IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 October 2024, accepted 19 November 2024, date of publication 20 November 2024,
date of current version 17 December 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3504378

RESEARCH ARTICLE

Using Information Extraction to Normalize the Training Data for Automatic Radiology Report Generation

YUXIANG LIAO[1], HAISHAN XIANG[2], HANTAO LIU[1], AND IRENA SPASIĆ[1]
[1]School of Computer Science and Informatics, Cardiff University, CF24 4AG Cardiff, U.K.
[2]Baoan Center for Disease Control and Prevention in Shenzhen, Baoan District, Shenzhen 518000, China

Corresponding author: Irena Spasić (spasici@cardiff.ac.uk)

The work of Yuxiang Liao was supported in part by a Ph.D. project funded by China Scholarship Council-Cardiff University Scholarship under Grant CSC202108140022. The project is supervised by Irena Spasić and Hantao Liu.

ABSTRACT High lexico-syntactic variation across radiology reports even when they convey the same diagnostic information complicates evaluation and hence the training of deep learning models for Automatic Radiology Report Generation. This problem can be addressed by 1) developing an internal standard for the structured representation of radiology reports; 2) automatically converting radiology reports to the structured representation prior to training; 3) training a deep learning model to generate a structured radiology report from an image, and finally 4) converting the structured report into a narrative one. In this study, we focus specifically on steps 1) and 2). First, we proposed a structured radiology report scheme based upon RadGraph, which serves to formally represent clinical entities, their attributes and relations discussed in a radiology report. Using the new scheme, we manually annotated a total of 550 MIMIC-CXR reports for model training and evaluation and 50 CheXpert reports for evaluating the model's generalization ability. We developed a joint entity and relation model and proposed a novel auxiliary component to enhance the model performance by interpreting token-level information. Using the annotated data, we trained the model for automatically converting information from a narrative radiology report into the structured representation, which achieved a micro-F1 of 96.6% and 96.1% on named entity recognition, 94.0% and 89.8% on entity attribute recognition, and 89.5% and 86.6% on relation extraction, on the MIMIC-CXR and CheXpert test sets, respectively. We then used this model to automatically annotate 227,835 MIMIC-CXR reports. We shared all data and software deliverables using PhysioNet Credentialed Health Data License 1.5.0 to enable further research on Automatic Radiology Report Generation.

INDEX TERMS Information extraction, natural language processing, named entity recognition, relation extraction, structured radiology report.

I. INTRODUCTION

Narrative radiology reports vary excessively in their language, length, and style, thereby limiting their utility in clinical research and other downstream applications. This issue has given rise to the idea of automatic structuring of radiology reports. It focuses on extracting key medical information from the free text, typically through named entity recognition (NER) and relation extraction (RE). Given a sequence of text, NER identifies text pieces as entities corresponding to predefined entity types; while RE assigns predefined relation types to pairs of entities. Completing this task involves efforts in three aspects: data, scheme, and model. The data defines the scope for model training and application, the scheme comprises predefined entity and relation types to guide human annotators in data annotation, and the model is trained using annotated data to ensure accurate predictions.

Automatic Radiology Report Generation (ARRG) focuses on utilizing deep learning methods to generate reports from

The associate editor coordinating the review of this manuscript and approving it for publication was Biju Issac.

radiology images, with an emphasis on recognizing normal and abnormal appearances and describing them accurately and comprehensively. Our recent review of ARRG proposed that structured reports can alleviate the inherent diversity of natural language, thus contributing to more accurate results in the model training and evaluation [1]. Although numerous studies have investigated extracting information from radiology reports [2], [3], [4], [5], [6], [7], they may require specific adaptations for ARRG. For instance, radiology reports commonly include comparisons of observations between the current and previous examinations in the findings and impression sections. However, de-identification processes applied in existing open-source large-scale datasets such as MIMIC-CXR [8] make it impossible to identify the corresponding prior references, leading to inconsistencies between the reported observation and those observed in the current image. Using such inconsistent data to train ARRG models will inevitably produce hallucinations [9]. In addition, the content of the report is usually related to the expression preferences of radiologists. For example, the impression sections of the following three reports all indicate that the patient's examination results are normal: (1) "Normal chest," (2) "Clear lungs. Heart size normal," and (3) "Both lungs are clear with no focal consolidation, pleural effusion, or pneumothorax. Normal cardiomediastinal and hilar silhouettes and pleural surfaces." However, training and testing ARRG models on plain reports may implicitly emphasise the wording of radiologists and suppress the importance of diagnostic accuracy.

To better adapt to the ARRG task, we propose an extension of the RadGraph scheme [2], which was originally designed to capture the most clinically relevant information within the report in a consistent manner. Our extended scheme, referred to as CXRGraph, introduces three entity attributes: one of which indicates the normality of observations, while the other two attributes indicate the interval change of observations that refer to the priors, and the action needed to convert these relative observations to direct observations. We define the entity attribute as a value that is chosen from a set of predefined values and bound to an entity. Additionally, we introduce one extra entity type and one extra relation type to address observed inconsistencies and disagreements in the annotated results within RadGraph. Fig. 1 provides two example sentences annotated using our CXRGraph scheme. With this scheme, we re-annotate the same data used in RadGraph, including 550 radiology reports from the MIMIC-CXR dataset and 50 radiology reports from the CheXpert dataset [10]. The data splits for training, validation, and testing remain unchanged. Moreover, we develop a pipeline method comprising an entity model and a relation model. We incorporate an auxiliary token-level classification component into the model. This component transforms entity span labels into token-wise labels, promoting the model to focus on the fine-grained information within entity spans, such as determining whether a token is part of an entity span and identifying its corresponding

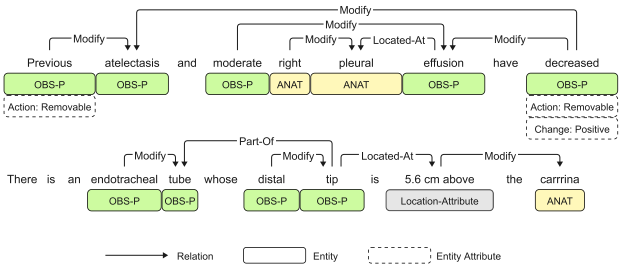


FIGURE 1. Example sentences annotated according to the CXRGraph scheme.

type. Our model achieves F1=96.6%/94.0%/89.5% (MIMIC-CXR) and F1=96.1%/89.8%/86.6% (CheXpert) on NER, entity attribute recognition (ATTR) and RE, respectively. Furthermore, our approach achieves results on par with the state-of-the-art in the joint NER and RE tasks across various domains.

Overall, our main contributions are summarized as follows:

- We propose a CXRGraph scheme specifically for the ARRG task, that can identify not only clinically relevant information in a report but also prior-referenced observations and the normality of observations. Using the scheme, we manually re-annotate a subset of radiology report. The scheme and annotated data can be used for the research in entity and relation extraction on chest X-ray reports.
- We develop a pipeline method comprising span-based entity and relation models with a novel token-level component incorporated into the models. Our method achieves comparable performance to state-of-the-art methods in the joint NER and RE tasks across various domains.
- We train and evaluate our model using the CXR-Graph data, and automatically annotate a widely used large-scale dataset – MIMIC-CXR. We release the trained model and inference data to facilitate further ARRG research, particularly the concerns on subtle yet critical issues in original reports, such as inter-subject expression bias and hallucinated prior references.

We believe future research can build upon the research results, advancing a wide range of medical practice activities in healthcare information management and radiology report generation.

II. RELATED WORK

A. AUTOMATIC RADIOLOGY REPORT GENERATION

Automatic Radiology Report Generation (ARRG) seeks to leverage deep learning techniques to automate the reporting process of radiologists. It aims to liberate radiologists from the repetitive reporting process, allowing them to focus on revising the reports and thereby enhancing the quality and efficiency of clinical communication. Recall that this study serves as a preliminary step for the ARRG research, intending

to provide accurate structured data for exploring the role and significance of such data in the ARRG task.

