer among others from poor explainability of their underlying models and the fact that they have been developed quite recently and can therefore not be used as a benchmark with earlier publications which only report traditional metric scores. In the following, we explain the different types of metrics that are among the most popular in usage (according to Blagec et al. [2022]) and give examples of them.

2.2.1 Word based (traditional) metrics

These metrics are based on lexical overlap or common subsequences of two texts. Examples of such are:

BLEU (Bilingual Evaluation Understudy by Papineni et al. [2002])
A word-overlap metric for evaluating the quality of machine translation systems. It compares the machine-generated output to one or more reference translations (i.e. ground truth) and calculates a score based on the n-gram overlap between the reference and machine-generated sentences. The BLEU score is calculated using the following formula:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

where BP is the brevity penalty, which adjusts the score based on the length of the

machine-generated output relative to the reference translations. The weights w_n and precision scores p_n are used to calculate the n-gram overlap.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation by Lin [2004])
A set of word-overlap metrics for evaluating the quality of summarization systems. It measures the n-gram overlap and longest common subsequence (LCS) between the summary generated by the system and the reference summary. The ROUGE-N (n-gram recall) score is calculated using the following formula:

$$\text{ROUGE-N} = \frac{\sum_{S \in \text{ReferenceSummaries}} \sum_{\text{gram}_n \in S} \text{Count}_{\text{match}}(\text{gram}_n)}{\sum S \in \text{ReferenceSummaries} \sum_{\text{gram}_n \in S} \text{Count}(\text{gram}_n)}$$

where gram_n is an n-gram of length n, $\text{Count}_{\text{match}}(\text{gram}_n)$ is the number of times an n-gram gram$n$ appears in both the generated summary and the reference summary(s), and $\text{Count}(\text{gram}_n)$, which is the total number of times an n-gram gram$n$ appears in the reference summary(s).

ROUGE has several other variants:

- ROUGE-1 and ROUGE-2 compute the overlap of unigrams and bigrams between the generated and reference summaries, respectively.
- ROUGE-L considers the longest common subsequence (LCS) between the generated and reference summaries, which is the longest sequence of words that appear in the same order in both summaries.
- ROUGE-S is a skip-bigram variant that takes into account the skipping of words between two words that appear consecutively in a sequence.

METEOR (Metric for Evaluation of Translation with Explicit ORdering by Banerjee and Lavie [2005])
A metric for evaluating machine translation systems. It compares the machine-generated output to one or more reference translations and calculates a score based on the number and the alignment of matched words. The score is calculated as follows:

$$\text{METEOR} = (1 - \text{p}) \cdot \text{M} + \text{p} \cdot \text{P}_{\text{align}}$$

where M is the unigram precision and recall scores, weighted equally, and P_{align} is the alignment-based precision score, which is a function of the number of word matches and the number of alignment errors. The hyperparameter p determines the

weight given to the alignment-based score.

SPICE (Semantic Propositional Image Caption Evaluation by Anderson et al. [2016])
An evaluation metric for image captioning that measures the semantic similarity between a generated caption and a reference caption. SPICE considers the scene graph structure of the image, which describes the objects in the image and their relationships, and uses it to create semantic propositions. These propositions capture the meaning of the image and serve as a basis for comparing the generated and reference captions.

$$\sum_{p \in \text{propositions}(I_i)} \min\left(\text{Count}_{\text{gen}}(p), \text{Count}_{\text{ref}}(p)\right) \cdot f(p)$$

where N is the number of images, $\text{Count}_{\text{gen}}(t)$ and $\text{Count}_{\text{ref}}(t)$ are the counts of term t in the generated and reference captions, respectively, and $\text{propositions}(I_i)$ is the set of semantic propositions for image I_i. The function $f(p)$ computes the weight of each semantic proposition p based on its salience score, which reflects how important the proposition is in describing the image.

2.2.2 Language model based metrics

These metrics are based on machine learning models / neural networks.
Examples include:

BLEURT (Bilingual Evaluation Understudy with Representations for Transformers by Sellam et al. [2020])
A variant of the BLEU metric used to evaluate the quality of machine translation output by comparing it to human translations. BLEURT is based on the Transformer architecture, commonly used in natural language processing tasks. The metric uses pre-trained Transformer models to generate sentence-level representations, which are then used to compute a similarity score between the machine-translated output and the human translation.

BERTScore (by Zhang et al. [2020a])
A metric to evaluate the similarity between two pieces of text. It uses a pre-trained BERT (Bidirectional Encoder Representations from Transformers) model to encode

the two pieces of text and measure their similarity based on their contextualized embeddings.

COMET (Cross-lingual Optimized Metric for Evaluation of Translation by Rei et al. [2020])
A machine learning based metric for evaluating machine translation systems. COMET is trained on human judgments of translation quality and uses a neural network to predict the quality of a translation based on its similarity to human translations.

2.3 Current state of metrics in medical image report generation

Various automatic and semi-automatic metrics have been developed for evaluating the quality and accuracy of reports generated from medical images. These metrics often emerge from recent publications on report generation, which aim to establish more relevant evaluation measures in addition to traditional ones such as BLEU and ROUGE.

Two such metrics include *Medical abnormality terminology detection* by Li et al. [2018] and *MeSH accuracy* by Huang et al. [2019]. The former measures the accuracy and error rate of a system's ability to identify common medical abnormalities in a report, while the latter calculates the accuracy of a report "[...] as the ratio of the number of MeSH correctly generated by a model to the number of all MeSH in the groundtruth [sic] [...]" [Huang et al., 2019, p. 154813]

Another metric, *Medical Image Report Quality Index* by Zhang et al. [2020b], utilizes multiple criteria to assess the quality of medical image reports, including the relevance of medical terms, syntactic structure, and semantic coherence. This approach is similar to the *Anatomical Relevance Score* (ARS) metric developed by Alsharid et al. [2019] for ultrasound image captioning. ARS, which is a type of F1 score, evaluates the anatomical relevance of report sentences by comparing them to a reference standard.

Even though those metrics exist, they are not widely adopted, and even new publications mostly rely on traditional NLP metrics, which are not considering medical aspects and thus give less meaningful evaluations than specialized metrics could be [Messina et al., 2022].

2.4 What makes a good metric?

When evaluating the performance of natural language generation systems for image captioning tasks in medical radiography, it is important to consider the unique properties of radiology reports.

For classic image captioning, Sai et al. [2022] propose to take into account factors such as evaluating the fluency of the style, the coverage of all important entities on the image, and their relationship amongst them. They also argue metrics should not be favoring long captions with unnecessary details over short and concise ones. While these guidelines by Sai et al. [2022] for metrics in classic image captioning tasks (in their example a Photo of a city street), provide a useful starting point, they may not be sufficient to capture the specific requirements of radiology reports.

Radiology reports need to reflect important properties of the entities depicted in the scan. They are narratives consisting of multiple sentences, including the exact position and severity of abnormalities, as well as concluding remarks summarizing the most prominent observations. Radiology report generation is a more challenging task, as the reports have their own distinctive characteristics and demand accurate clinical descriptions [Langlotz, 2015]. As current metrics like BLEU are not capturing those specific properties, there is a need for domain-specific metrics that take into account the unique properties and requirements of radiology reports [Chen et al., 2020].

In summary, a good metric for evaluating natural language generation systems for radiology report generation should take into account not only the fluency and coverage of the generated text but also its adherence to the required structure and the accuracy of the medical information it conveys.

2.5 Ranking-based human evaluation

Callison-Burch et al. [2007] compare commonly used evaluation methods based on their inter-annotator agreement performance and the time it takes to complete the evaluation.

In the past, metrics have been evaluated by using the model-assigned score for each text and then comparing this score to a score that the human evaluator gave the same text (e.g. on the quality of a generated report in our case).

Humans are, however, not very good at assigning precise scores to a text so the

approach that the authors discovered to be more effective is to task the human evaluators with ranking a number of texts (in their case 5) by their overall quality and then compare this ranking to the scores outputted by the model. This resulted in a higher inter-annotator agreement compared to the previous methods [Callison-Burch et al., 2007].

2.6 A radiology-specific knowledge Graph

In a 2021 publication by Jain et al. the authors implement a radiology-specific knowledge graph. RadGraph extracts the two types of entities *Anatomy (ANAT)* and *Observation (OBS)* together with one of three certainty quantifiers *definitely present (DP)*, *uncertain (U)* and *definitely absent (DA)* (see Table 1). Furthermore, three relationships that can be attributed to the entities are acquired, specifically: *suggestive of*, *located at*, and *modify* (see Table 2).

The knowledge graph was trained using data from radiology reports, which were annotated by board-certified radiologists using the aforementioned schema of entities and relations.

RadGraph is designed to support various NLP tasks, such as information extraction and text classification, and can be used to generate structured radiology reports. The system has been evaluated on a dataset of radiology reports and has demonstrated promising results in terms of accuracy and efficiency [Jain et al., 2021].

An example of a RadGraph annotation can be seen in Figure 1, where a radiology report has been processed to identify the relevant entities and their relationships in the form of a graph.

RadGraph is publicly available on PhysioNet [Goldberger et al., 2000] under certain conditions [1].

[1]The PhysioNet data use agreement contains, among others, the following points [PhysioNet]:

1. The user will not share the data
2. the user will make no attempt to re-identify individuals
3. any publication that makes use of the data will also make the relevant code available.

ANAT	"[. . .] an anatomical body part that occurs in a radiology report, such as a 'lung'." [Jain et al., 2021, p. 4]
OBS	"[. . .] observations made when referring to the associated radiology image [. . .] associated with visual features, identifiable pathophysiologic processes, or diagnostic disease classifications." [Jain et al., 2021, p. 4]

Table 1: Entities as defined in the RadGraph dataset

suggestive_of	(Observation, Observation)	"[. . .] is a relation between two Observation entities indicating that the presence of the second Observation is inferred from that of the first Observation." [Jain et al., 2021, p. 4]
located_at	(Observation, Anatomy)	"[. . .] is a relation between an Observation entity and an Anatomy entity indicating that the Observation is related to the Anatomy. While Located At often refers to location, it can also be used to describe other relations between an Observation and an Anatomy." [Jain et al., 2021, p. 4]
modify	(Observation, Observation) or (Anatomy, Anatomy)	"[. . .] is a relation between two Observation entities or two Anatomy entities indicating that the first entity modifies the scope of, or quantifies the degree of, the second entity. As a result, all Observation modifiers are annotated as Observation entities, and all Anatomy modifiers are annotated as Anatomy entities for simplicity." [Jain et al., 2021, p. 4]

Table 2: Relations as defined in the RadGraph dataset

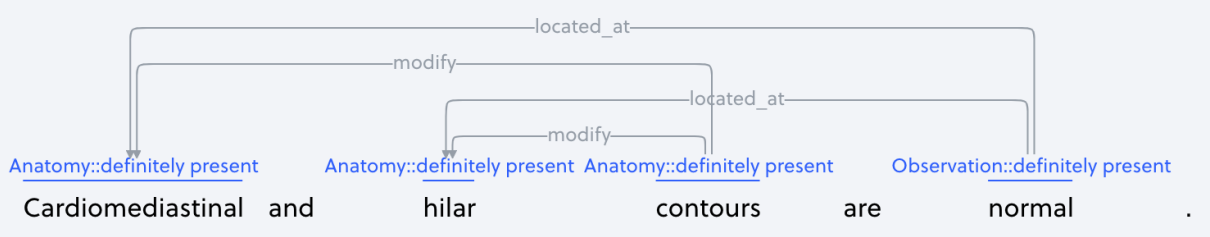


Figure 1: Example annotation using RadGraph labels and entities (image provided by the authors Jain et al. [2021])

3 Metric Proposition

As we have seen, there is a need for more specific and reliable metrics for evaluating the accuracy of texts, in particular image captions in the medical field. In our work, we focus specifically on radiology reports and therefore considered two different approaches for developing a new metric:

At first, we were considering an approach involving RadGraph [Jain et al., 2021] or a similar knowledge graph such as Abstract Meaning Representation (AMR) [Knight et al., 2014] to derive graphs for both the generated radiology reports and their corresponding ground truth. These graphs provide a structured and interpretable representation of the semantic meaning of the reports. Subsequently, we would employ graph similarity measures (e.g. sub-graph comparison) to compare the graph of the generated report with that of the ground truth, which would allow us to derive a numeric score of graph similarity. This score provides a quantitative measure of how well the generated report matches the ground truth, allowing for a more objective and consistent evaluation of report accuracy.

