RRED : A Radiology Report Error Detector based on Deep Learning Framework

Dabin Min[1*], **Kaeun Kim**[1*], **Jong Hyuk Lee**[1], **Yisak Kim**[123], **Chang Min Park**[123]
[1]Department of Radiology, Seoul National University Hospital
[2]Interdisciplinary Program in Bioengineering, Seoul National University Graduate School
[3]Department of Radiology, Seoul National University College of Medicine
{reonaledo, jonghyuklee, yisakk, morphius}@snu.ac.kr
kaeun.kim@uwaterloo.ca

Abstract

Radiology report is an official record of radiologists' interpretation of patients' radiographs and it's a crucial component in the overall medical diagnostic process. However, it can contain various types of errors that can lead to inadequate treatment or delay in diagnosis. To address this problem, we propose a deep learning framework to detect errors in radiology reports. Specifically, our method detects errors between findings and conclusion of chest X-ray reports based on a supervised learning framework. To compensate for the lack of data availability of radiology reports with errors, we develop an error generator to systematically create artificial errors in existing reports. In addition, we introduce a Medical Knowledge-enhancing Pre-training to further utilize the knowledge of abbreviations and key phrases frequently used in the medical domain. We believe that this is the first work to propose a deep learning framework for detecting errors in radiology reports based on a rich contextual and medical understanding. Validation on our radiologist-synthesized dataset, based on MIMIC-CXR, shows 0.80 and 0.95 of the area under precision-recall curve (AUPRC) and the area under the ROC curve (AUROC) respectively, indicating that our framework can effectively detect errors in the real-world radiology reports.

1 Introduction

Radiology report is a document containing official interpretation of patients' radiographs which is used as an important communication tool between radiologists and referring physicians (Wallis and McCoubrie, 2011). The major components of the report include basic demographic information (e.g. patient's name, identifying number), **findings** which explains the image findings along with pertinent clinical information, and **conclusion** (also called impression) which is a list of summary statements of radiographic study conclusion and recommendations for further evaluation and patient management (Wilcox, 2006). Medical treatment decisions are often based on the findings and conclusions of the radiology report (Sistrom and Langlotz, 2005). This explains how the radiologic contribution to inappropriate or delayed diagnosis overall is likely to be substantial (Bruno et al., 2015).

Radiology report errors can be categorized and defined in different ways, mostly based on their causes. Kim and Mansfield classified the errors in 12 types which include errors caused by underreading, location of the lesion, and faulty reasoning. Pinto et al. claim that radiology report errors can be classified based on 4 main reasons why radiologists are sued which include observer errors, errors in interpretation, and failure to suggest proper recommendations. Sangwaiya et al. has analyzed errors on location and size discrepancy of lesions in radiology reports. Combining these works, we conclude that the errors that contribute most to inappropriate or delayed diagnosis are radiologists failing to identify and interpret abnormalities, and discrepancies in size or location of the lesions reported.

Although there have been sufficient discussions in previous studies on methods to reduce errors in radiology reports, research on algorithms to directly detect such errors has been conducted at a very basic level. Lee et al. proposed a software that detects the laterality error for the side or sites between the radiology report and its examination name. Minn et al. proposed an algorithm to detect gender and laterality mismatch in report and its metadata. Zech et al. proposed a LSTM (Hochreiter and Schmidhuber, 1997) based neural model to detect inappropriate insertions, deletions, and substitutions of words in radiology reports. As such, existing studies on error detection in radiology reports were conducted only on local parts such as gender, laterality, and a single word, and most

[*] These authors contributed equally.

Proceedings of the 4th Clinical Natural Language Processing Workshop, pages 41 - 52
July 14, 2022 ©2022 Association for Computational Linguistics

of these were done by simple matching without considering deep contextual meanings. Unfortunately, in real life, radiologists' error occur due to more complicated reasons that cannot be covered by these approaches. Considering how radiologists record their interpretation and communicate with referring physicians, capturing and understanding the contextual meaning of each section in the report is an important part for practical error detectors that can be used in real life.

In the field of NLP, many pre-trained language models (PLMs) are showing remarkable achievement in various tasks of natural language understanding since the advent of ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018). Recently, several studies on PLM that utilize world knowledge for language understanding have appeared, showing outstanding results not only in the general domain but also in domain-specific tasks (Zhang et al., 2019; Sun et al., 2019, 2020, 2021; Wang et al., 2020). Specifically, PLMs specialized in the medical domain such as ClinicalBERT (Alsentzer et al., 2019) and BioMegatron (Shin et al., 2020) have shown notable performance in the medical NLP tasks.

Despite the remarkable achievement in the field of NLP, the main barrier to apply these technologies, is the absence of radiology reports with errors to perform PLM supervised learning. Two reasons are identified behind the lack of accessible data. First, identifying errors in radiology reports can only be done by well-trained radiologists which is time-consuming and requires costly manual work. Second, in fact, radiology report errors do not occur as often enough for them to be collected and used to train deep learning models. It is estimated that in a daily practice, the rate of radiology report errors that are substantial to result in inappropriate or delayed diagnosis is less than 4% (Berlin, 2007). Also, when considering the different types of errors, classifying and collecting enough data for each type of error is an unrealistic approach.