In early research on ARRG, many studies utilised disease labels into ARRG models. For example, some studies introduced a multi-label classification task on the disease labels as the auxiliary training target for the ARRG model [11], [12]. Sun et al. directly input the predicted labels to the text decoder, leaving aside the visual features [13]. More commonly, the embedding of the predicted labels were fused with the visual features, yielding contextual features that was subsequently passed to the text decoder [14], [15]. These disease labels comprised common thoracic diseases categorized by the Medical Text Indexing (MTI) [16], the CheXpert dataset [10], or the ChestX-ray8 dataset [17].

As research progresses in this field, there is a growing interest among researchers in employing graph structure to supersede disease labels, enabling a more granular representation of clinical observation in the reports [18], [19]. Comparing to disease labels which represent reports in terms of common disease categories as a high-level summary, graph structure provides a normalized representations of these reports regarding the key information entities and relationships, such as RadGraph [2]. To the best of our knowledge, RadGraph is the structured data processing method most widely adopted by current ARRG models. Consequently, our objective is to enhance RadGraph to produce high quality training data specified for ARRG models.

There are mainly two paradigms to utilize the graph structure in ARRG. There are two main paradigms that utilize graph structure in ARRG. One paradigm is modelling ARRG as a pipeline of image-to-graph and graph-to-text task [20], [21], [22]. This approach explicitly utilizes graph structures to improve the quality of generative language models. However, the generated report might suffer from the error propagated from the last stage, such as missing some information when the predicted graph is not accurate enough. Another paradigm interpreted the ARRG as an image-to-text task where a graph prediction module was appended to the visual encoder. In this paradigm, the graph features are typically fused with visual features and passed to the text decoder, allowing the text decoder to learn to attend to different features [23], [24], [25]. This approach enables visual features to supplement the predicted graph with missing information. Compared to the first paradigm, this approach enables visual features to supplement the predicted graph with missing information, yet requires the model to have more substantial learning and generation capabilities.

B. INFORMATION EXTRACTION

Numerous schemes have been proposed for extracting information from radiology reports obtained by different modalities including magnetic resonance imaging (MRI) [3], [6], [26], computed tomography (CT) [4], [5], [26] and X-ray [2], [6], [7], [27], [28]. These studies typically structure the information into four types of entities: observation, anatomy, and their corresponding modifiers. Observation refers to a clinical manifestation (e.g. disease) or an observed feature, while anatomy refers to the affected anatomical entity. Modifier is used to provide additional context or detail to observations or anatomical entities. However, the details of the information extraction scheme may vary from study to study.

For example, Hassanpour and Langlotz [5] introduced a new type of modifier specifically for certainty. Spasic et al. [3] assigned such a certainty modifier as an observation modifier and introduced an observation attribute for negation. Sugimoto et al. [4] redefined the common structure based on the RadLex radiology lexicon [29]. They extended observation to include both observation and clinical findings and subdivided the observation modifier into certainty, change, size, and characteristics. They merged anatomy and its modifier into a single location modifier. Jain et al. [2] extended the observation entity with three attributes (present, absent, and uncertain) to describe its certainty and negation. Based on [2], Khanna et al. [7] further defined eight types of observation entities to represent disease progression. Other schemes were designed to extract different types of information from reports such as spatial role [6], [27], [28] and clinical assertion and fact [26]. Ramesh et al. [9] identified the change statements in radiology reports and simply removed them to limit references to prior examinations. They demonstrated an improvement in ARRG models trained on such preprocessed data.

The utilization of relation types to group relevant entities can be classified into two broad categories. The first category considers an observation entity as the central element of an entity group, where relations are asserted between anatomy and observation entities and between the modifier and its corresponding subject entities [2], [3], [7]. The second category considers a spatial indicator entity as the central element and groups entities using a set of spatial relations [6], [27], [28].

To extract information from radiology reports and use it to populate templates based on the given schema, most approaches adopted it as a joint task of NER and RE by leveraging deep learning (DL) methods. Currently, there are two mainstream paradigms for designing span-based DL models for this joint task. The first one is an end-to-end approach, which begins by tokenising the input text and then constructing appropriate span/span-pair representations, and finally passing them to different feed-forward neural networks (FFNNs) to classify the entity and relation labels. Specifically, DYGIE++ [30] dynamically constructed graphs and updates span representation via graph propagation, where the re-contextualized representations were passed to different FFNNs. HGIE [7] introduced a conditional pre-training process with a hierarchical loss function to enhance the DYGIE++ model, specifically targeting entity types with hierarchical structures. SpERT.PL [31] introduced part-of-speech tags as additional features to

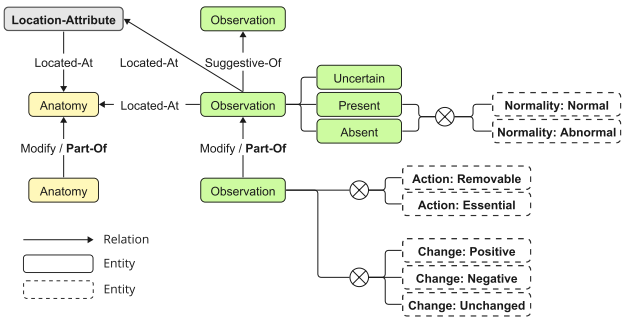


FIGURE 2. CXRGraph annotation scheme is based on RadGraph. The newly added components are indicated by the bold typeset.

produce span representations and combined two spans along with the tokens in between to form span-pair representations. TriFM [32] proposed a memory module to enhance the bi-directional interaction between entities.

The second dominant paradigm is a pipeline approach that uses two models to perform NER and RE tasks separately. The NER model encodes the original text into span representations and predicts entities. Subsequently, the RE model encodes the original text along with the predicted entities into span-pair representations and predicts the relations. In PURE [33], the NER model used a pre-trained language model to encode the input text, and used FFNNs as the NER classifier. The RE model had a similar architecture, however, the input contained additional typed entity markers inserted into the original text. PL-Marker [34] used the same framework though the markers were applied to both NER and RE models.

III. METHOD

A. ANNOTATION SCHEME

Our scheme is based on the structure of RadGraph as illustrated in Fig. 2. Its name, CXRGraph, indicates that it applies solely to chest X-ray reports. Just like RadGraph, CXRGraph is confined to the two sections of radiology reports – findings and impression, which are the primary targets of ARRG [1].

In RadGraph, an entity corresponds to a continuous span of text. These entities are assorted into four types: Anatomy, Observation-Present, Observation-Absent, and Observation-Uncertain. Anatomy (ANAT) refers to an anatomical body part corresponding to a specific observation while Observation (OBS) refers to an identifiable pathophysiologic process, a diagnostic disease, or an observed feature. Any observation is considered as Observation-Present by default unless a certainty or negation modifier is used in its context. For example, the pneumothorax in ''there is no pneumothorax'' is regarded as Observation-Absent; the pneumonia in ''possible left lower lobe pneumonia'' is regarded as Observation-Uncertain. Composite concepts are decomposed where possible and the underlying concepts are annotated. For example, ''left lower lobe'' is decomposed into three Anatomy entities.

A relation is defined as a directed link pointing from a subject entity to an object entity. Relations are assorted into three types: Suggestive-Of, Located-At, and Modify. The scope of permitted use of a relation is denoted in format: (subject entity type, relation type, object entity type). Suggestive-Of (OBS, Suggestive-Of, OBS) indicates that the presence of the subject Observation can infer, or is secondary to, that of the object Observation. Located-At (OBS, Located-At, ANAT) indicates spatial or other relations between Observation and Anatomy. Modify (OBS, Modify, OBS) or (ANAT, Modify, ANAT) indicates that the subject entity modifies the object entity with respect to size, change, degree, scope, etc.

CXRGraph adopts the above definitions of entity and relation from RadGraph and extends them incrementally with three objectives: 1) enhancing data reliability for ARRG; 2) unifying inconsistent annotation results; and 3) ensuring compatibility with RadGraph.