While working on this thesis, two papers were presented by Delbrouck et al. [2022] and Yu et al. [2022], respectively, which employed an approach similar to the one we were about to propose. In light of that, we decided to no longer pursue this direction and instead focus on a second approach. We directly incorporate the two new evaluation measures (RadGraph F1 and RadCliQ) proposed by Yu et al. [2022] into our model architecture to build a stable parallel corpus and training foundation for our automatic approach.

As the second approach, we decided to use an automatic model evaluation method that involves training a neural network on a parallel corpus of radiology reports. Specifically, we use the referenceless quality estimation (QE) metric architecture proposed by Rei et al. [2020], which does not require a reference input (see Table 3 for the mapping of inputs). The architecture of the QE metric consists of a pooling layer that aggregates the word embeddings of the radiology report, followed by a feed-forward neural network that maps the pooled representation to a scalar quality score. The network is trained using mean square error as the loss function, with the

quality score being predicted based on the similarity between the embeddings of the generated report and those of a reference report.

In order to train the model, a parallel corpus was constructed using the Indiana University X-Ray (IU X-Ray) dataset [Demner-Fushman et al., 2016], a widely utilized dataset within the radiology domain. The IU X-Ray dataset contains one to two chest X-Ray scans per data point, along with accompanying reports of actual findings, brief summaries of these findings (referred to as the "impression"), and assigned Medical Subject Headings (MeSH) labels. MeSH is a controlled vocabulary used by the National Library of Medicine database to index and organize biomedical information [National Library of Medicine, 2023]. These terms are used to categorize medical articles based on their content and encompass a broad range of medical topics, including anatomy, diseases, drugs, and procedures. An example can be seen in Figure 2.

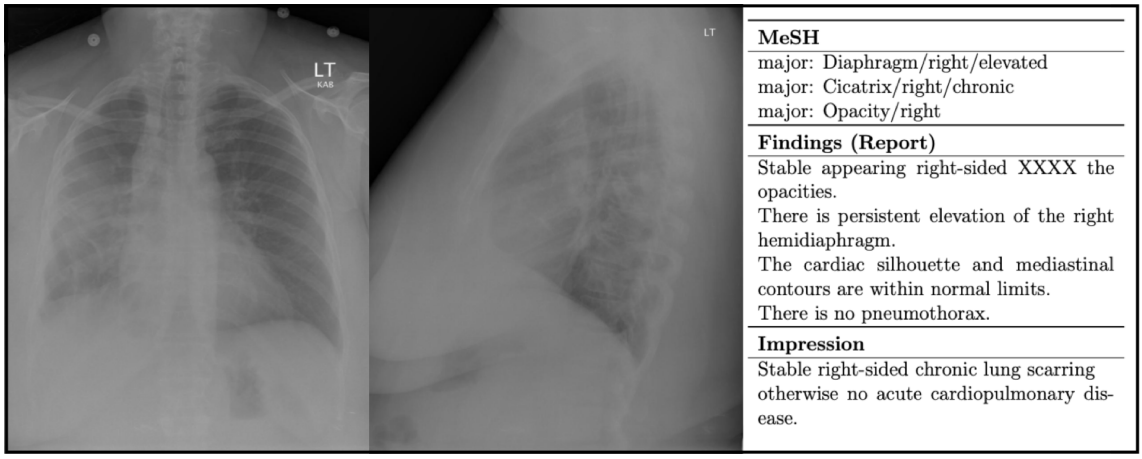


Figure 2: An example report showing the two images and the MeSH, findings and impression columns (Image constructed by the author with the data from Demner-Fushman et al. [2016]).

Once the model is trained, we can use it to evaluate the accuracy of a generated report by comparing the output score, which provides a quantitative measure of the similarity between the generated report and the ground truth. This score can be used to assess the accuracy of the generated report.

3.1 Automated Metric based on Parallel Corpus

The main approach we explore to develop our radiology-specialized metric is to use the COMET evaluation architecture by Unbabel AI [2020] (based on the work by

Rei et al. [2020]) and train our own model on radiology data with a focus on the technicalities that medical metrics must follow as outlined in section 2.4.

3.1.1 Available metric architectures

The COMET neural evaluation framework by Unbabel provides three different metric model types that can be used to develop novel metrics [Unbabel AI, 2020]:

- Regression metric: Given a source, hypothesis and reference performing a regression task to assign a quality score to a hypothesis.
- Referenceless metric: Given only a source and hypothesis, performing a quality estimation.
- Ranking metric: Given a source, a negative example, a positive example, and a reference, arranging the embedding space to give greater importance to positive examples and less attention to negative ones.

We are focusing on the referenceless Metric as our targeted input data will be consisting of two radiology reports, one ground truth (i.e. the source), and one model generated (i.e. the hypothesis). The architecture for the referenceless Metric is visualized in Figure 3.

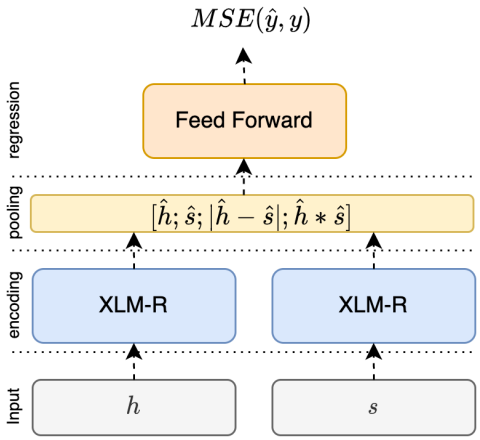


Figure 3: Model architecture for the referenceless metric in the COMET Framework (Image provided by Unbabel AI [2020]).

3.2 Parallel Corpus

Because the COMET architecture is built for assessing the quality of Machine Translation it requires a parallel corpus of source (i.e. the original text), hypothesis (i.e. the machine translation), and reference (i.e. the correct translation of the source) as input to train the model. In our case, we propose the mapping of these input concepts from the machine translation to report generation space as seen in table 3.

Input	**Machine Translation**	**Report Generation**
Source	Original Text (Source Language)	Human-written report
Hypothesis (1...n)	Machine Translation (Target Language)	Model Generated Report
Reference	Correct Translation (Target Language)	Human-corrected version of the generated report (not needed for our referenceless approach)

Table 3: Mapping proposition of COMET Architectures' input parameters to the report generation problem space

To ensure the reliability of our model, we require a sufficiently large number of reports for training. To construct the training data for our metric, we propose a parallel corpus of similar reports. We used the Indiana University X-Ray Collection as our initial dataset [Demner-Fushman et al., 2016]. To increase the likelihood of having similar reports and to reduce the number of reports to process, we performed a K-Means clustering, as described in subsection 3.2.1. This clustering process allowed us to take the cross-product of each cluster, instead of the entire dataset, thereby reducing computational load and the amount of data to score.

Next, we scored the similarity of reports in relation to all other reports in the same cluster, as explained in subsection 3.2.3. We partitioned the resulting dataset into test and train sets, ensuring an equitable distribution of normal and abnormal reports, as detailed in subsection 3.2.2. The output was a parallel corpus of reports that were similar to each other. This parallel corpus served as input for the referenceless QE framework proposed by Unbabel AI [2020]; Jain et al. [2021].

3.2.1 Clustering

We looked at clustering the IU X-Ray reports based on two different columns:

1. Using the **impression column** of the report
2. Using the **combined Medical Subject Headings (MeSH) columns (type: major)** of the report

The MeSH column in the IU X-Ray dataset contains major and automatic labels. MeSH is a controlled vocabulary used to index and organize biomedical information in the National Library of Medicine database. MeSH terms are assigned to medical articles to describe the content of the article, and the MeSH vocabulary covers various aspects of medicine, such as anatomy, diseases, drugs, and procedures, to name a few [National Library of Medicine, 2023]. In IU X-Ray the Major MeSH labels are assigned by human experts with medical knowledge by using expert judgment. Automatic MeSH labels, on the other hand, are generated using an NLP system that extracts keywords from the findings in the reports associated with each X-ray image. Type Major MeSH labels are therefore considered to be more reliable than automatic MeSH labels.

To create the combined MeSH column for our clustering, we concatenated all major MeSH labels and discarded all automatic ones. We also removed the MeSH values "no indexing" and "technical quality of image unsatisfactory" from our MeSH column, as they are not representative of the dimensions we want to cluster on. We aimed to cluster the reports based on the actual abnormality and not based on a technical label related to the image quality or indexing.

To prepare the data for clustering on the impression column, we removed all reports with a MeSH of "major: normal" and "major: No Indexing" as we want to keep all normal reports in one cluster and the abnormal ones clustered by type. We introduced this step after the first try of clusters showed that the majority of terms in all clusters were dominated by different "normal" terms (e.g. "no acute findings", "no acute disease", etc.)

For determining the optimal number of clusters to use, we followed the following steps:

1. Using a TF-IDF (Term Frequency-Inverse Document Frequency) Vectorizer to transform the MeSH and Impression columns respectively into vectors.
2. Apply a TruncatedSVD (Singular Value Decomposition) and Normalizer (both from sklearn) to reduce the dimensionality of the data.
3. Run K-means clustering for each cluster n with $2 \leq n \leq 40$ on the transformed data.

4. Calculate Silhouette, Elbow, and Calinski-Harabasz scores for each number of clusters n.

   The Silhouette Score measures how well each object fits within its own cluster compared to other clusters. It is calculated as the difference between the average distance of an object to all other objects in its own cluster and the average distance to all objects in the nearest cluster. A higher Silhouette Score indicates better-defined clusters.

   The Elbow score measures the amount of variance explained by a clustering algorithm as a function of the number of clusters. It aids in determining the optimal number of clusters by identifying the point where adding more clusters would only lead to marginal improvement in the explained variance. The optimal number of clusters is often identified as the "elbow point" where the rate of variance explained starts to level off.