Here, we introduce two novel approaches to identify errors in radiology reports based on the understanding of the nature of radiology reports while overcoming the challenge created by inadequate radiology report error data: 1) To compensate for the lack of data availability of radiology reports with errors, we introduce an artificial **error generator**. The error generator synthesizes errors that mimic radiologists behaviors that potentially cause errors in daily practice. It can generate different types of errors by employing appropriate and relevant medical knowledge. 2) In order to incorporate medical knowledge for detecting complex errors, we introduce a **Medical Knowledge-enhancing Pre-training** task, which is inspired by ERNIE1.0(Sun et al., 2019), to our BERT based error detector. This additional pre-training task allows the detector to directly learn medical abbreviations and frequent phrases in radiology report.

To validate our proposed approach, experiments are performed on MIMIC-CXR (Johnson et al., 2019) with part of it including intentionally generated error by a board-certified radiologists. Furthermore, through additional experiments, the proposed model was able to identify errors in original MIMIC-CXR which was verified by human evaluation. The experiment results show that the error detector can detect errors in real-world data while it is trained on artificial errors generated by the error generator. Additionally, external validation, experiments on domain adaptation, and several ablation studies well prove the generalizability of the error detector and the performance of the knowledge-enhancing pre-training task.

In summary, our main contributions are as follows:

1. We propose **RRED** (**R**adiology **R**eport **E**rror **D**etector) which is a deep learning framework that can detect radiology report errors based on rich understanding of context and medical knowledge.
2. We propose an **error generator** that systematically generates realistic errors in the radiology reports by integrating medical knowledge.

2 Method

Figure 1 illustrates the suggested complete framework. The following sections describe the Error Generator and the Error Detector independently.

2.1 Error Generator

While there can be many types of errors in radiology reports, this study aims to detect errors occurring when writing the conclusion section based on the findings section. In order for the error generator to synthesize realistic radiology report errors, we have categorized the errors into two types based on previous works on categorization of errors in radiology reports. For clarity, Appendix C, Table 9 provides examples of each type of error.

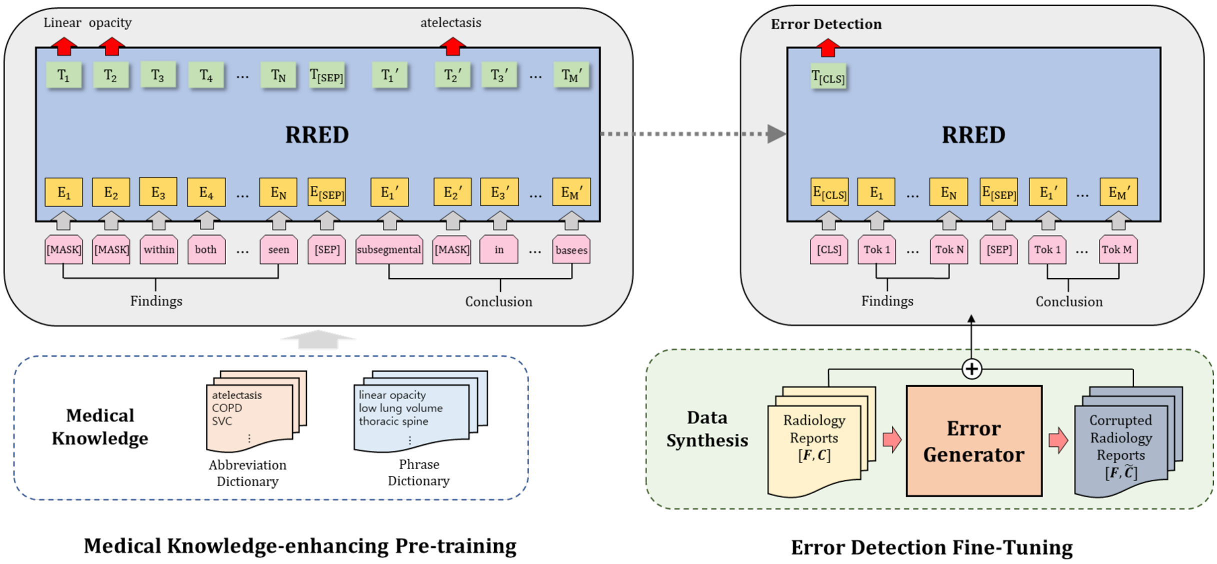

Figure 1: The overall framework of RRED.

2.1.1 Interpretive Error

Interpretive error is any error that changes the interpretation of the findings section in one way or another. This type of error can be subdivided into 3 classes based on their causes.

Faulty reasoning Errors in which findings were identified but attributed to the wrong cause. This occurs due to lack of knowledge or experience of the interpreter or due to lack of information provided in the findings section. For instance, when the conclusion section identifies cardiomegaly while the findings section only identifies pneumothorax, this is clearly an error.

Absence of abnormalities Errors in which abnormalities described in the findings sections are missed in the conclusion section.

Presence of incorrect findings Errors in which abnormalities are described in the conclusion section while the finding section clearly states that there are no findings.

2.1.2 Factual Error

Factual error is any error in which the interpretation and identification of abnormalities are correct while there are discrepancies in the description of the lesion itself. This can be subdivided into 2 classes:

Discrepancy in location of the lesion Errors in which the direction of the lesion location is mistaken (e.g. left $\rightarrow$ right, lower $\rightarrow$ upper).