For the first objective, we introduce three optional attributes for the Observation entity: Normality, Action, and Change. These attributes are designed to provide additional context and specificity to the Observation entities, giving more flexibility for researchers to utilize the data for ARRG. We define the entity attribute as a value chosen from a set of predefined values and can optionally be bound to an entity. Normality has two values: Normal and Abnormal, which are optional on Observation-Present and Observation-Absent, indicating whether a group of related entities are described as normal or not. By default, Observation-Present is considered to describe an abnormal observation while Observation-Absent describes a normal observation. Hence, no Normality value is assigned in these default cases due to their disproportionately large occurrence. When a non-default condition arises, a Normality value is assigned to the corresponding modifier. For example, the clear in ''lungs are clear'' is annotated as Observation-Present with Normal. This attribute is introduced based on findings from our recent review, which suggests that separate processing of normal and abnormal samples has a positive impact on ARRG [1]. Action and Change are designed to address the hallucinated references to prior examinations [9]. They are optional on an Observation entity. Action indicates how a group of Observation entities would be converted into a direct observation, containing two values: Removable and Essential. Removable means that the entity with this attribute, along with relevant entities pointing to it, can be removed without altering the original meaning of this group of entities. Essential indicates that the entity cannot be removed without altering the meaning of this Observation group. Change indicates the progression of an observation in the interval of time since a specific prior examination, having three values: Positive, Negative, and Unchanged. It is typically assigned to a change modifier along with an Action value. For example, the stable in ''cardiomediastinal contours are stable'' is considered as Observation-Present with Essential and Unchanged, whereas in ''the heart size is enlarged, but

stable'', the stable is considered as Observation-Present with Removable and Unchanged.

For the second objective, we introduce a new entity type: Location-Attribute, and a new relation type: Part-Of. In RadGraph, the annotation of some spatial descriptions such as ''course into'', and ''terminates 2.3 cm above'' are disagreed between annotators. For example, in the sentences similar to ''Nasogastric and Dobbhoff tubes course into the stomach'', we observed instances where ''course'' was labelled as either OBS or none. In the sentences similar to ''Endotracheal tube terminates xx cm above the carina'', we observed instances where 1) ''xx cm above the carina'' was labelled as OBS, 2) ''xx cm above'' and ''carina'' as ANAT, or 3) ''xx cm'' and ''carina'' as ANAT and ''above'' as OBS. To address this issue, we introduce a Location-Attribute (LOC-ATT) entity that serves as a complement to the Located-At relation, indicating a special spatial relation between Observation and Anatomy. Ideally, the LOC-ATT entity is expected to be bound to the Located-At relation. For instance, in ''tubes course into the stomach'', we would have (tubes, Located-At, ANAT) in which the Located-At relation is directly linking from ''tubes'' to ''stomach''. We expect the LOC-ATT entity instance ''course into'' to be bound to the Located-At relation of ''tubes'' and ''stomach'' to expand the relation detail. However, we found no annotation tools that support binding individual entities to relations. As a workaround, we split a single Located-At relation into two and bridge them by the LOC-ATT entity, which is presented as (OBS, Located-At, LOC-ATT, Located-At, ANAT). More details of the post-processing of this workaround are discussed in the third objective.

We also noticed disagreements in determining subject and object entities for Modify, which could further affect the selection of an object entity for Located-At. For example, in ''pulmonary vascular congestion'', we have seen three different combinations of relations: (pulmonary, Modify, vascular) + (congestion, Located-At, vascular), (vascular, Modify, pulmonary) + (congestion, Located-At, vascular), and (vascular, Modify, pulmonary) + (congestion, Located-At, pulmonary), where the last one is the most prevalent in RadGraph. However, we prefer the second pattern as we consider that pointing an Observation entity directly to its related Anatomy entity is more intuitive. To balance consistency, intuitiveness, and compatibility, we introduce a new relation, Part-Of, as an extension of Modify. It indicates that the subject entity is part of the object entity (e.g. vascular, Part-Of, pulmonary) or is the property of the object entity (e.g. silhouette, Part-Of, hilar). The Part-Of relation allows annotation to remain intuitive while maintaining consistency with RadGraph by applying some post-processing rules. For example, the above example can now be annotated as (vascular, Part-Of, pulmonary) + (congestion, Located-At, vascular) as expected. If we want to make the annotation consistent with RadGraph, we can easily replace Part-Of with Modify, resulting in (vascular, Modify, pulmonary); and replace the subject entity of Located-At to the last entity along the relation chain (congestion, Located-At, vascular, Part-Of, pulmonary), resulting in (congestion, Located-At, pulmonary).

For the third objective, we establish several rules to facilitate the conversion of CXRGraph to RadGraph. For the Location-Attribute entity, we convert it into Anatomy and change its outward-pointing Located-At relation (LOC-ATT, Located-At, ANAT) into a Modify relation. We then remove its inward-pointing Located-At relation (OBS, Located-At, LOC-ATT) and recover the Located-At relation between the corresponding Observation and Anatomy entities. For the Part-Of relation, we change it back to Modify. We define a transfer-pointing rule: in a directed chain of entities linked by Part-Of, a relation pointing to the intermediate entity of the chain is changed to pointing to the final entity of the chain. This rule is applied when the relation type is either Located-At or Suggestive-Of.

B. MODEL

Our method follows the pipeline approach, comprising an NER model and an RE model. The NER model takes as input a sentence and its context and predicts the entities in the sentence. Based on the predicted results, the RE model takes as input a sentence and its context, along with additional marker tokens corresponding to predicted entities. Each data instance of the RE model consists of a set of entity pairs formed by taking one entity in the sentence as the subject and the other entities as the objects.

We define $X = \{x_1, x_2, \ldots, x_n\}$ as a sentence of n tokens x_i $(i = 1, \ldots, n)$. Encoder() indicates a pre-trained language model used to obtain the contextualized representation g_x of token x: $g_x(i) = \text{Encoder}(x_i)$. $\text{FFNN}^{task}_{input}()$ denotes a feed-forward neural network (FFNN) classifier that computes non-linear mapping from input vectors to task-specific logit values. ε_{ner} denotes the entity types with a None value; ε_{attr} denotes the entity attribute values with a None value; ε_{re} denotes the relation types with a None value. Our NER model aims to predict the boundaries, types, and attribute values of entities. Our RE model aims to predict the relation types based on the predicted entities from the NER model.

1) NAMED ENTITY RECOGNITION MODEL

Our NER model is a span-based model based on the Princeton University RE (PURE) system [33]. An overview of the NER model is shown in Fig. 3. Let $S(X) = \{s_1, s_2, \ldots, s_m\}$ denote all possible spans no longer than a maximum length L within the given sentence X. For each span $s_i \in S$, let START (i), END (i), and LENGTH (i) denote its start token, end token, and total length in tokens, respectively. Our RE model is framed as follows:

$$g_s(i) = \begin{bmatrix} \text{Encoder}(\text{START}(i)); \\ \text{Encoder}(\text{END}(i)); \\ \text{Embedding}(\text{LENGTH}(i)) \end{bmatrix}, \quad (1)$$

$$p_s^{ner} = \text{softmax}\left(\text{FFNN}_{span}^{ner}\left(g_s\right)\right), \tag{2}$$

$$p_s^{attr} = \text{sigmoid}\left(\text{FFNN}_{span}^{attr}\left(g_s\right)\right), \tag{3}$$

where g_x and g_s are the representations of token x and span s respectively. Embedding () denotes a learnable embedding layer used to encode the span's length features. FFNN_{span}^{ner} is a multi-class classifier that predicts the probability distribution p_s^{ner} of entity type ε_{ner}, whereas $\text{FFNN}_{span}^{attr}$ is a multi-label classifier that on predicts entity attributes ε_{attr}.

It may be counter-intuitive not to utilize the tokens between the start and end tokens of a span. Such an approach seems to assume that the inner tokens of a span and those surrounding it have the same importance. To address this issue, we introduce an auxiliary task of token-level classification:

$$p_x^{ner} = \text{sigmoid}\left(\text{FFNN}_{tok}^{ner}\left(g_x\right)\right), \tag{4}$$

where FFNN_{tok}^{ner} is a multi-label classifier that predicts the probability distribution p_x^{ner} of entity type ε_{ner} from the token representation g_x. A token that is not contained in any entities is said to be of a None type, otherwise, it inherits the types of the entities that encompass it. We set threshold=0.5 for attribute classification and token classification. The model is updated during training using a combined cross-entropy loss of the three sub-tasks. During inference, when multiple values of an attribute are predicted, the one with the highest probability is selected.