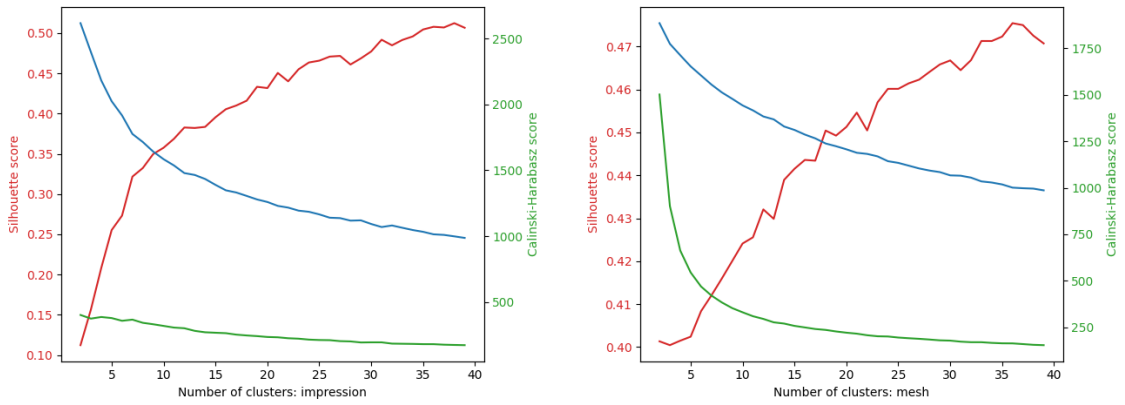
   The Calinski-Harabasz score (also known as the variance ratio criterion) measures the ratio of between-cluster variance to within-cluster variance. It provides a measure of how well-separated the clusters are, with higher scores indicating more distinct clusters. This score is particularly useful for complex datasets where the optimal number of clusters is not immediately apparent [Caliński and Harabasz, 1974]. Like the Silhouette and Elbow scores, a higher Calinski-Harabasz score indicates better clustering.

5. Visual inspection of Figure 4 and the resulting cluster contents for the different amounts of clusters (e.g. Table 4) to determine the optimal number of clusters to fit the report data. You can find the complete list of scores in the appendix (Table 12 and Table 13).

   In this step, we used the three scores together to get a more complete picture than we would get by just looking at one score.

For the impression column, we can see, that the Calinski-Harabasz Score stays relatively even with a slight decline until 20 clusters and the Silhouette Score takes a steep incline and then flattens after 10 clusters which are also close to the slight elbow of the inertia. As the Calinski-Harabasz Score is less relevant for the impression column as the data it contains is plain sentences with great variance, we weighted the two other scores higher and experimented with $12 \leq n \leq 15$ and got the best-defined results at $n = 14$.

For the MeSH column we observe, that the Calinski-Harabasz Score takes a very steep decline until 5 clusters. The Silhouette Score starts slow and then climbs



(a) Impression Column

(b) MeSH Column

Figure 4: Silhouette score (red), Elbow score/inertia (blue), and Calinski-Harabasz score (green) for increasing amounts of clusters on the two columns MeSH and Impression (the data backing this figure can be found in Appendix A: Table 12 and Table 13)

evenly starting at 5 clusters until 37 clusters. The Elbow score has an even and slow downward slope. We focus most on Calinski-Harabasz for the MeSH column, as the contents of it are more complex and nuanced (i.e. usually anatomy with different quantifiers), and don't follow a natural sentence structure. We therefore experimented with $2 \leq n \leq 12$ and achieved best results with $n = 6$

We can see that with those cluster numbers, MeSH clusters are in general better defined (+300% higher starting silhouette score and +300% higher starting Calinski-Harabasz score)

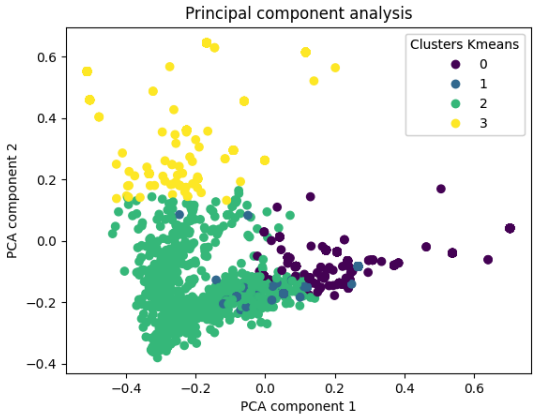
Using the optimal number of clusters, we ran the K-means (random_state : 0, n_init : 5) and K-modes random_state : 0, init : Cao, n_init : 5) algorithms to cluster the reports. Unfortunately, the latter algorithm did not produce useful clusters, so we focused on K-Means.

Looking at the principle component analysis (PCA) for the resulting clusters, we came to the following conclusions:

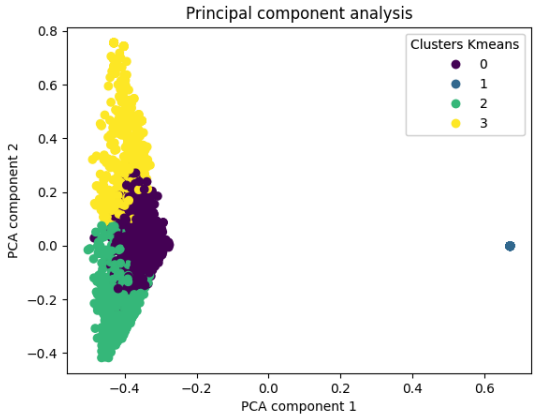
For the MeSH Clustering the normal reports are all grouped in one cluster, which is beneficial to our goal of creating a balanced corpus of normal and abnormal reports.

For both impression and MeSH there the clusters look well defined at 4 clusters each (see: Figure 5) but according to our score inspection, the values number of clusters for the impression should be higher.

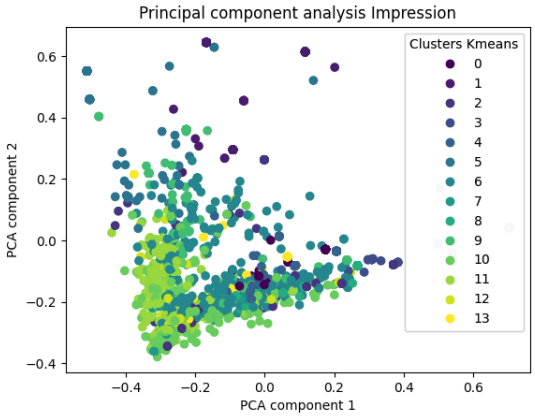
We ran our PCA also for the number of Clusters specified above for the two columns, the resulting clusters can also be seen in Figure 5. The most prominent labels for each cluster are shown in Table 4.



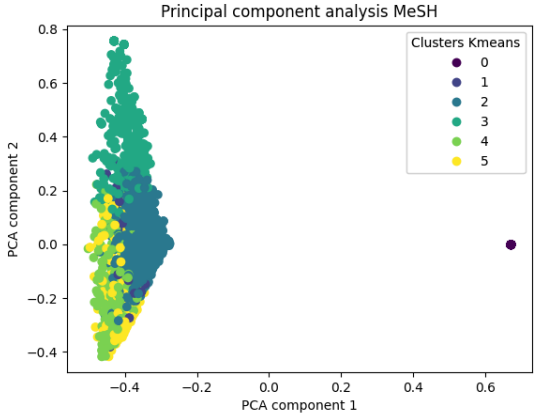
(a) **Impression** $n = 4$: Cluster 2 is significantly bigger than the other clusters.



(b) **MeSH** $n = 4$: The outlier cluster 1 contains the normal reports.



(c) **Impression** $n = 14$: The clusters are not very well defined. Cluster 7 are the normal reports (i.e. "Normal chest" label).



(d) **MeSH** $n = 6$: Outlier (Cluster 1) are the normal reports. Clusters 2 and 3 are well defined, 4 and 5 have a lot of overlap.

Figure 5: Visualization of the clusters generated by KMeans using PCA with $n_{MeSH} = n_{Impression} = 4$ clusters in the first row and $n_{MeSH} = 6$ clusters and $n_{Impression} = 14$ in the second row. The data has been reduced to two dimensions using PCA and the clusters are color-coded.

The analysis indicates that the impression clusters primarily consist of normal findings (e.g., "no acute/active [...]", "Clear [...]") across all clusters. To achieve cleaner impression clusters, we attempted to remove such instances using a combination of filters based on the MeSH and impression columns. We first tried removing all rows with MeSH label "normal" and "No Indexing" to reintroduce them as one single cluster after the remaining reports had been clustered on less normal impres-

ID	Most prominent term	Second most prominent term
	Impression Clusters	
0	No acute cardiopulmonary abnormalities.	No acute cardiopulmonary abnormalities. .
1	No acute disease.	No acute cardiopulmonary disease.
2	Negative for acute abnormality.	Negative for acute cardiopulmonary abnormality.
3	No acute cardiopulmonary abnormality.	No acute cardiopulmonary abnormality. .
4	No acute radiographic cardiopulmonary process.	No acute cardiopulmonary process.
5	No active disease.	No evidence of active disease.
6	No acute intrathoracic abnormality.	Emphysema without acute disease.
7	Normal chest	Normal chest.
8	No acute cardiopulmonary findings.	No acute cardiopulmonary findings. .
9	No acute pulmonary disease.	No acute pulmonary abnormality.
10	Low lung volumes, otherwise clear.	XXXX change. Hypoinflation with no visible active cardiopulmonary disease.
11	Right upper lobe pneumonia.	No focal lung opacity, pleural effusion of pneumothorax.
12	Clear lungs.	Clear lungs. No acute cardiopulmonary abnormality. .
13	No acute cardiopulmonary finding.	No acute cardiopulmonary finding. Emphysema and atherosclerosis.
	MeSH Clusters	
0	normal	*No value*
1	lung/hypoinflation	lung/hypoinflation markings/bronchovascular
2	granulomatous disease	cardiomegaly/mild
3	thoracic vertebrae/degenerative/mild	thoracic vertebrae/degenerative
4	calcified granuloma/lung/base/right	calcified granuloma/lung/base/left
5	calcified granuloma/lung/upper lobe/left	calcified granuloma/lung/upper lobe/right

Table 4: The most common and second most common terms for each cluster in Impression and MeSH Clusters by numeric cluster Identifier (ID)

sion terms. Unfortunately the resulting clusters still contained mainly variations of the aforementioned "no acute/active [...]" sentences. Therefore we tried removing rows containing these terms in the impression column. However, after applying all the filters, only 1049 out of the original 3414 rows remained, with clusters containing 100 reports or less, which lead us to focus mainly on the MeSH clusters for the further creation of the corpus.

```
df['mesh-0'] != "{'type': 'major', 'label': 'normal'}"
df['mesh-0'] != "{'type': 'major', 'label': 'No Indexing'}"
not df['IMPRESSION'].str.contains("No acute", case=False).
not df['IMPRESSION'].str.contains("No active", case=False).
```

Listing 3.1: Filters applied to the data to obtain cleaner impression clusters

3.2.2 Data Preparation

To prepare for the training and validation of our referenceless QE metric, we performed the following steps:

We removed single word or empty reports, as these reports do not contain enough information to be useful for training or validation and can lead to breaking the RadGraph modeling code we rely on to score our parallel corpus.