Discrepancy in numerical measurement of the lesion This type of error includes errors in which the measured unit is incorrectly recorded in the conclusion section (e.g. cm $\rightarrow$ mm, mm $\rightarrow$ m) or when decimal points are misplaced or missed (e.g. 12.20 $\rightarrow$ 1.220, 8.25 $\rightarrow$ 82.5).

When factual errors occur, surgeries and biopsies can be operated on the wrong side of the body which can potentially harm the patient physically.

The error generator generates realistic errors from existing radiology reports which can create synthesized datasets that can be used to train the error detector. The synthesized data is required to be realistic enough to train a robust error detector that can detect errors in real life radiology practice. The following sections will describe the details of the error generator.

2.1.3 Error Generator Overview

The error generator consists of two steps: 1) Labeling each report using CheXpert labeler (Irvin et al., 2019) 2) Applying errors based on the tree structure which is based on the CheXpert classes mentioned in the following subsection.

2.1.4 CheXpert Labeler and its tree structure

CheXpert labeler predicts the probability of 14 different classes shown in Figure 2. The error generator first uses this to label each of the radiology reports provided. Two board certified radiologists expanded the labels of CheXpert to group similar labels which creates a tree structure. These similar labels can be interchangeable depending on the interpretation of the radiologist, therefore, cannot be considered incorrect when a different label is used within the similar label group. Figure 2 indicates the similar groups in different colors (other than

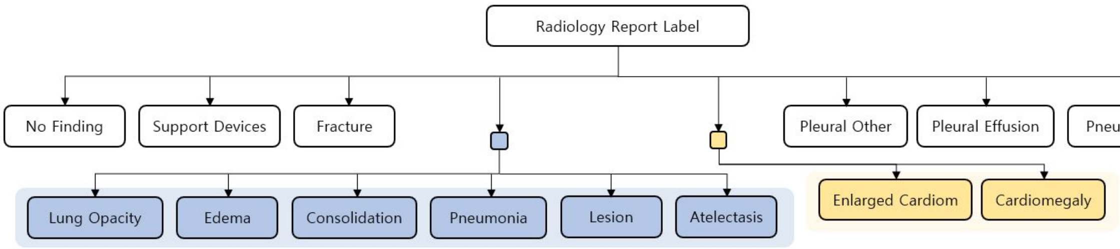

Figure 2: Diagram that illustrates the tree structure of the CheXpert labeler's labels. The labels highlighted in blue and yellow are two different similar label groups, in which the labels can be interchangeable within their own group.

white). In other words, labels within the blue region are interchangeable and the labels within the yellow region are also interchangeable.

2.1.5 Applying errors

Each type of error is applied using the labels labeled by the CheXpert labeler. To avoid any uncertainties, the entire set of labels CheXpert can label is indicated by U. Also, any report that has a label "No Finding" is noted by NF. In order to generate realistic errors, generating faulty reasoning error, absence of abnormalities, or factual error in reports in NF should be avoided. When there are no findings in the provided report, there is no medical significance in misidentifying the cause, removing findings, or creating errors in measurement or location of a lesion.

Faulty reasoning error is applied by randomly swapping the conclusion of the report with other reports' conclusion which has a different label from the original label. Since the CheXpert labeler is a multi-label labeler, the generator precisely and randomly selects a conclusion from the set $\{\mathrm{U} - \mathrm{NF} - \{original\ labels\}\}$. **Absence of abnormalities** is applied by randomly swapping the conclusion of the report with reports in NF. **Presence of incorrect findings** can only be applied when the given radiology report is in NF. It is applied by randomly swapping the conclusion of the reports in $\{\mathrm{U} - \mathrm{NF}\}$.

Discrepancy in location of the lesion is applied by detecting the keywords that indicate the location of the lesion. The keywords are the following: left, right, upper, lower, high, low, big, and small. When the keywords are identified, they are replaced by their counter-keywords which are: right, left, lower, upper, low, high, small, and big, respectively. **Discrepancy in numerical measurement of the lesion** is applied by first detecting any numerical measurement of a lesion (with its unit). Then, by a 50-50 chance, either the unit or the numerical value is changed.

2.2 Error Detector

The Error Detector uses a BERT-base architecture, which showed remarkable achievement in natural language understanding, to detect errors based on syntactic and semantic understanding of the radiology report. The parameters are initialized to the ClinicalBERT parameters which showed better performance in the medical domain.

2.2.1 Medical Knowledge-enhancing Pre-training Task

Radiology reports frequently contain medical abbreviations and phrases with specific meanings and we want the model to be able to capture richer local and global contexts for these. So, we introduce a Medical Knoweldge-enhancing Pre-training task (MKP), inspired by the Knowledge Integrated Masked Language Modeling task of ERNIE1.0, to obtain an integrated representation of such medical knowledge.

Specifically, we selected abbreviations and phrases from a radiology report to directly mask the corresponding tokens and predict the whole masked tokens. Abbreviations were identified using a medical abbreviation dictionary from imantsm and Aristotelis, and phrases were identified using a phrase dictionary created by keyBERT (Grootendorst, 2020) on MIMIC-CXR. For each report, one of the abbreviations or phrases detected in the dictionaries was randomly selected and all corresponding tokens were masked. For other tokens, probabilistic masking strategy was applied in the same way as BERT's masked-language modeling (MLM). To assist the model to capture the meaning of abbreviations and phrases effectively, border tokens of the abbreviation and phrase tokens were not masked.