2) RELATION EXTRACTION MODEL

Our RE model is a span-based model based on packed levitated (PL) marker [34]. An overview of the RE model is shown in Fig. 4. Let $E = \{e_1, e_2, \ldots, e_t\}$ denote t entity spans in the sentence X. The RE model iteratively considers each entity in E as the subject entity with the remaining entities as its objects, thereby constructing t input instances. Let $\langle subj\rangle$, $\langle/subj\rangle$, $\langle obj\rangle$, $\langle/obj\rangle$ denote the start and end markers for the subject and object entities. The input sequence is defined as $\hat{X} = \{\ldots, \langle subj\rangle, e_j, \langle/subj\rangle, \ldots, e_n, \ldots, \langle obj\rangle, \langle/obj\rangle, \ldots\}$, where a pair of subject markers are inserted to enclose a subject entity e_j and t pairs of object markers are appended to the end of the input sequence. We define START* (i) and END* (i) as the start and end markers of an entity e_i in $\hat{X}$. Our RE model works as follows:

$$g_e\left(i\right) = \begin{bmatrix}\text{Encoder}\left(\text{START}^*\left(i\right)\right); \\ \text{Encoder}\left(\text{END}^*\left(i\right)\right)\end{bmatrix}, \tag{5}$$

$$h_e\left(j,k\right) = \text{FFNN}_{subj}^{re}\left(g_e\left(j\right)\right) + \text{FFNN}_{obj}^{re}\left(g_e\left(k\right)\right), \tag{6}$$

$$p_e^{re} = \text{softmax}\left(h_e\left(j,k\right)\right), \tag{7}$$

where $g_e\left(i\right)$ denotes the representation of an entity span $e_i \in E$. Here, $h_e\left(j,k\right)$ denotes the representation of an entity pair of which j indicates the subject entity and k indicates the object entity. The representations of the subject and object entities are passed into different FFNN classifiers and added afterwards to compute the probability distribution p_e^{re} of relation type ε_{re} for an entity pair $e_{j,k}$. As the input of the model encoder, the subject markers are considered as a normal text token while the object markers have additional constraints. For the position embedding, the object markers share the position indices with the corresponding tokens. For the attention layer, the text tokens only attend to text tokens while the object markers attend to all text tokens and the associated markers corresponding to the same object entity.

Additionally, we employ the inverse relation approach utilized in PL-Marker, which was initially proposed for coreference relation [35]. This approach enhances the uni-directional relations by adding extra relations from object entities to subject entities. Specifically, the relation types are classified into symmetric and asymmetric ones. The direction of a symmetric relation does not affect the meaning, unlike the direction of an asymmetric relation. The inverse relation from object entity to subject is assigned the same type as the original relation if it is symmetric, otherwise, it is assigned a newly created dummy type. Therefore, the relation types $\varepsilon_{re} = \{None\} \cup \{sym - types\} \cup \{asym - types\}$ are extended to $\{None\} \cup \{sym - types\} \cup \{asym - types\} \cup \{dummy - asym - types\}$ for training. During inference, we compute the final prediction score of a pair of two entities by adding the logit value of two permutations of the entity pair:

$$p_p^{re} = \text{softmax}\left(h_e\left(j,k\right) + h_e^*\left(k,j\right)\right), \tag{8}$$

where we swap the value of an asymmetric type with that of its original type in $h_e^*\left(k,j\right)$, making it consistent with $h_e\left(j,k\right)$.

We add two auxiliary tasks to the model, one of which focuses on token-level classification as we discussed in our NER model:

$$p_x^{re} = \text{sigmoid}\left(\text{FFNN}_{tok}^{re}\left(g_x\right)\right), \tag{9}$$

where the classification target is the relation type ε_{re}. The other task is inherited from PL-Marker, which aims to predict entity types ε_{ner} of the object markers:

$$p_e^{ner} = \text{sigmoid}\left(\text{FFNN}_{obj}^{ner}\left(g_e\right)\right). \tag{10}$$

The threshold for both tasks is set to 0.5. The model is updated during training using a combined cross-entropy loss from the relation classification and the two auxiliary tasks.

IV. EXPERIMENTS AND RESULTS

A. DATA ANNOTATION AND INFERENCE

We manually annotated 600 radiology reports taken from RadGraph, including 425, 75, and 50 MIMIC-CXR reports to be used for training, validation, and testing, respectively. The test also included 50 CheXpert reports. Annotation was performed using the Brat Rapid Annotation Tool [36]. Two annotators first independently annotated 50 test reports from MIMIC-CXR by adapting original RadGraph annotations. Inter-annotator agreement (IAA) was measured using the F1-score as suggested by Grouin et al. [37], resulting in 97.2% for entity types, 87.3% for entity attributes and 91.0% for relation types. However, we still noticed a

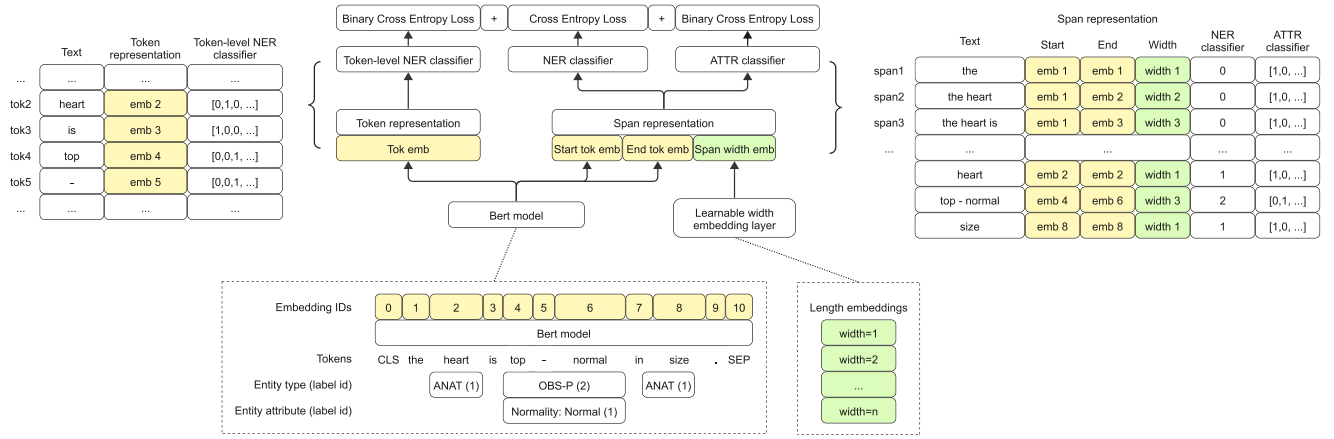


FIGURE 3. An overview of the named entity recognition model.

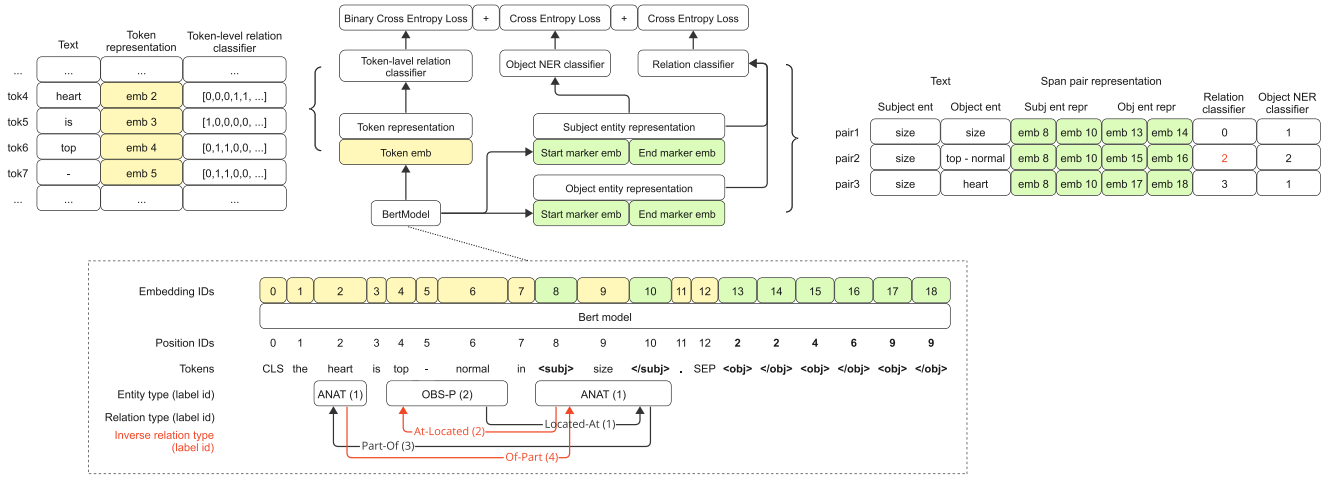


FIGURE 4. An overview of the relation extraction model.

not trivial inconsistency in the original RadGraph annotations. Given that only 45 pilot reports were utilized for training the annotators and refining the scheme for RadGraph, we attribute such inconsistency primarily to the difficulty in designing an annotation scheme to cover the diversity of natural language expressions thoroughly. It results in the differences in understanding between different annotators on how the scheme is applied to uncommon samples.