To avoid bias towards normal reports, which make up one-third of the IU X-Ray dataset, we balanced the dataset by including an equal number of normal and abnormal reports. This was achieved by looking at the clustering of the reports. As for the MeSH-clustering, all normal reports are in the same cluster (because of them having the MeSH tag *major: normal*. We are selecting at most n normal reports out of that cluster to include in the completed parallel corpus, where n is the maximum number of abnormal reports in the largest cluster of said reports.

To evaluate the effectiveness of the balanced dataset, we compared the results of our QE metric when trained on this dataset to those obtained from a random 80/20 split of the original dataset (see chapter 4).

3.2.3 Similarity Scoring

3.2.3.1 Method 1: BERTalign

We first looked at scoring our corpus using BERTalign by Liu and Zhu [2022], a sentence-aligner that works by first encoding each sentence into a sequence of embeddings and then computing a similarity score between each pair of embeddings. The alignment is ultimately obtained by maximizing a joint probability that considers both the similarity scores and the length of the aligned phrases. We were able to match some of the reports but with very low accuracy. We considered all reports, that have the same impression column to be part of our gold-standard parallel corpus. Unfortunately, not even half of the matches produced by BERTalign were found in the gold-standard. To improve the method we tried replacing the encoding layer from standard BERT with the BioClinical BERT encoder from Alsentzer et al. [2019]. This improved the results by a small margin but did not yield a much better matching in total.

3.2.3.2 Method 2: RadCliQ (Radiology Report Clinical Quality)

The RadCliQ Metric is a novel evaluation measure for the similarity of clinical reports, which leverages the BLEU-2 score and the RadGraph F1 metric. The latter ""computes the overlap in clinical entities and relations that RadGraph extracts from machine- and human-generated reports" [Yu et al., 2022, p. 4]. To assess the similarity between reports within the same cluster, we adopted the RadCliQ Metric. The input to the inference process consists of two files, one containing the source reports and the other containing the reference reports. Each row in these files corresponds to a unique study ID, which is employed to match the source and reference reports. The output of this process is a table, where each row contains the source report, its corresponding study ID, and the following four evaluation scores: BLEU-2, BERTscore, CheXbert labeled vector similarity (s_emb Score) [Smit et al., 2020], and Radgraph F1, as well as the RadCliQ score.

From our RadGraph inference on the IU X-Ray dataset, we can see that RadGraph recognized on average 17.3 entities and 10.7 relations in the reports (see Table 5).

Initially, we developed a parallel corpus by selecting the top-scored (i.e. most similar) match for each report (based on the RadGraph metric), resulting in a corpus that encompasses all reports of the cleaned IU X-Ray dataset at least once (i.e. the corpus size is equal to the size of the cleaned IU X-Ray dataset and each report in the dataset has one corresponding report, which matches best in terms of similarity).

RadGraph	Entity		Relation	
	Count	F1	Count	F1
Average	17.300	0.286	10.688	0.118

Table 5: The average count of RadGraph relations and entities including their corresponding F1 score for the inference run to obtain the RadCliQ score

To further explore the space of possibilities, we explored the generation of a secondary parallel corpus with a different approach: We allow the appearance of multiple instances of single reports in the corpus, in the event that they have multiple best matches. Conversely, we intended to limit the corpus by applying a score threshold so as to not artificially blow up the corpus. In this configuration, only a subset of all cleaned reports is still in the corpus. To obtain the most similar reports, we select the rows with the highest scores in the CXR (RadCliQ) combined metric evaluation. These selected reports are again corrected to include an equal share of normal and abnormal reports.

3.2.4 Train/Test Split

After having created the parallel corpus we divided it into two distinct subsets, a training set and a test set. We created a random split of 80/20 using the pandas method `sample` with random_state : 0 to extract 20% of the data into the test set and keep the remaining 80% as the training set. This ensured that our model can be trained and evaluated on two distinct sets of data. With the training process in mind, we also split the training data set further into two subsets, the primary training set and the validation subset, using the same 80/20 split to `sample` (random_state : 0) the validation data out of the training set. This validation set is provided to the model trainer to fine-tune its hyper-parameters on each epoch.

The complete large-scale corpus of reports has comprised a total of 737,015 rows. After applying the filtering process to pick the top 10% of the corpus, the filtered output corpus resulted in a size of 73,692 rows.

The further split of the filtered corpus into the three subsets resulted in the training set containing 47,163 rows, the validation set containing 11,792 rows, and the testing set containing 14,739 rows (see Table 6).

Corpus	Total rows
Raw Corpus	737,015
Filtered (Top 10%)	73,692
Training	47,163
Validation	11,792
Testing	14,739

Table 6: The size of the different resulting corpora

The final corpora all consist of 12 columns: report_id (integer), report (string), mesh (string), reference_report_id (integer), reference_report (string), reference_mesh (string), study_id (integer), bleu_score (float), bert_score (float), semb_score (float), radgraph_combined (float), and cxr_metric_score (float).

3.3 Training our Metric Model

In our thesis, we employ the COMET Architecture by Unbabel AI [2020] to train a proof-of-concept metric. To implement the training process, we followed the authors' recommendations and utilized PyTorch lightning configuration files. The complete list of parameters used in the configuration files is provided in Appendix B.

Most importantly, during the training of our model, we configured the early stopping and model checkpoints to monitor and optimize the Kendall Tau value, which is a commonly used measure of the correlation between two ranked lists to evaluate the similarity between the predicted and ground truth rankings. By optimizing the Kendall Tau value, our model learns to predict semantic similarity scores that are more similar to the ground truth scores we see in the parallel corpus, which helps improve the accuracy and robustness of the model's predictions.

Initially, we set the maximum number of training epochs to 5. However, we observed that during the training process, we consistently reached the maximum number without triggering early stopping, prompting us to increase the maximum number of epochs first to 20 and eventually to 40 epochs, which in return yielded much higher Kendall Tau values (see section 4.2 and Table 7 for details).

We have trained our model on both the parallel corpus based on the impression clusters and on the one based on the MeSH clusters. But as the MeSH corpus

quickly outperformed the impression corpus on the correlation to other metrics, we are focusing on the latter.

Furthermore, we investigated the performance impact of replacing the encoder layer in the COMET Architecture from the default XLM-RoBERTa [Conneau et al., 2020] to BioClinical BERT [Alsentzer et al., 2019]. Also, for the top 10% corpus we compared the performance of training the model on the RadCliQ Score vs. on the RadGraph F1 score.

4 Results

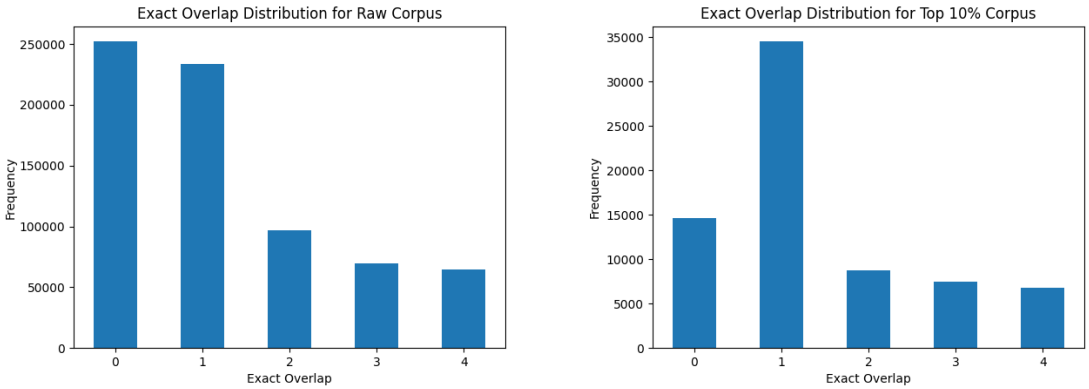
4.1 Parallel Corupus Validation

Our motivation for providing a parallel corpus is to assist future researchers in training their own metrics using a "Source - Reference' model architecture in their research. The parallel corpus built on RadGraph F1 Similarity offers the advantages of this graph-based method without the need for a long wait. Although the inference time of RadGraph F1 is about 5x longer compared to the proof-of-concept model we present based on the COMET architecture, the scores our model outputs still show a high correlation with RadGraph F1, making our model a faster alternative.

To ensure the quality of our corpus, we have compared the exact overlap on MeSH labels among source and reference reports (i.e. the number of overlapping tokens). Our analysis of the Top 10% corpus revealed that 80.2% of the rows had overlap in their MeSH labels, with 46.9% having one token overlapping and 33.3% having more than one. Only 19.83% of rows had no exact overlaps in MeSH tokens. Similarly, when we examined the extent of overlap between MeSH labels in the complete corpus (i.e. among all scores), we found that 34.20% of rows had no overlap between their MeSH labels. In contrast, 31.69% of rows had only one overlap, and 34.11% had more than one overlap between their MeSH labels. We, therefore, see that the scores in the Top 10% corpus reflect the contents of the reports well. The complete chart of scores can be found in Figure 6.

4.2 Trained Model Checkpoints

During our experiments with different clustering and similarity score methods, we have generated many parallel corpora and also already trained several models to benchmark their performance. Out of all models, we have decided to focus on a couple of best-performing checkpoints (based on the highest Kendall τ value while training) in our evaluation (see Table 7). We used our two corpora (best match and top 10% as described in subsection 3.2.3) and combined them each once with the



(a) **Complete Corpus**: Rows with zero overlap form the largest group

(b) **Top 10% Corpus**: Most rows have one overlapping token.

Figure 6: Visualization of the overlap of MeSH tokens in the two corpora

XLM-RoBERTa [Conneau et al., 2020] encoder layer and once with the medical-specific BioClinical BERT [Alsentzer et al., 2019]. Also, as already discussed, we trained the models on two scores: Once on the Radgraph F1 score, and once on the RadCliQ metric score to compare how they differ in correlation performance.

It is important to notice, that the **RadCliQ score is a measure of how many errors a report will contain** [**Yu et al., 2022**] (i.e. **lower is better**) and **RadGraph F1 is a measure of graph similarity** (i.e. **higher is better**). Our model checkpoints will behave accordingly when giving their predicted scores.

We trained the following checkpoints (see also Table 7):

Match XLM-R RadCliQ Based on the Best Match corpus, with XLM-RoBERTa as the encoder layer and RadCliQ as the training score. A lower score indicates a better report. The Scores are unbounded but typically fell within -3.5 and +0.5 in our tests.

Match Clinic RadCliQ Based on the Best Match corpus, with BioClinical BERT as the encoder layer and RadCliQ as the training score. A lower score indicates a better report. The Scores are unbounded but typically fell within -3.5 and +0.5 in our tests.

Top Clinic RadCliQ Based on the Top 10% corpus, with BioClinical BERT as the encoder layer and RadCliQ as the training score. A lower score indicates a better report. The Scores are unbounded but typically fell within -3.0 and +1.5 in our tests.

Top Clinic RadGraph Based on the Top 10% corpus, with BioClinical BERT as the encoder layer and RadGraph F1 as the training score. A higher score indicates a better report. The Scores are unbounded but typically fell within -0.2 and +1.5 in our test.