2.2.2 Training Process

Pre-training MKP is performed on the MIMIC-CXR dataset which includes 91,544 chest X-ray reports. Because ClinicalBERT, which shows a sufficient level of understanding of medical domain text, is used as the initial weight, heavy pre-training for large-scale corpus is not performed. The maximum sequence length, batch size and training epochs were set to 512, 32 and 50 respectively. We performed experiments with models pre-trained for 100 and 150 epochs, but there were no significant differences observed in error detection task performance between these models.

Fine-tuning The training objective of the error detector is to perform a binary classification between original reports and corrupted reports. The training set, namely machine-synthesized dataset, consists of original reports and corrupted reports generated by the error generator. The error detector takes the concatenation of the findings and conclusion sections of the radiology report with a separator token as an input. The input representation is created by adding different segment embedding to distinguish them from each other. Also, positional embedding is added in the same way as BERT. Taking into account the general length of each section in a radiology report, the maximum lengths of findings and conclusion are 338 tokens and 172 tokens, respectively.

3 Experiments

In this section, we describe the datasets, implementation details, and experiment results of the error detection task on several datasets.

3.1 Datasets

3.1.1 MIMIC-CXR

MIMIC-CXR is a publicly available dataset consisting of chest X-rays and corresponding radiology reports, collected from patients between 2011 and 2016 at the Beth Israel Deaconess Medical Center Emergency Department. We used the train-test split disclosed in THE MEDIQA 2021 shared task (Abacha et al., 2021), which consists of 91,544 train sets and 2,000 test sets sampled by simple criterion such as acceptable length. Out of the 91,544 training examples, errors were generated on 88,388 examples (96.55% of the training set) where the percentage of interpretation error and factual error were 85.06% and 14.94%, respectively. For the test set, errors were generated on 1,933 examples (96.65% of the test set) where the percentage of the interpretation error was 79.51% and the percentage of the factual error was 20.49%.

3.1.2 Open-I

Open-I (Demner-Fushman et al., 2016) is another publicly available chest X-ray and radiology report dataset. It is collected from the Indiana Network for Patient Care, consisting of 2,928 reports. We used this dataset as an external dataset to check the generalizability of the model, meaning that both pre-training and fine-tuning is only performed on MIMIC-CXR, and the Open-I is tested in a completely unseen state. Using the error generator, errors were generated on 2,813 examples (96.07% of the dataset) where the percentage of the interpretation error and factual error were 89.69% and 10.31%, respectively.

3.1.3 Radiologist-synthesized dataset

To verify that the error detector trained on the dataset generated by the error generator can work on the real-world error generated by the radiologist, we prepared a dataset in which two board-certified radiologists manually injected errors into the MIMIC-CXR test set of THE MEDIQA 2021 split. Errors were injected into the conclusion section of 111 randomly selected reports out of a total of 2,000 reports, and 7 types of errors were generated to comprehensively verify the various types of errors that could actually occur.

The following types were considered as interpretive errors: Written as a wrong cause that is easy to confuse due to lack of knowledge or experience (Type 1-A, 18%), written as a completely nonsensical disease (Type 1-B, 18%), written in the absence of abnormalities (Type 1-C, 18%), written in the presence of incorrect findings (Type 1-D, 18%). The following types were considered as factual errors: Discrepancy in location of the lesion (Type 2-A, 19%), discrepancy in the numerical measurement of the lesion (Type 2-B, 4%). Additionally, free-form errors that do not fall into any of the six categories (Type 3, 5%).

3.2 Experimental Setups

After generating corrupted MIMIC-CXR using the error generator, we fine-tune the detector model on machine-synthesized data. This machine-synthesized data has 141,420 reports for the training set and 35,356 reports for the validation set. We tune the initial learning rate $\in$ {1e-6, 5e-6, 1e-5, 5e-5, 2e-4, 2e-3}, batch size ∈ {16, 32}, number of epochs ∈ {1, 3, 5, 10, 20}. Adam optimizer is used and other hyperparameters are fixed to their default values. The optimal setting is determined by AUPRC on MIMIC-CXR and the decision threshold for binary classification is set to a value representing precision 0.99 on the training set.

3.3 Experimental Results on datasets with Synthesized Error

The performance of our proposed framework for each dataset is shown in Table 1. Test results on MIMIC-CXR and Open-I, which are machine-synthesized datasets using our error generator, showed very high scores in all metrics including the area under precision recall curve (AUPRC) and the area under the ROC curve (AUROC). Showing these results even without training on Open-I, which is collected from a completely different hospital, means that the proposed framework has high generalizability to unseen data. A domain adaptation strategy can be attempted to further improve performance on the external dataset, and the experimental results are provided in Appendix B.

Experimental results on the radiologist-synthesized dataset also showed a significant level of performance. This means that the proposed framework that learns from errors generated by the error generator is highly applicable to real-world data. According to the experimental results, it is expected that the proposed model can detect 63% of all reports with errors with 87% of precision in the actual field. As shown in the precision-recall curve in Appendix A, Figure 3 and Figure 4, precision and recall can be set to an appropriate level by adjusting the decision threshold. Recall by each type of error with different precision criterion is also provided in Appendix A, Table 7.