We, therefore, asked two annotators to jointly annotate the whole dataset, resolve any disagreements through discussion, and subsequently improve the scheme. The annotated data were revised if unseen patterns of disagreement emerged. Specifically, we simultaneously presented the original report with RadGraph labels to the annotators. One of the annotators was asked to annotate the raw report using our CXRGraph scheme, while the other annotator reviewed the newly annotated results and compared them with the RadGraph results. Since CXRGraph inherits the definitions of entity and relation labels from RadGraph, we expect the inherited labels to be consistent in the two schemes. During the annotation process, we observed that some of the original RadGraph entity and relationship labelswere inconsistently assigned to the same entity. For example, the entity "border" in "right heart border" has both instances labelled as OBS and instances labelled as ANAT; some spatial entities (e.g. area, region, zone, field, part) have instances inconsistently labelled as OBS, ANAT, or None. When the two annotators disagreed or an unseen inconsistency pattern was observed between our results and RadGraph, we would review the entire dataset, discuss, and update our annotation rules on specific terminologies, and then re-review and modify all previously annotated data.

To quantify the differences between the two annotation schemes, Table 1 shows the IAA between annotations in CXRGraph and RadGraph, respectively. To be able to compare the annotations, we converted CXRGraph annotations into a format compatible with RadGraph.

TABLE 1. The inter-annotator agreement between CXRGraph and RadGraph annotations.

Subset	Size	Entity type F1	Relation type F1
MIMIC-train	425	92.363	77.085
MIMIC-dev	75	92.270	77.306
MIMIC-test (labeler1)	50	95.991	83.753
MIMIC-test (labeler2)	50	96.674	85.024
CheXpert-test (labeler1)	50	91.110	71.479
CheXpert-test (labeler2)	50	90.891	73.134

The newly annotated data were used to train a joint NER and RE model whose details will be discussed in the next section. The newly trained model was used to automatically annotate 227,835 reports from MIMIC-CXR. Table 2 shows the distribution of data across entity type, entity attribute, and relation type. Conspicuously, during the annotation process, no entity was identified that could be clearly defined as Observation-Absent with its Normality attribute annotated as Abnormal. In addition, certain annotations exhibited low F1 scores, including the entity types ''Observation-Uncertain'' (78.3%) and ''Location-Attribute'' (78.57%), the entity attribute ''Change: Negative'' (73.7%), and the relation type ''Suggestive-Of'' (68.7%). The ''Observation-Uncertain'' demonstrated an 8 percent point increase in the F1 score compared to the F1 score of 70% reported in RadGraph, which can be attributed to the manual data curation. However, the prediction performance for ''Suggestive-Of'' labels remains similar to that in RadGraph (F1=68.5%). We assume the reason is related to the relatively long-range context between the subject and object entities of this relation type. Insufficient training samples could also contribute to the sub-optimal prediction performance of these labels.

B. MODEL SETUPS

This section provides details of models used to perform NER and RE. This includes both our own model and the baselines. Here, we specify the datasets used for training and evaluation, metrics used for evaluation and the large language models (LLMs) used to encode text. Finally, we conclude this section by providing specific implementation details of our own model.

1) BASELINE MODELS

The related work section describes these models in greater detail. Here, we briefly re-iterate these models. DYGIE++, originally designed for diverse domains, was employed in the RadGraph dataset. HGIE, an enhancement of DYGIE++, was tailored for a new radiology dataset with hierarchical structures. PURE and PL-Marker, pipeline models designed for diverse domains, are the foundation for our model. SpERT.PL and TriFM, are end-to-end models that demonstrate performance consistency with State-of-the-Art (SOTA) results.

2) DATASETS

We evaluated our model on four datasets having NER and RE labels. Two of these datasets are not directly related to our application domain. Specifically, these datasets include ACE05 [38], which contains 500 documents collected from various sources from the general domain including news, weblogs, conversation, etc., and SciERC [39], which contains 500 scientific abstracts collected from AI conference/workshop proceedings. The main reason for using these data is to evaluate the generalisability of the models considered. Finally, to compare the performance on chest radiology reports, we used RadGraph and CXRGraph datasets. Both are composed of the same 600 radiology reports collected from MIMIC-CXR and CheXpert, but annotated against two different annotation schemes. We split these datasets into training, validation and test subsets following previous works [2], [33], [34].

3) METRICS

We used a micro-averaged F1 score to evaluate the performance of the model [33], [40]. For NER, a predicted entity span is considered correct if its boundaries and entity type are correct. Entity attribute recognition (ATTR) was evaluated using the same criteria. For RE, a predicted relation is considered correct if the boundaries of two entity spans and the relation type are correct. Additionally, there is a strict evaluation (RE+) that considers a predicted relation as correct if the boundaries and entity types of two entity spans are correct and the relation type is correct. For the RadGraph dataset, we computed the average F1 score between its two labeler subsets.

4) LARGE LANGUAGE MODELS

The NER model uses BERT (bert-base-uncased) [41] as the encoder for ACE05; SciBERT (scibert-scivocab-uncased) [42] as the encoder for SciERC; and PubMedBERT (biomedbert-base-uncased) [43] for RadGraph and CXRGraph. The RE model uses the corresponding fine-tuned encoder obtained from the NER model for each task. PubMedBERT is identified as one of the top-performing encoders on the RadGraph dataset [2], [7]. As shown in Table 3, our experiment compares it against other biomedical variants of BERT [41], [44], [45], [46] using both micro- and macro-F1 score, and come to the same conclusion on the CXRGraph dataset.

5) IMPLEMENTATION DETAILS

For the NER model, we extended the input sentence with its context to a maximum of 512. The maximum length of the span was set to 8. The size of the length embedding layer was set to 150. We applied a weight decay of 0.01 to the entity encoder. For the RE model, we set the maximum length of the extended sentence to 256. Both models were trained with the AdamW optimizer of a linear scheduler with a warmup ratio of 0.1. A dropout of 0.1 was applied to the outputs of the model encoders. All FFNNs in the models had only one layer. The attribute classifier on the NER model was activated only on the CXRGraph dataset. In all experiments, the NER

TABLE 2. Annotation statistics and the breakdowns of inference results.