Checkpoint name	Encoder	Corpus	max($Kendall_\tau$)
match_xlm-r_radcliq	XLM-RoBERTa	best match (RadCliQ-trained)	0.696 (Epoch 3)
match_clinic_radcliq	BioClinical BERT	best match (RadCliQ-trained)	0.714 (Epoch 10)
top_clinic_radcliq	BioClinical BERT	top 10% (RadCliQ-trained)	0.830 (Epoch 24)
top_clinic_radgraph	BioClinical BERT	top 10% (RadGraph F1-trained)	0.714 (Epoch 18)

Table 7: Best performing model Checkpoints used for evaluation

4.3 Evaluation

We evaluated the performance of our model metric using the test set of our parallel corpus (as detailed in subsection 3.2.4) and a set of generated reports produced by the R2Gen (Radiology Report Generation Chen et al. [2020]) and M2Tr (M^2: Meshed-Memory Transformer by Cornia et al. [2020]) models on the IU X-Ray dataset.

Both of those mentioned models have been recently developed and are using a Transformer-based architecture. R2Gen has been specifically developed for Radiology Report generation, whereas M2Tr is a general image captioning model. They have both achieved good results either on IU X-Ray (for R2Gen) or the more general image captioning COCO dataset (Common Objects in Context by Lin et al. [2014] for M2Tr) [Chen et al., 2020; Cornia et al., 2020].

Unfortunately for the duration of this thesis, we did not get the chance to talk to field professionals and obtain a ranking of reports to compare to our scores (as described in section 2.5) for assessing the correlation of a proposed metric with human judgment. We therefore are evaluating the performance of our metric by examining its correlation with established metric scores claming to have high correlation with human judgment. The Pearson correlation coefficient [Pearson, 1896] and Spearman's rank correlation coefficient [Spearman, 1910] were used as evaluation measures. Both assess the strength and direction of the relationship between two variables. Unlike Pearson correlation, which measures the linear relationship between two variables, Spearman correlation can capture non-linear relationships as well. Earlier works using and comparing the two correlation measures suggest

Spearman's corelation to be more stable and less variable than Pearson's [de Winter et al., 2016; Halawi et al., 2012]. The coefficient values are in the range $[-1, 1]$, with a value of -1 being a perfectly negative correlation and +1 indicating a perfectly positive correlation. If the correlation coefficient is 0, there is no correlation at all.

To provide a comprehensive comparison, we calculated the traditional BLEU-2 Score, as well as the newer scores BERTScore [Zhang et al., 2020a] and S_emb Score (CheXbert, Smit et al. [2020]). Yu et al. [2022] establish the hierarchy of overlap with human (i.e. radiologist) judgement to be as follows (from high to low):

1. RadGraph F1
2. BERTScore
3. BLEU-2
4. S_emb (CheXbert)

Because of that, we were particularly interested in seeing how well our model would perform in relation to RadGraph F1 and BERTscore metrics. Given that our model was trained on RadCliQ, which is a combination of RadGraph F1 and BLEU-2, we would anticipate a high correlation with these two metrics. For this reason, we also evaluate the model checkpoint trained directly on RadGraph F1 instead of RadCliQ to explore the performance difference.

4.3.1 Evaluating on the test corpus

The performance evaluation on the test dataset of the parallel corpus was conducted by splitting the report and reference report column into two text files, `src` and `hyp1`, respectively, for inference. Additionally, the precalculated BLEU-2, BERTscore, and S_emb scores were extracted from the parallel corpus into a separate dataframe.

We then performed the inference on the two files using the different model checkpoints we listed in Table 7 to obtain the predicted "RadEval" scores.

Note: Our RadCliQ-trained checkpoints predict the number of errors in the hypothesis report (i.e. lower is better) whereas our RadGraph-trained checkpoints predict the similarity (i.e. higher is better). The RadCliQ-trained checkpoints will therefore have a negative correlation with the other metrics (BLEU-2, BERTScore, S_emb, and the RadGraph F1-trained checkpoint's predictions). To ensure consistency when comparing the two types of checkpoints, we multiply the RadCliQ scores by -1, so that all correlation values are positive.

Subsequently, we calculated the two correlation values (Pearson and Spearman) between our RadEval Score and the other metrics' scores for the different checkpoints.

For easier reading, we normalized the values by representing the correlation scores as percentages in the paragraphs going forward. This means 0.1 correlation is represented as 10%.

Furthermore, when referring to correlation scores in the paragraphs of this subsection, we are referencing Spearman correlation if not explicitly marked otherwise.

A detailed summary of the scores (including Pearson) can be found in Table 8. We have not included the scores on which the checkpoints were trained when determining the highest and second-highest scores.

Model	**BLEU-2**	**BERTscore**	**S_emb**	**RadGraph F1**	**RadCliQ**
Pearson Correlation					
Match XLM-R RadCliQ	**90.23%**	74.07%	31.01%	*84.82%*	97.37%
Match Clinic RadCliQ	**91.36%**	74.84%	31.03%	*84.75%*	98.08%
Top Clinic RadCliQ	**91.16%**	74.11%	30.75%	*84.23%*	97.71%
Top Clinic RadGraph	61.52%	*62.29%*	29.07%	95.41%	**83.80%**
Spearman Correlation					
Match XLM-R RadCliQ	**86.26%**	66.98%	27.75%	*71.38%*	95.37%
Match Clinic RadCliQ	**87.99%**	67.80%	27.80%	*71.05%*	96.52%
Top Clinic RadCliQ	**88.76%**	67.03%	27.45%	*67.22%*	95.51%
Top Clinic RadGraph	41.35%	*48.86%*	24.45%	87.92%	**67.57%**

Table 8: Correlations between the RadEval score of our model checkpoints and the other metrics based on the test dataset of our parallel corpus. The **highest correlation (both negative and positive) is marked in bold** and *the second highest (both negative and positive) in italics*. The score on which the specific model checkpoint was trained is printed in light grey.

After completing the evaluation process and looking at the Spearman correlation scores, we found that all RadCliQ-trained models (Match Clinic RadCliQ, Match XLM-R RadCliQ and Top Clinic RadCliQ) exhibited a high correlation of over 85% with the BLEU-2 score, which was according to our anticipation as described above. Additionally, these model checkpoints showed the second-highest correlation of approximately 69% with the RadGraph F1 score, which was also in line with our initial expectations.

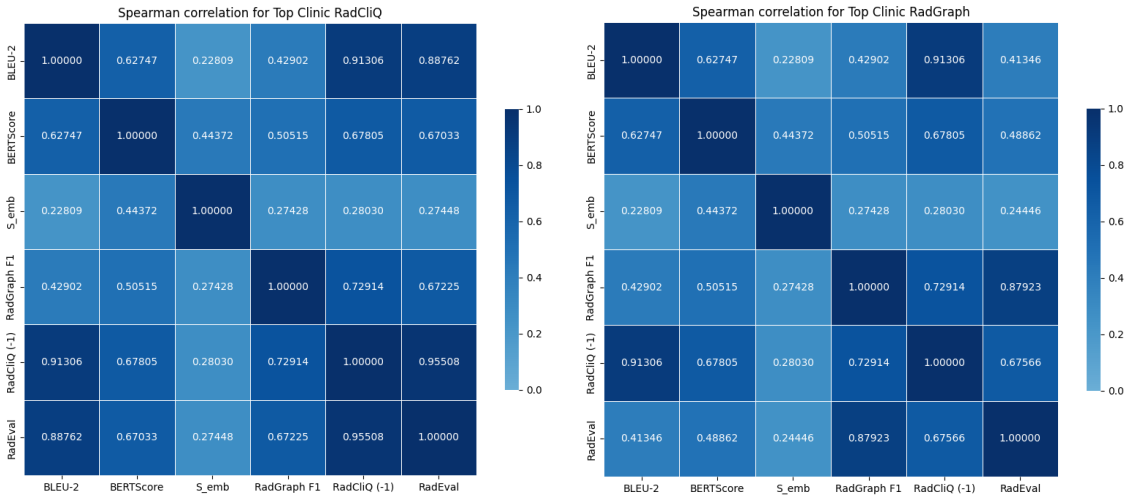
Interestingly, we found that our RadCliQ-trained models also displayed a reasonably high correlation of approximately 67% with the BERTscore metric.

The RadGraph F1-trained checkpoint (Top Clinic RadGraph) on the other hand showed the highest correlation with the RadCliQ score at 67.57% and the second highest correlation with BERTscore at 48.86%, with BLEU-2 following at 41.35%.

It is worth noting that none of our model checkpoints exhibited a high correlation with the S_emb score, with correlations ranging between 24% and 28%.

Even though the correlation with BLEU-2 for the RadGraph F1-trained checkpoint was much lower compared to the RadCliQ-trained checkpoints (-45 percentage points), the RadGraph F1-trained checkpoint also showed a lower correlation with BERTscore (-19 percentage points) and S_emb score (-3 percentage points) at the same time, albeit less drastic than the drop in BLEU-2 correlation.

A visual representation of the difference in the correlation matrices between the RadCliQ-trained and RadGraph F1-trained checkpoints is shown in Figure 7.



(a) **Top Clinic RadCliQ**: Most correlation with BLEU and second most with BERTscore

(b) **Top Clinic RadGraph**: Better correlation with BERTScore than BLEU, highest with RadCliQ

Figure 7: Comparison of Spearman correlations for the Model Checkpoints trained on RadCliQ (left) and RadGraph F1 (right) scores. Both based on the Top 10% corpus with BioClinical BERT

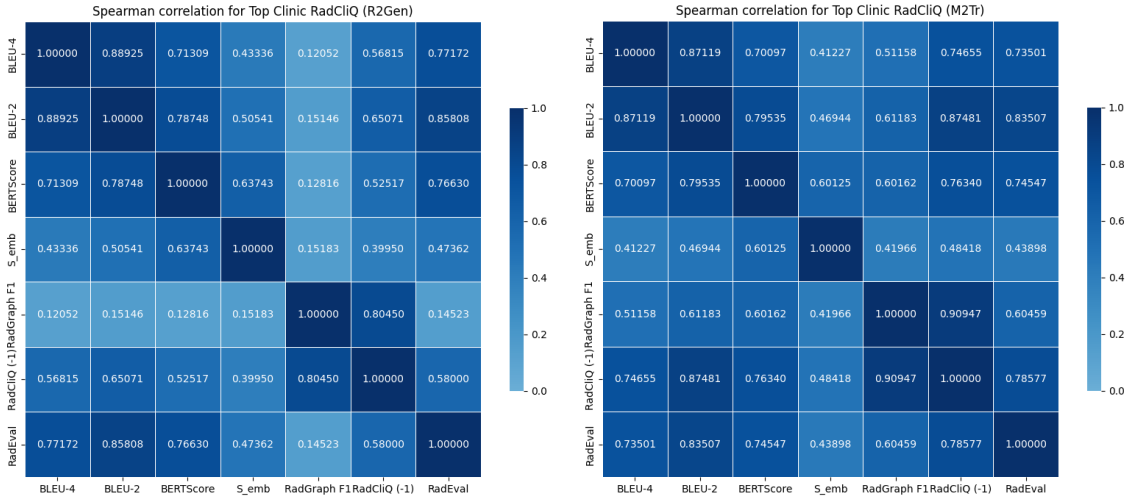
4.3.2 Evaluating on generated reports

We used the generated reports for IU X-Ray images created by the two models R2Gen [Chen et al., 2020] and M2Tr [Cornia et al., 2020]. For each model, we had

590 reports at our disposal, which contained, on every row in the data, the ground truth report, the predicted report as well as the BLEU-4 score for said prediction. It needs to be noted, that the size of the parallel corpus test data evaluated in subsection 4.3.1 is a lot (13'000 more reports) larger than this set of generated reports.