3.4 Human Evaluation of RRED

To evaluate the practical ability of the proposed framework detecting actual errors in real world dataset, the trained model was inferred to the entire original MIMIC-CXR and the results are evaluated. As a result of inference, it is predicted that errors exist in 408 reports, which is 0.44% of the 93,544 reports. For 100 randomly selected cases, a board-certified radiologist was asked to answer 'Yes'/'No' to the following questions:

1. **Question 1**: There is an error between the findings and the conclusion.
2. **Question 2**: Among those where the answer to Question 1 is 'Yes', factual error that is not appropriate for findings, exists in conclusion. (e.g., discrepancies in laterality, numbers and the existence of unreported facts.)
3. **Question 3**: Among those where the answer to Question 1 is 'Yes', interpretive error that is not appropriate for findings, exists in conclusion. (e.g., faulty reasoning, missing important interpretation.)

The percentages of 'Yes' for the three questions are shown in Table 2. It can be seen that the actual error rate is 81% for the 100 selected cases, which is fairly consistent with the evaluation result on radiology-synthesized data showing about 87% of precision. In addition, it is observed that about 73% of the detected errors are factual errors, about 65% are interpretive errors and 31% are both. The detected examples of report with errors in MIMIC-CXR is shown in Table 3. Through this human evaluation result, we can expect that the proposed framework can be effectively applied in real radiology practice to detect factual errors and interpretive errors.

3.5 Effect of Medical Knowledge-enhancing Pre-training

Three experiments are performed to verify the effectiveness of the proposed Medical Knowledge-enhancing Pre-training (MKP) in various aspects. Table 4 shows the mean performance improvement by MKP for each dataset. The improvement for the machine-synthesized datasets (MIMIC-CXR, Open-I, and Open-I*) seem to be marginal as they are already scoring close to 1.0, but they show a consistent improvement for most metrics. For radiologist-synthesized dataset, the performance gains are more noticeable. Table 5 shows that the level of recall for types 1-A and 1-D, which are interpretive errors, increased. These observations suggest that MKP gives the model a higher level of understanding of medical context and knowledge, allowing the model to detect more complex types of errors.

Table 6 is the ablation result showing the performance change when each component is excluded from MKP. We can see that the masking strategy on medical abbreviations and phases is highly useful. When compared to the result of MLM only, it pushes the AUPRC score from 0.773 to 0.798 on

	AUPRC	AUROC	Precision(ppv)	Recall(sensitivity)	Specificity	Accuracy
MIMIC-CXR	0.998 (0.00)	0.998 (0.00)	0.992 (0.00)	0.964 (0.00)	0.993 (0.00)	0.979 (0.00)
Open-I	0.993 (0.00)	0.994 (0.00)	0.986 (0.00)	0.935 (0.01)	0.988 (0.00)	0.963 (0.00)
Radi-synth*	0.798 (0.03)	0.950 (0.02)	0.870 (0.05)	0.633 (0.03)	0.994 (0.00)	0.974 (0.00)

Table 1: Performance on MIMIC-CXR, Open-I and our Radiologist-synthesized dataset(*). These are the mean performance and its standard deviation from 10 random bootstrap experiments. Since there are no other studies to compare the performance, we only showed the performance of the proposed model without the baseline.

	Percentage of 'Yes'
Question 1	81.00
Question 2	72.84
Question 3	65.43
Question 2 & 3	31.00

Table 2: Quality of RRED assessed by a board-certified radiologist evaluator.

radiology-synthesized dataset. In particular, when the phrase masking is excluded, the performance is significantly reduced(AUPRC 0.798→0.769) showing that knowledge integration for phrases can provide significant understanding ability. Finally, the MLM on radiology report also seems to have an important effect on improving the overall understanding of the report itself.

4 Discussion

Through the evaluation of RRED, we mainly focus on precision rather than recall. This is because this study aims to develop a practical and reliable error detector that can be used in daily practice with a low false alarm rate. We believe that minor errors are worth missing if the alarm can provide a strong guarantee of actual errors to the radiologists.

Despite the fact that the experimental results show notable effectiveness of our approach, there are some limitations. First, the types of errors that have been implemented and experimented with, do not represent the entire scope of radiology report errors. While interpretive and factual errors are critical in the process of diagnosis, expanding the type of errors would be beneficial to reflect real-life errors in radiology practice. Second, the error generator relies on simple random swapping to generate interpretive errors. Although the experimental results show how this method is effective in large dataset like MIMIC-CXR, it is evident that this does not fully reflect the true nature of the real-life interpretive error. If the error generator can improve its' ability to imitate the behavior of radiologists, the error detector is expected to capture complex interpretive errors more precisely.

5 Conclusion

In this paper, we present **RRED**, a **R**adiology **R**eport **E**rror **D**etector based on a rich understanding of context and medical knowledge with supervised deep learning framework. We also propose a error generator for generating synthetic report data with errors to train the detector model. Through various types of evaluations, we showed that our framework can be effectively applied to real world data to detect errors that could cause inappropriate or delayed diagnosis. We also showed the significant effects of the MKP which is a proposed pre-training task to integrate medical knowledge into pre-training language model.

To the best of our knowledge, this is the first work proposing PLM based error detection model for radiology reports. In future works, we plan to develop **RRED2.0** with improved error generator and detector: 1) We will investigate more systematic approaches to generate a broader range of errors in radiology reports, in an effort to expand and improve the usability of the radiology report error detector. 2) We will expand this work to develop a vision-language error detector that can detect errors also in the findings section which is intended to record findings when reading radiographs.