Labels	Train	Dev	Test (MIMIC)	Test (CheXpert)	Inference	Inference F1 on Test (MIMIC)	
						Micro	Macro
Total entities	12,765	2,272	1,352	1,510	6,617,806	96.62	90.44
Anatomy	5,399	977	530	662	2,789,834	98.78	*
Observation-Present	5,133	859	511	605	2,564,969	96.21	*
Observation-Absent	1,369	281	251	162	828,680	97.23	*
Observation-Uncertain	693	98	46	66	339,338	78.26	*
Location-Attribute	171	57	14	15	94,985	78.57	*
Total entity attributes	2,304	399	222	241	1,196,000	94.04	92.35
Total Normality	417	68	63	58	268,956	95.31	/
Normality: Normal	417	68	63	58	268,956	95.31	*
Normality: Abnormal	0	0	0	0	0	/	/
Total Action	969	174	84	95	480,864	93.25	/
Action: Removable	803	149	64	89	403,518	94.40	*
Action: Essential	166	25	20	6	77,346	89.47	*
Total Change	918	157	75	88	446,180	93.79	/
Change: Positive	164	31	11	10	81,411	95.24	*
Change: Negative	179	38	10	13	87,610	73.68	*
Change: Unchanged	575	88	54	65	277,159	97.14	*
Total Relations	9,622	1,717	987	1,151	4,811,916	89.50	86.10
Modify	4,637	802	457	565	2,245,105	91.74	*
Located-At	3,242	596	365	353	1,641,878	87.06	*
Suggestive-Of	439	66	36	57	183,072	68.66	*
Part-Of	1,304	253	129	176	741,861	93.80	*
Total reports	425	75	50	50	227,835	/	/
Total sentences	3,720	675	392	618	1,954,717	/	/
Average sentences per Report	8.8	9	7.8	12.4	8.6	/	/
Annotated sentences	2,571	465	289	283	1,372,803	/	/
Average annotated sentences per report	6.1	6.2	5.8	5.7	6	/	/
Total tokens	46,554	8,033	4,814	7,390	23,996,770	/	/
Average tokens per report	109.54	107.1	96.3	147.8	105.3	/	/
Annotated tokens	13,141	2,372	1,373	1,561	6,740,235	/	/
Average annotated tokens per report	30.9	30.3	27	30.2	29.6	/	/

The asterisk indicates the classes we used to calculate the macro-F1 scores for the entity, entity attribute and relation categories.

TABLE 3. Macro-F1 score.

	NER		ATTR		RE+	
	MIMIC	CheXpert	MIMIC	CheXpert	MIMIC	CheXpert
BERT [41]	95.52	94.57	91.28	88.56	87.27	84.68
BioBERT [44]	95.96	94.95	90.53	89.08	88.17	84.84
ClinicalBERT [45]	96.18	94.65	92.94	**89.87**	88.03	83.95
PubMedBERT [43]	**96.62**	**96.06**	**94.04**	89.79	**89.50**	**86.64**
BlueBERT [46]	95.93	93.55	91.63	86.90	88.03	83.30

(a) Micro-F1 score

	NER		ATTR		RE+	
	MIMIC	CheXpert	MIMIC	CheXpert	MIMIC	CheXpert
BERT [41]	89.26	88.74	88.42	82.46	82.86	79.85
BioBERT [44]	88.88	90.01	87.92	82.22	82.92	80.20
ClinicalBERT [45]	90.01	89.19	88.71	**84.50**	84.41	78.78
PubMedBERT [43]	90.44	**90.57**	**92.35**	84.49	**86.10**	**80.38**
BlueBERT [46]	**90.86**	87.44	89.30	79.97	85.97	77.96

(b) Macro-F1 score

model was trained for 100 epochs using a learning rate of 1e-5 for the encoder and 5e-4 for classifiers, with a batch size of 16. The RE model was trained for 20 epochs with a learning rate of 2e-5, a batch size of 8, and a gradient clipping set to 1.

C. MODEL RESULTS

We compared our results to the baselines results achieved by other relevant approaches as illustrated in Table 4. For each task, we present results from models that utilized the same encoder. The performance of our own approach is aligned with the State-of-the-Art (SOTA) results on the ACE05, SciERC datasets, although it underperforms compared to HGIE [7] on the RadGraph dataset. On the other hand, on the CXRGraph dataset, our approach achieves good performance on MIMIC-CXR reports with F1 scores of 96.62%, 94.04%, and 89.50% in NER, ATTR, and RE, respectively. Our approach also demonstrates feasible generalization performance on CheXpert reports, achieving F1 scores of 96.06%, 89.79%, and 86.64% in NER, ATTR, and RE, respectively. However, on RadGraph, our approach shows noticeable performance drops compared to the results on CXRGraph (MIMIC-CXR/CheXpert), with F1 score decreasing 2.3/4.8 percent points on NER and 6.2/13.3 percent points on RE. We attribute this gap to the unsolved annotation disagreements in RadGraph, which may confuse the model during training.

To evaluate the impact of our token-level auxiliary component, we conducted a series of ablation experiments across multiple datasets. As in Table 5, when the component was removed from the entity model, it negatively affected the results of both NER and RE tasks, such as a decrease of 2% and 5% micro F1 scores on the SciERC dataset in NER and RE+, respectively. However, if the component had been removed from the entity model, continuously removing it from the relation model did not show notable impacts on the model's performance in RE+. We assume the drop in RE+ scores was mainly attributed to error propagation

TABLE 4. The micro-averaged F1 scores on the test data.

	Ace05	SciERC	RadGraph		CXRGraph	
			MIMIC	CheXpert	MIMIC	CheXpert
DYGIE++ [30]						
NER	88.60	67.50	94.00	90.50	/	/
RE	63.40	48.40	/	/	/	/
RE+	/	/	82.30	72.50	/	/
PURE [33]						
NER	90.10	68.90	/	/	/	/
RE	67.70	50.10	/	/	/	/
RE+	64.80	36.80	81.20	72.90	/	/
HGIE [7]						
NER	/	/	/	/	/	/
RE	/	/	/	/	/	/
RE+	/	/	**84.90**	**75.20**	/	/
PL-Marker [34]						
NER	89.80	69.90	/	/	/	/
RE	69.00*	**53.20***	/	/	/	/
RE+	66.50*	41.60*	/	/	/	/
TriMF [32]						
NER	87.61	70.17	/	/	/	/
RE	66.49	52.44	/	/	/	/
RE+	62.77	/	/	/	/	/
SpERT.PL [31]						
NER	/	70.53	/	/	/	/
RE	/	51.25	/	/	/	/
RE+	/	/	/	/	/	/
Ours						
NER	**90.30**	**70.70**	**94.36**	**91.23**	96.62	96.06
ATTR	/	/	/	/	94.04	89.79
RE	**69.89**	52.90	85.20	75.37	92.00	88.25
RE+	**67.14**	**43.71**	83.27	73.30	89.51	86.64

The * mark indicates a probable overestimation of model performance since the PL-Marker regards each symmetric relational instance as two data instances during evaluation.

stemming from inaccuracies in the NER output. We also notice that the component is less prominent in a more complex model. For instance, the overall performance of the CXRGraph model, which had an additional task of entity attribute recognition, showed an insignificant performance decrease with the removal of the component compared to models trained on other datasets.

D. MODEL VARIANTS

Considering the variations across the test datasets, particularly those containing shorter sentences, our vanilla NER model may incur time costs as it treats each sentence as a separate data instance. To address this issue, we introduced a variant of the NER model, denoted as Ent-doc, which considers the entire document as a single data instance. In cases where the document length exceeds the maximum sequence length of 512, we partitioned the document into multiple data instances using a length threshold of 384. Each document segment overlaps with its adjacent segments by 128 tokens. As shown in Table 6, the Ent-doc model accelerated the training and evaluation process by 1.9x to 8.7x compared to the vanilla NER model. However, this efficiency gain comes at the expense of a slight decrease in performance.

We observed a small percentage of cross-sentence relations within the RadGraph (0.74%) and CXRGraph (0.7%) datasets as illustrated in Table 7, where the subject and object entities are located in different sentences. However, our vanilla RE model was designed to extract the relation of entities within the same sentence, making it unable to capture cross-sentence relations. To address this limitation, we introduced a variant of the RE model, denoted as Rel-cross, which extends the range of target sentences for obtaining candidate object entities. Specifically, we extend the range from one sentence to include the two directly adjacent sentences. The results presented in Table 8 indicate a similar performance between the Rel-cross and the vanilla RE model. Given the extremely rare occurrence of cross-sentence relations, we suggest ignoring the cross-sentence relations to simplify the RE task without significantly impacting the overall performance.