We followed the same sequence as for the parallel corpus test to do the inference and get the RadEval score. We included the model-provided BLEU-4 score alongside BLEU-2 and the other scores in the detailed Table 9 to maintain comparability between the scores for the generated reports and the scores for the parallel corpus test dataset in Table 8.

As before, we multiply the RadCliQ scores by -1, such that all correlation values are positive and we normalized the values by representing them as percentages in the paragraphs going forward. Furthermore, when referring to correlation scores in the paragraphs of this subsection, we are referencing Spearman correlation if not explicitly marked otherwise.



(a) **R2Gen generated reports** (Top Clinic RadCliQ): Slightly lower correlations in general and particularly low correlation with RadGraph F1.

(b) **M2Tr generated reports** (Top Clinic RadCliQ): Lowest correlation with S_emb score but even correlations in general.

Figure 8: Comparison of Spearman Correlations for the Model Checkpoints trained on RadCliQ (left) and RadGraph F1 (right) scores. Both based on the Top 10% corpus with BioClinical BERT

Looking at the correlation when inference is run on our two model-generated datasets, we can see a much different set of correlations than we had on the parallel corpus test dataset. We still see a high correlation with both BLEU scores for the RadCliQ-trained (Match Clinic RadCliQ, Match XLM-R RadCliQ, and Top Clinic RadCliQ)

Model	BLEU-4	BLEU-2	BERTscore	S_emb	RadGraph F1	RadCliQ
Pearson Correlation						
R2Gen reports						
Match XLM-R RadCliQ	84.57%	**92.89%**	*86.19%*	55.30%	22.80%	73.22%
Match Clinic RadCliQ	87.03%	**93.95%**	*87.24%*	54.51%	18.64%	70.98%
Top Clinic RadCliQ	**92.08%**	*91.96%*	81.49%	48.32%	17.40%	68.88%
Top Clinic RadGraph	**83.76%**	*83.57%*	78.26%	45.63%	12.57%	60.35%
M2Tr reports						
Match XLM-R RadCliQ	81.21%	*90.49%*	83.19%	51.29%	**90.85%**	97.51%
Match Clinic RadCliQ	83.64%	**92.05%**	*85.55%*	49.30%	80.42%	92.34%
Top Clinic RadCliQ	*88.62%*	**89.11%**	78.76%	44.68%	72.97%	86.60%
Top Clinic RadGraph	76.48%	*79.45%*	75.91%	40.90%	77.47%	**84.30%**
Spearman Correlation						
R2Gen reports						
Match XLM-R RadCliQ	78.08%	**86.85%**	*79.54%*	52.69%	24.74%	66.37%
Match Clinic RadCliQ	81.84%	**88.94%**	*80.95%*	51.95%	19.36%	63.03%
Top Clinic RadCliQ	77.17%	**85.81%**	*76.63%*	47.36%	14.52%	58.00%
Top Clinic RadGraph	**61.37%**	*66.33%*	65.09%	39.09%	5.17%	40.96%
M2Tr reports						
Match XLM-R RadCliQ	74.71%	*84.88%*	76.58%	47.66%	**85.90%**	95.28%
Match Clinic RadCliQ	79.72%	**87.60%**	*79.83%*	45.54%	71.73%	87.70%
Top Clinic RadCliQ	73.50%	**83.51%**	*74.55%*	43.90%	60.46%	78.58%
Top Clinic RadGraph	58.12%	*64.29%*	64.01%	33.64%	65.60%	**71.76%**

Table 9: Correlation between the RadEval score of our model checkpoints and the other metrics based on the generated reports by M2Tr and R2Gen. The **highest correlation is marked in bold** and *the second highest in italics*. The score on which the specific model checkpoint was trained is printed in light grey.

checkpoints, with correlations ranging from 73% to 86% on both R2Gen and M2Tr reports. For the same checkpoints however, we see a strong difference in correlation towards the other metrics when looking at the two generation models: For R2Gen we see a really low correlation with RadGraph F1 with values ranging from 14% to 25%. For M2Tr on the other hand we note between 60% and 85%. It is also worth noting, that R2Gen has the lowest correlation (5.17%) with RadGraph F1 for the model checkpoint (Top Clinic RadGraph), which was trained on the said score.

Correlation with BERTscore and S_emb scores are generally higher than the values we have seen when evaluating the parallel corpus test dataset with values ranging from 33% up to 53% for S_emb score and 64% to 80% for BERTscore.

When looking at the RadGraph F1-trained checkpoint for both generation models, we no longer see the consistently bad correlation with BLEU-2 and BERTscore but instead, they are aligning more closely with the correlation values of the other RadCliQ-trained model checkpoints, being at most 19 percentage points away from the highest value for BLEU-2 of the RadCliQ-trained checkpoints and at most 11 percentage points for BERTscore. Compared to the maximal distance in scores for the parallel corpus test dataset of 45 percentage points for BLEU-2 and 19 percentage points for BERTscore.

For the S_emb score on the other hand we see a more steep decline in scores looking at the generated reports' evaluation, with a maximal drop of 10 percentage points, compared to the parallel corpus test dataset, where the maximal drop is of just 3 percentage points.

We assembled the mean, standard deviation, minimum, maximum, and quartile values for the spearman correlation values of each metric over all checkpoints and both setups (parallel corpus test dataset and generated reports dataset) in Table 10.

4.4 Interpreting the scores

Upon analyzing the Spearman correlation values, particularly for RadGraph F1 and BERTscore metrics, we observed that among the four metric checkpoints, those trained on RadCliQ (Match XLM-R RadCliQ, Match Clinic RadCliQ, and Top Clinic RadCliQ) performed reasonably well on both test dataset of the parallel corpus and the generated reports datasets (R2Gen and M2Tr). The RadCliQ-trained checkpoints showed a more consistent correlation with the other metrics when scored on the test dataset of the parallel corpus.

	BLEU-2	BERTScore	S_emb	RadGraph F1	RadCliQ (-1)	BLEU-4
count	12	12	12	9	3	8
mean	0.794	0.707	0.391	0.540	0.601	0.731
std	0.146	0.092	0.104	0.268	0.167	0.087
min	0.413	0.489	0.244	0.145	0.410	0.581
25%	0.792	0.665	0.278	0.247	0.543	0.705
50%	0.860	0.712	0.415	0.672	0.676	0.759
75%	0.877	0.774	0.474	0.714	0.697	0.785
max	0.889	0.809	0.527	0.859	0.718	0.818

Table 10: Summary statistics for different metrics. The statistics presented include count, mean, standard deviation, minimum, maximum, and quartile values. The score on which the specific model checkpoint was trained has been removed.

Among the generated reports datasets, we found that the correlation values for inference on the M2Tr dataset were substantially higher, particularly for RadGraph F1, surpassing the correlation values observed for this checkpoint on the test dataset of the parallel corpus. Conversely, for the R2Gen generated dataset, the correlation with RadGraph was considerably low compared to other metrics.

Furthermore, our RadGraph F1 checkpoint generally demonstrated good performance on the generated reports compared to its performance on the test dataset. This is evident from the higher correlation observed with BERTscore and BLEU-2/4, while the correlation values on target metrics for the RadCliQ-trained checkpoints remained relatively consistent among the two datasets.

We observed that the performance of our metric on the generated reports was slightly better compared to the test set of the parallel corpus, as the former exhibited a smaller difference in correlation values between BERTscore and BLEU-2/4 as well as much higher correlation with S_emb score.

To create a performance score that aligns with the desired correlation pattern (greater correlation with metrics that Yu et al. [2022] has defined as being more closely linked to human judgment), we propose two weighted performance scores for our model checkpoints, denoted as S_{RadCliQ} and S_{RadGraph}. These scores are calculated by excluding the score that each checkpoint was trained on and using a weighted sum of the remaining metrics. The weights are set to start at 2.5 (RadGraph F1) and then decrease in 0.5 steps until 1.0 (S_emb) for RadCliQ-trained

checkpoints. In the case of $\mathrm{S}_{\mathrm{RadGraph}}$, we set the weight for the RadCliQ correlation to 2.0, which is the average of w_{BLEU_2} and $w_{RadGraph_{F1}}$, to avoid double-counting $BLEU_2$ (since RadCliQ is a combination of $BLEU_2$ and $RadGraph_{F1}$).

$$\mathrm{S}_{\mathrm{RadCliQ}} = \frac{2.5 \times \mathrm{RadGraph}_{\mathrm{F1}} + 2.0 \times \mathrm{BERTScore} + 1.5 \times \mathrm{BLEU}_2 + 1.0 \times \mathrm{S\_emb}}{\max(\mathrm{RadGraph}_{\mathrm{F1}}, \mathrm{BERTScore}, \mathrm{BLEU}_2, \mathrm{S\_emb})}$$

$$\mathrm{S}_{\mathrm{RadGraph}} = \frac{2.0 \times \mathrm{RadCliQ}^{-1} + 2.0 \times \mathrm{BERTScore} + 1.5 \times \mathrm{BLEU}_2 + 1.0 \times \mathrm{S\_emb}}{\max(\mathrm{RadCliQ}^{-1}, \mathrm{BERTScore}, \mathrm{BLEU}_2, \mathrm{S\_emb})}$$

where $BLEU_2$, $BERTScore$, S_emb, $RadGraph_{F1}$ and $RadCliQ^{(-1)}$ are the correlation values for said model, and the numbers are the weights assigned to each metric. The max function in the denominator is used to normalize the weighted sum of the correlations.

It should be noted that the weighted score used in this analysis was proposed by us and it has not been validated whether this score is meaningful or effective for evaluating the performance of the model checkpoints. Therefore, the results of this analysis should be interpreted with caution, and further research is needed to determine the validity of this approach.

experiment	dataset	$\mathrm{S}_{\mathrm{RadCliQ}}$	$\mathrm{S}_{\mathrm{RadGraph}}$
Match XLM-R RadCliQ	M2Tr	6.320	*n/a*
Top Clinic RadGraph	M2Tr	*n/a*	5.597
Match Clinic RadCliQ	M2Tr	5.889	*n/a*
Top Clinic RadCliQ	M2Tr	5.621	*n/a*
Match XLM-R RadCliQ	Parallel Corpus	5.444	*n/a*
Match Clinic RadCliQ	Parallel Corpus	5.376	*n/a*
Top Clinic RadGraph	R2Gen	*n/a*	5.287
Top Clinic RadCliQ	Parallel Corpus	5.213	*n/a*
Top Clinic RadGraph	Parallel Corpus	*n/a*	4.726
Match XLM-R RadCliQ	R2Gen	4.650	*n/a*
Match Clinic RadCliQ	R2Gen	4.449	*n/a*
Top Clinic RadCliQ	R2Gen	4.261	*n/a*

Table 11: Weighted scores for all checkpoints on the different datasets. For the RadCliQ-trained checkpoints, we calculate $S_{RadCliQ}$ and for the RadGraph-trained checkpoints, we calculate $S_{RadGraph}$

Utilizing the correlation scores, an attempt can be made to establish a performance hierarchy for the model checkpoints. The results indicate that the Match XLM-R RadCliQ checkpoint had the best performance for M2Tr output and the test dataset

of the parallel corpus, as well as the second-best on the R2Gen output, with a mean score of 5.47. Following this is the Match Clinic RadCliQ with a mean score of 5.24 (ranking third on M2Tr, second on the parallel corpus test dataset, and third on R2Gen). The Top Clinic RadCliQ had the lowest mean score of 5.032 and performed worst on the two generated datasets and came in third on the parallel corpus test dataset. Top Clinic RadGraph came in best on R2Gen data, second on M2Tr data, and last on the parallel corpus data with a mean score of 5.20 (see Figure 9).