We expect this work to become a practical foolproof system that can reduce critical errors in radiology reports to improve the quality of radiology reporting process and further, the entire diagnosis process.

References

Asma Ben Abacha, Yassine M'rabet, Yuhao Zhang, Chaitanya Shivade, Curtis Langlotz, and Dina Demner-Fushman. 2021. Overview of the mediqa 2021 shared task on summarization in the medical domain. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 74–85.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew

Error Type	Findings	Conclusion
Factual	The cardiomediastinal and hilar contours are stable. There has been interval increase in the **right pleural effusion** with a rounded contour concerning for loculation. There is no left pleural effusion. There is no pneumothorax. There is no focal consolidation concerning for pneumonia. Pulmonary vasculature is within normal limits.	Enlarged **left pleural effusion**, now possibly loculated.
Interpretive	A portable frontal chest radiograph shows the **large left lower lobe mass seen on recent CT chest**. New opacity adjacent to the aortic knob could represent pneumonia or fluid tracking up into the fissure. There is no appreciable pleural effusion or pneumothorax. The visualized upper abdomen is unremarkable.	Possible small left upper lobe pneumonia or pleural effusion extending into the major fissure. Large left lung mass, **less likely malignant**.
Both	Elevation of the **left hemidiaphragm** is new since prior exams, **with minimal adjacent relaxation atelectasis of the left lower lobe**. The cardiomediastinal contours are within normal limits. The bilateral hila are unremarkable. The lungs are clear without focal consolidation. There is no evidence of pulmonary vascular congestion. There is no pneumothorax or pleural effusion.	New **right hemidiaphragmatic** elevation. **Consider evaluation right hemidiaphragm function**. Otherwise, no acute cardiopulmonary process.

Table 3: Examples of actual errors detected in MIMIC-CXR by RRED. We can see that the error actually exists in the highlighted area for each error type. In the example of factual error, the location is described differently. In the example of interpretive error, the mass of the left lung is overestimated as less malignant. In the last example, there is not only a discrepancy of location, but also the important information of the findings is over-summarized.

	AUPRC	AUROC	Precision(ppv)	Recall(sensitivity)	Specificity	Accuracy
MIMIC-CXR	0.001 (<.001)	0.001 (<.001)	0.003 (<.001)	0.018 (<.001)	0.003 (<.001)	0.010 (<.001)
Open-I	0.001 (.065)	0.002 (<.001)	-0.001 (.410)	0.011 (<.001)	-0.001 (.300)	0.005 (<.001)
Open-I*	0.001 (.016)	0.000 (.071)	0.003 (<.001)	-0.007 (<.001)	0.002 (<.001)	-0.002 (<.001)
Radi-synth	0.057 (<.001)	0.006 (<.001)	0.061 (<.001)	0.047 (.005)	0.003 (<.001)	0.005 (<.001)

Table 4: Mean performance improvement by Medical Knowledge-enhancing Pre-training for each dataset and p-values of paired t-test. Open-I* indicates the performance of RRED tested after domain adaptation on the Open-I.

	1-A	1-B	1-C	1-D	2-A	2-B	3	Total
w/o MKP	0.20	**0.55**	0.65	0.50	**0.81**	0.20	0.00	0.50
w/ MKP	**0.50**	**0.55**	**0.70**	**0.75**	**0.81**	**0.60**	**0.2**	**0.64**

Table 5: Comparison of recall for each type of error between models with and without MKP.

	AUPRC	AUROC	Precision(ppv)	Recall(sensitivity)	Specificity	Accuracy
Full MKP	**0.798 (0.03)**	0.950 (0.01)	0.870 (0.05)	**0.633 (0.03)**	0.994 (0.00)	**0.974 (0.00)**
− Abbreviation	0.792 (0.03)	**0.951 (0.01)**	**0.873 (0.04)**	0.609 (0.03)	**0.995 (0.00)**	0.972 (0.00)
− Phrase	0.769 (0.04)	**0.951 (0.01)**	0.863 (0.05)	0.632 (0.03)	0.994 (0.00)	0.973 (0.00)
− Abbreviation & Phrase *	0.773 (0.04)	0.945 (0.02)	0.841 (0.05)	0.618 (0.03)	0.993 (0.00)	0.971 (0.00)
No Pre-training	0.741 (0.03)	0.944 (0.01)	0.810 (0.06)	0.586 (0.03)	0.992 (0.00)	0.968 (0.00)

Table 6: Ablation results of Medical Knowledge-enhancing Pre-training (MKP). These are the mean performance and its standard deviation from 10 random bootstrap experiments on Radiologist-synthesized Dataset. − Abbreviation & Phrase * indicates the case where only MLM is considered.

McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Aristotelis. 2018. wordlist-medicalterms-en.

Leonard Berlin. 2007. Radiologic errors and malpractice: a blurry distinction. *American Journal of Roentgenology*, 189(3):517–522.

Michael A Bruno, Eric A Walker, and Hani H Abujudeh. 2015. Understanding and confronting our mistakes: the epidemiology of error in radiology and strategies for error reduction. *Radiographics*, 35(6):1668–1676.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

imantsm. 2022. medical_abbreviations.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 590–597.

Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. 2019. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8.