V. DISCUSSION

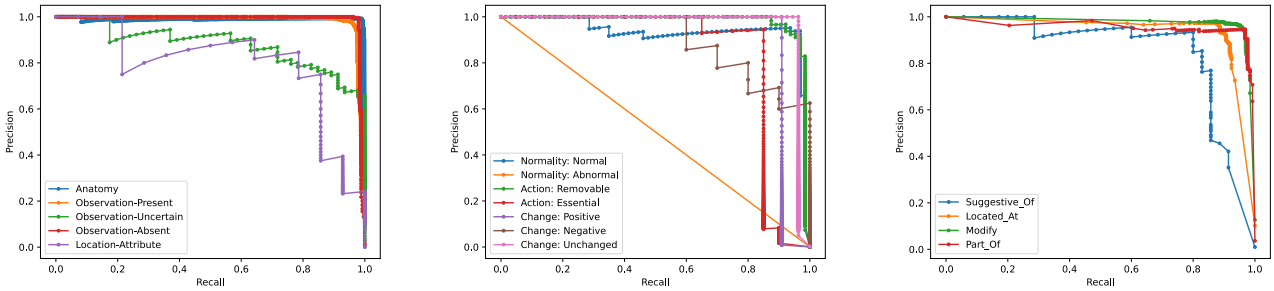
We employed precision-recall curves to break down the model predictive results on CXRGraph to provide a more detailed understanding of the model performance, as shown in Fig. 5. Note that the PR curve is drawn by calculating the precision and recall of a class under different thresholds of prediction probability. Each point on the curve represents a precision-recall value pair of a specific probability threshold. However, the PR curve is designed for binary classification. Therefore, it might not be perfectly precise in a multi-class classification scenario. We adapted it by taking the target class as positive and all other classes as negative. In Fig. 5, the curves of each class are consistent with their F1 scores illustrated in Table 2. For example, in Fig. 5a, the F1 scores of Anatomy, Observation-Present, and Observation-Absent were above 95%, while the micro-F1 scores of Observation-Uncertain and Location-Attribute were only about 78%. In Fig. 5b, there is a diagonal line revealing that the Normality: Abnormal class has no data instance in the CXRGraph dataset. Furthermore, some curves exhibit a recurring pattern of an abrupt decline followed by a gradual rise, such as Suggestive-Of. This phenomenon likely arises when a subset of negative samples with specific characteristics is predicted to have similar probabilities. Once the threshold drops to a certain value, these negative samples are misclassified as positive, resulting in a significant decrease in precision while recall remains unaffected.

During the annotation process, we observed different annotation patterns in RadGraph. We incorporated the most intuitive and interpretable patterns into our scheme. First, some neutral observation modifiers such as pre-existing, known, previous, and old were not labelled consistently in RadGraph. Some of them were labelled as OBS, whereas others were not. In our scheme, we labelled them all as OBS with Action: Removable attribute, making these entities easier to process afterwards as required. Second, some relations between OBS entities involving general terms such as changes, disease, and process were labelled in a different pattern. For example, "mild atelectatic changes" had been either labelled as (mild, Modify, changes), (changes, Modify, atelectatic) or (mild, Modify, changes), (atelectatic, Modify, changes). We adopted the second pattern as it

TABLE 5. Model ablation.

Model		SciERC		RadGraph		CXRGraph		
		NER	RE+	NER	RE+	NER	ATTR	RE+
Ent	Rel	**70.69**	**43.70**	**94.35**	**83.67**	**96.47**	**91.57**	**88.01**
	Rel w/o TA		41.60 (-2.10)		82.56 (-1.11)			87.92 (-0.09)
Ent w/o TA	Rel	68.58 (-2.11)	38.60 (-5.10)	93.69 (-0.66)	81.96 (-1.71)	96.19 (-0.28)	91.07 (-0.50)	87.55 (-0.46)
	Rel w/o TA		38.15 (-5.55)		81.91 (-1.76)			87.69 (-0.32)

TA indicates the auxiliary task of token-level classification. w/o TA indicate the token-level auxiliary component is removed from the model. For CXRGraph, the results are mean scores of the MIMIC and CheXpert test subsets.



(a) PR curves on entity classes (b) PR curves on entity attribute classes (c) PR curves on relation classes

FIGURE 5. Precision-recall (PR) curves of our model prediction on CXRGraph test (MIMIC).

TABLE 6. Comparison of NER model variants with respect to the micro-averaged F1 scores and time costs.

Dataset	Model	NER	ATTR	Time (mins)	Average sentences per document
SciERC	Ent	70.70	/	56	5.4
	Ent-doc	70.22	/	30	
ACE05	Ent	90.30	/	567	28.4
	Ent-doc	89.22	/	65	
RadGraph	Ent	92.33	/	117	9.0
	Ent-doc	91.82	/	36	
CXRGraph	Ent	96.32	91.83	119	9.0
	Ent-doc	96.29	91.41	58	

For RadGraph and CXRGraph, the evaluation dataset contains both the MIMIC-CXR and CheXpert reports. The Time column measures the overall time cost of model training and evaluation.

TABLE 7. Distribution of relations across sentences.

Dataset	Data split	K=0	K=1	K=2	K=3
RadGraph	Train	9192	59		
	Dev	1626	60		
	Test (labeler1)	2012	61	1	1
	Test (labeler2)	1959	62	2	
CXRGraph	Train	9556	63		
	Dev	1706	64		
	Test	2122	65	1	

K denotes the distance between the sentences that contain the subject and object entities respectively. K=0 indicates that both subject and object entities are within the same sentence.

TABLE 8. Comparison of relation model variants regarding micro-averaged F1 scores and time costs.

Dataset	Model	RE	RE+	Time (mins)
RadGraph	Rel	79.53	77.26	100
	Rel-cross	79.70	77.29	106.00
CXRGraph	Rel	89.98	87.96	102.00
	Rel-cross	89.66	87.71	108.00

The evaluation dataset contains both the MIMIC-CXR and CheXpert reports. The Time column measures the overall time cost of model training and evaluation.

appeared more prevalent. Third, we observed various disagreements related to the modifiers that are part of an entity or describing its characteristic/attribute, including size, amount, degree, appearance, pattern, volume, position, place, course, area, region, margin, junction, body, border, angle, field, portion, level, part, silhouette, contours, and so on. For example, we have seen annotations of ((appearance, OBS), Located-At, (chest, ANAT)) in "overall appearance of the chest", ((appearance, ANAT), Modify, (cardiac, ANAT)) in "normal appearance of the cardiac silhouette", and unlabelled "appearance" in similar phrases. To distinguish them from other modifiers (e.g. mild and left), we considered them as entity modifiers and introduced a Part-Of relation to link them with the entities they modify. Consequently, one of the above instances was annotated as ((appearance, ANAT), Part-Of, (chest, ANAT)) on our scheme. In addition to the above common modifiers, specialized terms such as silhouette, contours, venous, vascular, vasculature, alveolar, arteries, perihilar, sinus, parenchymal, interstitial, knob, arch, and valve have similar issues. We addressed these in the same manner; for instance, the phrase "appearance of the cardiac silhouette" was actually annotated as ((appearance, ANAT), Part-Of, (silhouette, ANAT)) and ((silhouette, ANAT), Part-Of, (cardiac, ANAT)) according to our scheme. Fourth, our annotation scheme differs from the RadGraph preference in

terms of external devices such as tube, line, catheter, clip, sheath, device and PICC. For example, in phrases such as ''right internal jugular line'', RadGraph preferred ''internal'' and ''jugular'' to be OBS entities. However, we observed that the modifier ''right'' was mainly annotated as an ANAT entity in such phrases, making it to be an isolated pattern and hard to follow. To make it more consistent with general cases, we considered ''internal'' and ''jugular'' to be ANAT entities.