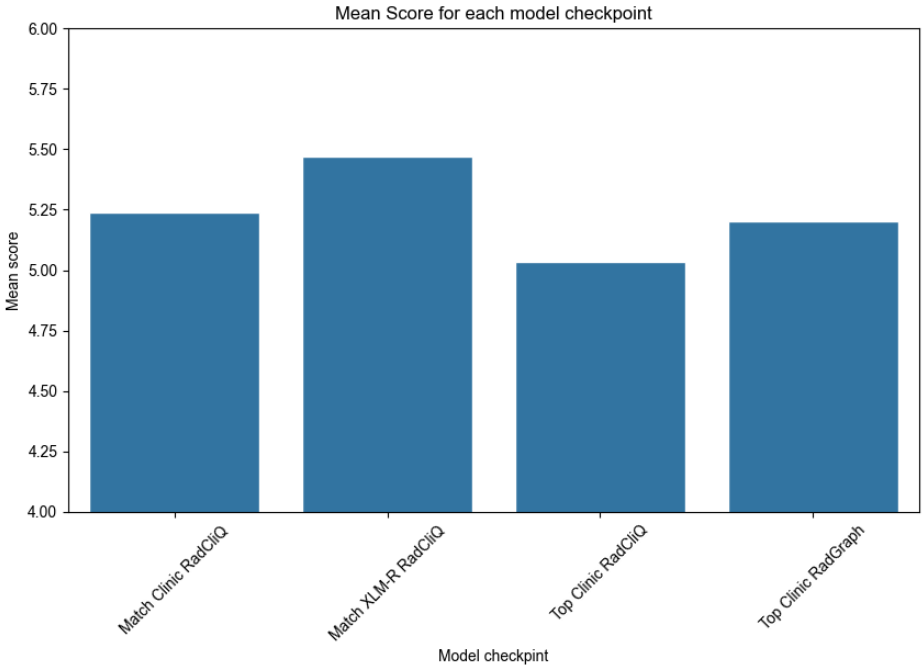


Figure 9: Mean of the weighted scores over all three datasets

5 Conclusion

The aim of our thesis was to investigate the feasibility of developing a new radiology-specific evaluation metric. To achieve this goal, we pursued two research questions.

The first question focused on adapting and optimizing an existing, successful metric from a non-radiology domain to fit our problem of evaluating generated radiology reports. We used the architecture of the popular machine-translation evaluation framework COMET by Unbabel AI to train our own metric on radiology data and explored several optimization techniques such as encoder layer replacement using a BioClinical BERT encoder and training on two types of scores (error count / RadCliQ and graph similarity / RadGraph F1).

Additionally, we created several similarity-scored parallel corpora, which we used as training data for our model checkpoints. Our correlation analysis shows that the checkpoints generally correlate well with the other metrics we used to compare. However, due to the absence of a human reference to evaluate the performance of our metric in real-life conditions, we cannot claim with certainty that our metric performs better than other popular metrics in the field. Nonetheless, our trained metric showed moderate to high correlation with BERTscore and S_emb score, which suggests that it has the potential to be an effective radiology-specific evaluation metric.

The second research question focused on whether the integration of a radiology-specific knowledge graph could improve the evaluation quality. We answered this question in two ways. Firstly, we trained one of our metric model checkpoints solely on the RadGraph F1 score and compared its performance with the performance of the other metric model checkpoints, which were trained on RadCliQ. Based on our interpretation of the score hierarchy (see section 4.4), the RadGraph F1-trained checkpoint performed well, particularly for generated reports, suggesting that RadGraph F1 has the potential to be used standalone in an automatic setting. Secondly, we incorporated the RadGraph F1 and RadCliQ scores into the parallel corpora and validated them to demonstrate how filtering based on radiology-aware scores could improve the quality of the corpus.

In sum, we provide the following contributions to the scientific community: two parallel corpora that synthesize a "source-reference" set of reports, four model checkpoints to score generated radiology reports and compare them to the reference reports, a proposed weighted score performance measure to evaluate the correlation towards human judgement of developed metrics, and code to replicate our experiments and further improve the model's performance.

By providing these resources, we hope to facilitate future research and development in the field of NLG for radiology and the medical domain in general. Moreover, we hope that our proposed evaluation framework can be utilized as a benchmark for evaluating the quality of generated radiology reports by other researchers.

We encourage the scientific community to utilize our resources to advance the state-of-the-art in natural language generation for radiology reports and to further improve the quality of generated reports for use in clinical practice.

5.1 Further Research

In future studies, it may be possible to enhance the clustering approach (subsection 3.2.1) by utilizing RadGraph during data preparation to extract anatomy labels from the findings and cluster reports based on these labels. This could potentially provide more comprehensive information about the reports than merely utilizing the entire unprocessed MeSH column.

During our study in section 3.3, we did not conduct any hyper-parameter optimization except for replacing the encoder and increasing the maximum epochs, leaving it as a potential future research avenue to further enhance model performance.

As our evaluation framework was originally designed for machine translation, additional research could explore the use of model checkpoints to score translations of radiology reports. Furthermore, there are two other model architectures available when using COMET (see section 3.1). Future researchers could be looking into the potential of training a ranking instead of a referenceless metric on our proposed parallel corpus.

Currently, we only validate our metrics by measuring correlation using the newly proposed weighted score performance measure. The next step in gaining a more thorough understanding of model performance would be to include human evaluation, as described in section 2.5. Moreover, our weighted performance score requires further validation and testing.

If at a future point, a big enough parallel corpus of machine- and human-generated reports is being published, this corpus could be scored using the methods described in subsection 3.2.3 and could potentially replace the parallel corpus we presented in our work to improve the training data for the model.

We make our parallel corpora and code (see section 5.2) available to future researchers for use in further works and for the possible training of evaluation models on architectures other than COMET.

5.1.1 Outlook

Following the completion of the thesis, we intend to further investigate the subject matter and subsequently release a paper disclosing our findings later this year. Part of this paper we will include a small study of human judgement correlation in collaboration with radiologists from Kyoto University (京都大学) as part of the inter-university academic exchange agreement with the University of Zurich.

5.2 Additional Material

The code for each step outlined in chapter 3 and chapter 4 has been made publicly available on GitHub at github.com/amoscalamida/rad-eval.

The larger files (such as model checkpoints, parallel corpora, and score tables) have been uploaded to SwitchDrive[1] at drive.switch.ch/index.php/s/RW1362mGhhi8VY8 and can be downloaded under the same license terms as the original IU X-Ray data.

The contents of both repositories may be updated in the future to reflect new developments.

[1] SwitchDrive is a cloud storage and file-sharing service offered by SWITCH, a Swiss foundation that provides IT infrastructure services for higher education institutions

Glossary

Evaluation The process of measuring the effectiveness or quality of a system, metric, or algorithm. In natural language processing, evaluation is commonly used to compare the quality of machine-generated text with human-written text.

Machine learning A type of artificial intelligence that involves teaching machines to learn from data and make predictions. Machine learning algorithms are commonly used in natural language processing to optimize models and make predictions.

Precision The proportion of relevant items among the total number of items retrieved or generated.

Recall The proportion of relevant items retrieved or generated among the total number of relevant items in the corpus.

F1-score A measure of the performance of a system that combines both precision and recall into a single score. F1-score is a balanced measure that takes into account both the number of correct positive predictions and the number of false negative predictions.

Parallel corpus A collection of texts, usually in multiple languages aligned for direct comparison. A parallel corpus is often used in natural language processing to train and evaluate machine translation and other language-related models.

Word embedding A technique used to represent words as vectors in a high-dimensional space. Word embeddings are often used to improve the performance of machine learning models by capturing the semantic relationships between words.

Encoder Layer A component of neural network models, that is responsible for processing input text and producing contextualized word representations. The encoder layer typically consists of several stacked layers of self-attention and feed-forward neural networks, and is trained on large corpora of text using unsupervised learning techniques.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-1909.

Mohammad Alsharid, Harshita Sharma, Lior Drukker, Pierre Chatelain, Aris T. Papageorghiou, and J. Alison Noble. Captioning ultrasound images automatically. In Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pages 338–346, Cham, 2019. Springer International Publishing. ISBN 978-3-030-32251-9.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 382–398, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46454-1.

Satanjeev Banerjee and Alon Lavie. 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, June 2005. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, March 2021. doi: 10.1145/3442188.3445922.

Kathrin Blagec, Georg Dorffner, Milad Moradi, Simon Ott, and Matthias Samwald.

A global analysis of metrics used for measuring performance in natural language processing. pages 52–63, May 2022. doi: 10.18653/v1/2022.nlppower-1.6.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

Tadeusz Caliński and Jack C. Harabasz. A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27, 1974. doi: 10.1080/03610927408827101.

Chris Callison-Burch, Cameron Fordyce, Philipp Koehn, Christof Monz, and Josh Schroeder. (Meta-) evaluation of machine translation. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 136–158, Prague, Czech Republic, June 2007. Association for Computational Linguistics.

Zhihong Chen, Yan Song, Tsung-Hui Chang, and Xiang Wan. 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, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.112.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-Memory Transformer for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Joost C. F. de Winter, Samuel D. Gosling, and Jeff Potter. Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A

tutorial using simulations and empirical data. *Psychological Methods*, 21(3): 273–290, September 2016. doi: 10.1037/met0000079.

Jean-Benoit Delbrouck, Pierre Chambon, Christian Bluethgen, Emily Tsai, Omar Almusa, and Curtis Langlotz. Improving the factual correctness of radiology report generation with semantic rewards. pages 4348–4360, December 2022.

Dina Demner-Fushman, Marc D. Kohli, Marc B. Rosenman, Sonya E. Shooshan, Laritza Rodriguez, Sameer Antani, George R. Thoma, and Clement J. McDonald. Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2): 304–310, March 2016. doi: 10.1093/jamia/ocv080.

Chenhe Dong, Yinghui Li, Haifan Gong, Miaoxin Chen, Junxin Li, Ying Shen, and Min Yang. A survey of natural language generation. *ACM Comput. Surv.*, 55 (8), dec 2022. ISSN 0360-0300. doi: 10.1145/3554727.

Roger Evans, Paul Piwek, and Lynne Cahill. What is NLG? In *Proceedings of the International Natural Language Generation Conference*, pages 144–151, Harriman, New York, USA, July 2002. Association for Computational Linguistics.

Albert Gatt and Emiel Krahmer. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Int. Res.*, 61(1): 65–170, January 2018. ISSN 1076-9757.

Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet. *Circulation*, 101(23):e215–e220, June 2000. doi: 10.1161/01.cir.101.23.e215.

Guy Halawi, Gideon Dror, Evgeniy Gabrilovich, and Yehuda Koren. Large-scale learning of word relatedness with constraints. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, page 1406–1414, New York, NY, USA, August 2012. Association for Computing Machinery. ISBN 9781450314626. doi: 10.1145/2339530.2339751.