Young W Kim and Liem T Mansfield. 2014. Fool me twice: delayed diagnoses in radiology with emphasis on perpetuated errors. *AJR Am J Roentgenol*, 202(3):465–470.

Young Han Lee, Jaemoon Yang, and Jin-Suck Suh. 2015. Detection and correction of laterality errors in radiology reports. *Journal of digital imaging*, 28(4):412–416.

Matthew J Minn, Arash R Zandieh, and Ross W Filice. 2015. Improving radiology report quality by rapidly notifying radiologist of report errors. *Journal of digital imaging*, 28(4):492–498.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Antonio Pinto, Luca Brunese, Fabio Pinto, Riccardo Reali, Stefania Daniele, and Luigia Romano. 2012. The concept of error and malpractice in radiology. In *Seminars in Ultrasound, CT and MRI*, volume 33, pages 275–279. Elsevier.

Minal Jagtiani Sangwaiya, Shyla Saini, Michael A Blake, Keith J Dreyer, and Mannudeep K Kalra. 2009. Errare humanum est: frequency of laterality errors in radiology reports. *American Journal of Roentgenology*, 192(5):W239–W244.

Hoo-Chang Shin, Yang Zhang, Evelina Bakhturina, Raul Puri, Mostofa Patwary, Mohammad Shoeybi, and Raghav Mani. 2020. BioMegatron: Larger biomedical domain language model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4700–4706, Online. Association for Computational Linguistics.

Chris L Sistrom and Curtis P Langlotz. 2005. A framework for improving radiology reporting. *Journal of the American College of Radiology*, 2(2):159–167.

Yu Sun, Shuohuan Wang, Shikun Feng, Siyu Ding, Chao Pang, Junyuan Shang, Jiaxiang Liu, Xuyi Chen, Yanbin Zhao, Yuxiang Lu, et al. 2021. Ernie 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. *arXiv preprint arXiv:2107.02137*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Xuyi Chen, Han Zhang, Xin Tian, Danxiang Zhu, Hao Tian, and Hua Wu. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Yu Sun, Shuohuan Wang, Yukun Li, Shikun Feng, Hao Tian, Hua Wu, and Haifeng Wang. 2020. Ernie 2.0: A continual pre-training framework for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8968–8975.

A Wallis and P McCoubrie. 2011. The radiology report—are we getting the message across? *Clinical radiology*, 66(11):1015–1022.

Ruize Wang, Duyu Tang, Nan Duan, Zhongyu Wei, Xuanjing Huang, Guihong Cao, Daxin Jiang, Ming Zhou, et al. 2020. K-adapter: Infusing knowledge into pre-trained models with adapters. *arXiv preprint arXiv:2002.01808*.

John R Wilcox. 2006. The written radiology report. *Applied Radiology*, 35(7):33.

John Zech, Jessica Forde, Joseph J Titano, Deepak Kaji, Anthony Costa, and Eric Karl Oermann. 2019. Detecting insertion, substitution, and deletion errors in radiology reports using neural sequence-to-sequence models. *Annals of translational medicine*, 7(11).

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. *arXiv preprint arXiv:1905.07129*.

A Appendix

Figure 3 and Figure 4 shows the Precision-Recall curve and ROC curve on Radiologist-synthesized dataset, respectively.

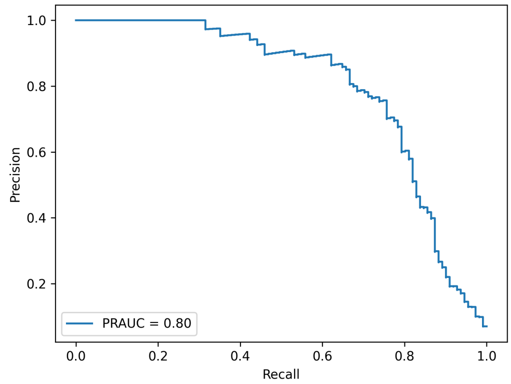

Figure 3: Precision-Recall curve on Radiologist-synthesized dataset

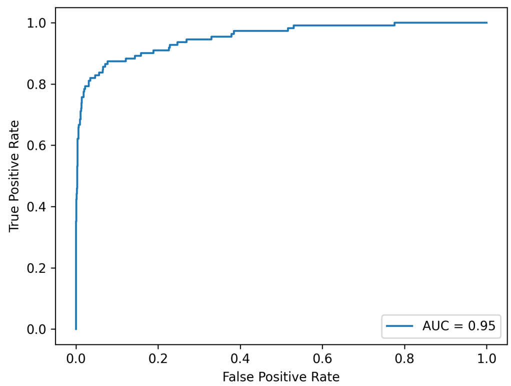

Figure 4: ROC curve on Radiologist-synthesized dataset

Table 7 shows that, even with an precision of 1.0, about 31% of errors can be detected, which means that a certain amount of errors can be detected even with a false-alarm rate close to zero in the real world. When precision is set to 0.96, the recall for factual error (type 2) rises remarkably. Also, the recall of interpretive error (type 1) is increased when it is set to 0.8 to 0.9. The recommended precision settings for a false alarm rate is around 0.96, and for a better detection of interpretive errors is around 0.8 to 0.9.

B Appendix

Table 8 shows the experimental results regarding the effectiveness of the domain adaptation strategy that can ideally improve the performance on

Precision(ppv)	1-A	1-B	1-C	1-D	2-A	2-B	3	Total
1	0.1	0.35	0.5	0.35	0.38	0.2	0	0.31
0.96	0.1	0.35	0.55	0.4	0.71	0.6	0	0.41
0.9	0.35	0.5	0.65	0.6	0.81	0.6	0	0.56
0.8	0.55	0.6	0.7	0.8	0.86	0.6	0.2	0.68

Table 7: Recall by error type with different precision criterion

	AUPRC	AUROC	Precision(ppv)	Recall(sensitivity)	Specificity	Accuracy
Domain Adaptation X	0.993 (0.00)	0.994 (0.00)	0.986 (0.00)	0.935 (0.01)	0.988 (0.00)	0.963 (0.00)
Domain Adaptation O	**0.997 (0.00)**	**0.998 (0.00)**	**0.994 (0.00)**	**0.944 (0.00)**	**0.995 (0.00)**	**0.970 (0.00)**
Difference (P Value of paired t-test)	0.004 (<.001)	0.004 (<.001)	0.008 (<.001)	0.009 (<.001)	0.007 (<.001)	0.008 (<.001)

Table 8: Performance increase by domain adaptation on Open-I dataset

external dataset. For domain adaptation, 1 epoch training is performed on 500 reports of Open-I after fine-tuning on MIMIC-CXR. Even with this light training, a statistically significant level of consistent performance improvement is observed for all metrics. Therefore, when applying the proposed framework to the real life scenarios, performance improvement can be expected if domain adaptation is performed with synthetic data generated by the error generator.

C Appendix

Table 9 provides some examples of errors generated by the error generator using MIMIC-CXR.

Error Class	Sub-class	Examples		
		Findings	Error-free Conclusion	Conclusion with synthesized error
Inter-pret-ation Error	**Faulty Reasoning**	Heart size is mildly enlarged. The mediastinal and hilar contours are normal. The pulmonary vasculature is normal. Lungs are clear. There is minimal blunting of the left costophrenic sulcus suggestive of a trace left pleural effusion. No right-sided pleural effusion is present. There is no pneumothorax. No acute osseous abnormalities detected.	Small left pleural effusion. Otherwise, no acute cardiopulmonary abnormality.	Improved left lower lobe atelectasis but are persistent consolidation is concerning for aspiration pneumonia. The same is true for the smaller region can of consolidation in the left lung base medially. NG tube in appropriate position.
	Absence of abnormalities	AP portable chest radiograph demonstrates interval placement of a nasogastric tube, which appears to descend the thorax in an uncomplicated course. The terminal tip appears at the anticipated location of the gastroesophageal junction. For standard placement within the stomach advance approximately 8 cm. Streaky opacity in the left lung base is reflective of atelectasis. Bibasilar atelectasis is persistent on the right and slightly improved on the left. Lung volumes are overall low. There is no pneumothorax or pleural effusion. Note is made of chronic deformity of the right humeral neck.	Interval placement of an enteric tube. Recommend advancement approximately 8 cm for more appropriate positioning within the gastric lumen. Bibasilar atelectasis.	No acute cardiopulmonary abnormality.
	Presence of incorrect findings	PA and lateral views of the chest were provided demonstrating clear lungs without focal consolidation, effusion or pneumothorax. The cardiomediastinal silhouette is normal. Bony structures are intact. No free air below the right hemidiaphragm.	No acute intrathoracic process.	In comparison to ___, there is increased opacity of the left mid lung, which may represent fluid trapped in an accessory minor fissure. This could be confirmed with a additional lateral view of the chest or CT. Increased left pleural effusion
Factual Error	**Discrepancy in location of the lesion**	A right PICC ends in the mid SVC. There has been a significant decrease in size of the right pleural effusion but no change in marked right lower lobe atelectasis. There is no pneumothorax. Apical bullous disease is stable. Left basilar atelectasis has improved. There is no new consolidation. The cardiomediastinal silhouette is normal.	Decrease in size of right pleural effusion after thoracentesis. No pneumothorax. Persistent marked right lower lobe atelectasis. Near resolution of left basilar atelectasis. Resolved results were discussed with ___ at 4:30 p.m. on ___ via telephone by Dr ___.	Decrease in size of right pleural effusion after thoracentesis. No pneumothorax. Persistent marked right upper lobe atelectasis. Near resolution of left basilar atelectasis. Resolved results were discussed with ___ at 4:30 p.m. on ___ via telephone by Dr ___.
	Discrepancy in numerical measurement of the lesion	Bilateral atelectasis is mild. An approximately 2.6 x 2.1-cm lobulated opacity projecting over the left apex is new since ___ and has a mass-like appearance. No pleural effusion, pneumothorax, or edema. The heart is top-normal in size, unchanged. No acute osseous abnormality. Biapical pleural thickening is worse on the left.	New 2.6-cm lobulated opacity projecting over the left apex since ___ could be an underlying mass. Chest CT non-emergent is recommended to further evaluate left apex lobulated mass in setting of reported history of a right breast mass.	New 26.0 cm lobulated opacity projecting over the left apex since could be an underlying mass Chest CT non emergent is recommended to further evaluate left apex lobulated mass in setting of reported history of a right breast mass

Table 9: Examples of different types of errors in radiology reports. The column, **Conclusion with synthesized error**, shows the synthesized error from the error generator for each type.