Xin Huang, Fengqi Yan, Wei Xu, and Maozhen Li. Multi-attention and incorporating background information model for chest x-ray image report generation. *IEEE Access*, 7:154808–154817, 2019. doi: 10.1109/ACCESS.2019.2947134.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven QH Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew P Lungren, Andrew Y Ng, Curtis P Langlotz, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. *PhysioNet*, 2021. doi: 10.13026/hm87-5p47.

Kevin Knight, Martha Palmer, Daniel Marcu, Kira Griffitt, Laura Baranescu, Claire Bonial, Madalina Georgescu, Ulf Hermjakob, and Nathan Schneider. Abstract meaning representation (AMR) annotation release 1.0. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria, August 2014. Linguistic Data Consortium. doi: 10.35111/0YNC-7404.

Curtis P Langlotz. *Radiology report: a guide to thoughtful communication for radiologists and other medical professionals*. Independent Publishing Platform, San Bernardino, CA, March 2015. ISBN 9781515174080.

Christoph Leiter, Piyawat Lertvittayakumjorn, Marina Fomicheva, Wei Zhao, Yang Gao, and Steffen Eger. Towards explainable evaluation metrics for natural language generation. March 2022.

Christy Y. Li, Xiaodan Liang, Zhiting Hu, and Eric P. Xing. Hybrid retrieval-generation reinforced agent for medical image report generation. page 1537–1547, 2018.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Computer Vision – ECCV 2014*, pages 740–755. Springer International Publishing, 2014. doi: 10.1007/978-3-319-10602-1_48.

Lei Liu and Min Zhu. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, December 2022. doi: 10.1093/llc/fqac089.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of SPIDEr. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, October 2017. doi: 10.1109/iccv.2017.100.

Pablo Messina, Pablo Pino, Denis Parra, Alvaro Soto, Cecilia Besa, Sergio Uribe, Marcelo Andía, Cristian Tejos, Claudia Prieto, and Daniel Capurro. A survey on deep learning and explainability for automatic report generation from medical images. *ACM Comput. Surv.*, 54(10s), sep 2022. ISSN 0360-0300. doi: 10.1145/3522747. URL https://doi.org/10.1145/3522747.

National Library of Medicine. Medical subject headings files 2023, 2023. URL https://www.nlm.nih.gov/mesh/meshhome.html. [Online; accessed March 10th 2023].

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. Why we need new evaluation metrics for NLG. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1238.

OpenAI. Introducing chatgpt, 2022. URL http://openai.com/blog/chatgpt. [Online; accessed March 1st 2023].

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135.

Karl Pearson. VII. mathematical contributions to the theory of evolution.—III. regression, heredity, and panmixia. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 187:253–318, December 1896. doi: 10.1098/rsta.1896.0007.

PhysioNet. Physionet credentialed health data use agreement 1.5.0. URL https://physionet.org/content/radgraph/view-dua/1.0.0/. [Online; accessed March 20th 2023].

XiPeng Qiu, TianXiang Sun, YiGe Xu, YunFan Shao, Ning Dai, and XuanJing Huang. Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, 63(10):1872–1897, September 2020. doi: 10.1007/s11431-020-1647-3.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. COMET: A neural framework for MT evaluation. pages 2685–2702, November 2020. doi: 10.18653/v1/2020.emnlp-main.213.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2), jan 2022. ISSN 0360-0300. doi: 10.1145/3485766.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. pages 7881–7892, July 2020. doi: 10.18653/v1/2020.acl-main.704.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Ng, and Matthew Lungren. 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, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.117.

Charles Spearman. Correlation calculated from faulty data. *British Journal of Psychology, 1904-1920*, 3(3):271–295, October 1910. doi: 10.1111/j.2044-8295.1910.tb00206.x.

Unbabel AI. Comet: High-quality machine translation evaluation, 2020. URL https://unbabel.github.io/COMET/. [Online; accessed December 10th 2022].

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation, 2022.

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

Yixiao Zhang, Xiaosong Wang, Ziyue Xu, Qihang Yu, Alan Yuille, and Daguang Xu. When radiology report generation meets knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12910–12917, April 2020b. doi: 10.1609/aaai.v34i07.6989.

A Tables

Cluster Amount	Silhouette Score	Elbow Score	Calinski Harabasz Score
2	0.1123	2615.6462	403.1788
3	0.1565	2397.0003	375.7110
4	0.2083	2181.2283	387.7774
5	0.2553	2023.7439	379.7914
6	0.2732	1916.1540	359.1262
7	0.3217	1775.6006	367.8817
8	0.3324	1714.6355	343.7689
9	0.3498	1642.0339	332.8517
10	0.3576	1584.2830	320.3707
11	0.3687	1536.7594	307.7021
12	0.3828	1479.0708	302.6347
13	0.3822	1464.7734	282.8111
14	0.3835	1434.4429	272.0363
15	0.3955	1389.6302	268.5154
16	0.4055	1348.2333	265.1993
17	0.4101	1331.8275	254.2322
18	0.4161	1305.8520	247.9447
19	0.4333	1279.1687	242.9247
20	0.4318	1259.8251	236.3508
21	0.4506	1230.3285	233.9220
22	0.4401	1217.2425	226.8507
23	0.4551	1194.3380	223.5873
24	0.4636	1185.5286	216.4887
25	0.4658	1165.7390	213.3289
26	0.4709	1140.7286	212.1999
27	0.4717	1137.6182	204.8926
28	0.4610	1118.5256	202.7565
29	0.4687	1120.6646	194.8534
30	0.4773	1092.9619	195.8080
31	0.4918	1069.8641	195.7486
32	0.4849	1081.8287	186.0754
33	0.4916	1064.2109	184.9423
34	0.4960	1047.2711	183.8442
35	0.5046	1033.6481	182.0471
36	0.5079	1014.7761	181.8789
37	0.5071	1010.3364	177.9641
38	0.5124	999.2764	176.0303
39	0.5067	987.6090	174.4223

Table 12: Cluster Scores for different amounts of clusters on the Impression Column of the Reports (see subsection 3.2.1 for details and Figure 4 for a visualization).

Cluster Amount	Silhouette Score	Elbow Score	Calinski Harabasz Score
2	0.4013	1885.8854	1502.0259
3	0.4004	1774.6551	901.9194
4	0.4015	1713.2610	662.2957
5	0.4024	1653.3468	544.6356
6	0.4083	1604.4589	469.0691
7	0.4120	1555.5883	420.4176
8	0.4159	1513.4368	383.4765
9	0.4200	1479.0646	352.8658
10	0.4241	1443.4558	330.3829
11	0.4256	1416.5990	309.1715
12	0.4321	1383.9273	294.7217
13	0.4299	1368.2138	276.3505
14	0.4389	1329.9459	269.6799
15	0.4415	1311.4290	257.2134
16	0.4436	1286.4803	248.9243
17	0.4434	1266.0982	240.3790
18	0.4504	1238.4890	235.5478
19	0.4493	1223.7864	227.2723
20	0.4513	1207.1610	220.6048
21	0.4546	1188.5917	215.3645
22	0.4505	1182.4505	206.9285
23	0.4570	1169.2967	201.3721
24	0.4602	1143.2380	200.2180
25	0.4602	1134.1673	194.4508
26	0.4614	1119.2459	190.8628
27	0.4623	1104.2314	187.6852
28	0.4641	1092.9885	183.7933
29	0.4658	1084.5539	179.4685
30	0.4668	1066.9198	177.9679
31	0.4645	1065.2668	172.4206
32	0.4668	1054.9632	169.4743
33	0.4713	1034.5247	169.4028
34	0.4713	1028.4956	165.7665
35	0.4723	1018.2670	163.4295
36	0.4755	1001.0423	163.0592
37	0.4750	997.8114	159.2904
38	0.4726	995.7692	155.4380
39	0.4708	986.3996	153.5598

Table 13: Cluster Scores for different amounts of clusters on the MeSH Column of the Reports (see subsection 3.2.1 for details and Figure 4 for a visualization).

B Code Extracts / Configuration

The Lightning Configuration parameters listed here have been used to train our model on the COMET Architecture by Unbabel AI [2020] (see section 3.3). They provide information on various variables of the training process, including the optimizer settings, learning rate and data augmentation techniques.

```
referenceless_regression_metric:
  class_path: comet.models.ReferencelessRegression
  init_args:
    nr_frozen_epochs: 0.3
    keep_embeddings_frozen: True
    optimizer: AdamW
    encoder_learning_rate: 5.0e-06
    learning_rate: 1.5e-05
    layerwise_decay: 0.95
    encoder_model: BERT
    pretrained_model: emilyalsentzer/Bio_ClinicalBERT
    pool: avg
    layer: mix
    #loss: mse
    dropout: 0.1
    batch_size: 4
    train_data: "input_train.csv"
    validation_data: "input_val.csv"
    hidden_sizes:
      - 2048
      - 1024
    activations: Tanh

trainer: trainer.yaml
early_stopping: early_stopping.yaml
model_checkpoint: model_checkpoint.yaml
```

Listing B.1: Model Configuration

```
class_path: pytorch_lightning.callbacks.EarlyStopping
init_args:
  monitor: val_kendall
  min_delta: 0.
  patience: 2
  verbose: False
  mode: max
  strict: True
  check_finite: True
  stopping_threshold: null
  divergence_threshold: null
  check_on_train_epoch_end: False
```

Listing B.2: Early Stopping

```
class_path: pytorch_lightning.callbacks.ModelCheckpoint
init_args:
  dirpath: null
  filename: '{epoch}-{step}-{val_kendall:.3f}'
  monitor: val_kendall
  verbose: True
  save_last: False
  save_top_k: 2
  mode: max
  auto_insert_metric_name: True
  save_weights_only: True
  every_n_train_steps: null
  train_time_interval: null
  every_n_epochs: 1
  save_on_train_epoch_end: null
```

Listing B.3: Model Checkpoint Config

```
class_path: pytorch_lightning.trainer.trainer.Trainer
init_args:
  accelerator: gpu
  devices: auto
  accumulate_grad_batches: 4
  amp_backend: native
  auto_lr_find: False
  auto_scale_batch_size: False
  auto_select_gpus: False
  check_val_every_n_epoch: 1
  deterministic: True
  fast_dev_run: False
  gradient_clip_val: 1.0
  gradient_clip_algorithm: norm
  limit_train_batches: 1.0
  limit_val_batches: 1.0
  limit_test_batches: 1.0
  limit_predict_batches: 1.0
  log_every_n_steps: 50
  overfit_batches: 0
  precision: 16
  max_epochs: 20
  min_epochs: 1
  max_steps: -1
  num_nodes: 1
  num_sanity_val_steps: 10
  reload_dataloaders_every_n_epochs: 0
  replace_sampler_ddp: True
  sync_batchnorm: False
  detect_anomaly: False
  track_grad_norm: -1
  val_check_interval: 1.0
  enable_model_summary: True
  move_metrics_to_cpu: True
  multiple_trainloader_mode: max_size_cycle
```

Listing B.4: Trainer Config[0]

[0]Properties with *null* values removed (amp_level, benchmark, default_root_dir, profiler, plugins, min_steps, max_time, tpu_cores)