There are several limitations of this work. First, the annotations are limited to chest X-ray radiology reports. This study paves the way to exploring the impact of structured data on the ARRG model, aiming to obtain accurate structured data. However, different radiology examinations are somewhat different in terms of their images and reports. Therefore, we are trying to simplify the problem by focusing on one radiology examination first to develop a proof of concept. According to our recent review, most current ARRG models have been trained and evaluated on the CXR dataset. As a result, we chose to focus on CXR to enable our subsequent work to be compared with the existing ARRG models. In addition, we also found that CXR has the largest publicly available image-report dataset, MIMIC-CXR, which further explains why CXR is the primary data source for current ARRG studies. Second, considering our reference scheme, RadGraph, was annotated by three board-certified radiologists, we employed only one cross-disciplinary expert with experience in clinical data processing and one expert with a medical background to speed up the collaborative annotation process. During the annotation process, the RadGraph results were provided as references whereas radiology images were not consulted. If we were unsure of the label for an entity or relation, we adopted the label from RadGraph. Third, the model prediction on entity type ''Observation-Uncertain'' (F1=78.3%) and relation type ''Suggestive-Of'' (F1=68.6%) performs worse than that on overall entity types (F1=96.6%) and relation types (F1=89.5%). We assume this can be attributed to the expression bias of natural language in the original report. Moreover, the annotation inconsistencies of these data are more perceptually noticeable than those of other data in RadGraph, which indicates that identifying these data is possibly more challenging than other data even for human experts. Additionally, reasons regarding the long-range context between the subject and object entities of this relation type and the insufficient training samples cannot be excluded. Fourth, as stated in RadGraph, the annotations do not capture the clinical context in a radiology report, such as the information stated in the Comparison or History section of the report. Fifth, our CXRGraph scheme was designed based on the reports from MIMIC-CXR, and our model was trained MIMIC-CXR. Although we evaluated the generalization of our model on CheXpert, it might not represent the diversity of radiology reports from different hospitals or regions. Researchers should be aware of the limitations when applying our method to other datasets. Finally, although we annotated the same data as RadGraph, the downloaded reports were de-identified, readers are encouraged to refer to the RadGraph dataset [2] for the demographic breakdowns regarding sex, age, and race, and potential bias in data selection.

VI. CONCLUSION

We described a method to transform the training text data into a format that simplifies the process of training deep learning models for ARRG. This method proposes a scheme designed to assist researchers in structuring narrative radiology reports, thereby mitigating the data quality issue caused by the inherent diversity of natural language, and hallucinated references, among others. It also includes a model to automatically populate such scheme with data extracted from the narrative reports. In order to train this model, we first manually annotated a ground-truth dataset consisting of 550 radiology reports from MIMIC-CXR and 50 radiology reports from CheXpert. Using this dataset, we designed and trained a joint NER and RE model, which achieved prediction performance of micro-averaged F1=96.6% (NER), 94.0% (ATTR) and 89.5% (RE) on the MIMIC-CXR dataset and achieved performance of micro-averaged F1=96.1% (NER), 89.8% (ATTR) and 86.6% (RE) on the CheXpert dataset. Utilizing the model, we automatically annotated an inference dataset consisting of 227,835 MIMIC-CXR reports. Our model also demonstrated performance in line with the state-of-the-art results in various domains within the joint NER and RE tasks. We release the ground-truth and inference datasets on PhysioNet [47]. The trained models and corresponding codes are available on GitHub [48]. Our future work will leverage the research results to verify the effectiveness of employing structured reports in ARRG while also enhancing the practicality and applicability of our method in medical practices.

REFERENCES

[1] Y. Liao, H. Liu, and I. Spasić, ''Deep learning approaches to automatic radiology report generation: A systematic review,'' *Informat. Med. Unlocked*, vol. 39, Jan. 2023, Art. no. 101273. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S235291482300117X

[2] S. Jain, A. Agrawal, A. Saporta, S. Truong, D. N. D. N. Duong, T. Bui, P. Chambon, Y. Zhang, M. Lungren, A. Ng, C. Langlotz, P. Rajpurkar, and P. Rajpurkar, ''Radgraph: Extracting clinical entities and relations from radiology reports,'' in *Proc. Neural Inf. Process. Syst. Track Datasets Benchmarks*, vol. 1, J. Vanschoren and S. Yeung, Eds., 2021. [Online]. Available: https://datasets-benchmarksproceedings.neurips.cc/paper_files/paper/2021

[3] I. Spasić, B. Zhao, C. B. Jones, and K. Button, ''KneeTex: An ontology–driven system for information extraction from MRI reports,'' *J. Biomed. Semantics*, vol. 6, no. 1, p. 34, Dec. 2015.

[4] K. Sugimoto, T. Takeda, J.-H. Oh, S. Wada, S. Konishi, A. Yamahata, S. Manabe, N. Tomiyama, T. Matsunaga, K. Nakanishi, and Y. Matsumura, ''Extracting clinical terms from radiology reports with deep learning,'' *J. Biomed. Informat.*, vol. 116, Apr. 2021, Art. no. 103729. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046421000587

[5] S. Hassanpour and C. P. Langlotz, ''Information extraction from multi-institutional radiology reports,'' *Artif. Intell. Med.*, vol. 66, pp. 29–39, Jan. 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365715001244

[6] S. Datta, M. Ulinski, J. Godfrey-Stovall, S. Khanpara, R. F. Riascos-Castaneda, and K. Roberts, "Rad-SpatialNet: A frame-based resource for fine-grained spatial relations in radiology reports," in *Proc. 12th Lang. Resour. Eval. Conf.*, N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 2251–2260. [Online]. Available: https://aclanthology.org/2020.lrec-1.274

[7] S. Khanna, A. Dejl, K. Yoon, S. Q. Truong, H. Duong, A. Saenz, and P. Rajpurkar, "RadGraph2: Modeling disease progression in radiology reports via hierarchical information extraction," in *Proc. 8th Mach. Learn. Healthcare Conf.*, vol. 219, K. Deshpande, M. Fiterau, S. Joshi, Z. Lipton, R. Ranganath, I. Urteaga, and S. Yeung, Eds., 2023, pp. 381–402. [Online]. Available: https://proceedings.mlr.press/v219/khanna23a/khanna23a.pdf

[8] A. E. W. Johnson, T. J. Pollard, S. J. Berkowitz, N. R. Greenbaum, M. P. Lungren, C.-Y. Deng, R. G. Mark, and S. Horng, "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports," *Sci. Data*, vol. 6, no. 1, p. 317, Dec. 2019.

[9] V. Ramesh, N. A. Chi, and P. Rajpurkar, "Improving radiology report generation systems by removing hallucinated references to non-existent priors," in *Proc. 2nd Mach. Learn. Health Symp.*, vol. 193, 2022, pp. 456–473.

[10] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. 33rd AAAI Conf. Artif. Intell. 31st Innov. Appl. Artif. Intell. Conf. 9th AAAI Symp. Educ. Adv. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 590–597, doi: 10.1609/aaai.v33i01.3301590. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/issue/view/246

[11] P. Harzig, Y.-Y. Chen, F. Chen, and R. Lienhart, "Addressing data bias problems for chest X-ray image report generation," in *Proc. Brit. Mach. Vis. Conf.*, 2019, p. 144. [Online]. Available: https://dblp.org/rec/conf/bmvc/HarzigCCL19.html?view=bibtex

[12] X. Wang, Y. Peng, L. Lu, Z. Lu, and R. M. Summers, "TieNet: Text-image embedding network for common thorax disease classification and reporting in chest X-Rays," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9049–9058.

[13] L. Sun, W. Wang, J. Li, and J. Lin, "Study on medical image report generation based on improved encoding-decoding method," in *Intelligent Computing Theories and Application*, D.-S. Huang, V. Bevilacqua, and P. Premaratne, Eds., Cham, Switzerland: Springer, 2019, pp. 686–696.

[14] C. Yin, B. Qian, J. Wei, X. Li, X. Zhang, Y. Li, and Q. Zheng, "Automatic generation of medical imaging diagnostic report with hierarchical recurrent neural network," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 728–737.

[15] J. Yuan, H. Liao, R. Luo, and J. Luo, "Automatic radiology report generation based on multi-view image fusion and medical concept enrichment," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2019*, D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, and A. Khan, Eds., Cham, Switzerland: Springer, 2019, pp. 721–729.

[16] J. G. Mork, A. Jimeno-Yepes, and A. R. Aronson, "The NLM medical text indexer system for indexing biomedical literature," in *Proc. BioASQ@CLEF*, 2013. [Online]. Available: https://ceur-ws.org/Vol-1094/

[17] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3462–3471.

[18] Y. Zhang, X. Wang, Z. Xu, Q. Yu, A. L. Yuille, and D. Xu, "When radiology report generation meets knowledge graph," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34. Palo Alto, CA, USA: AAAI Press, 2020, pp. 12910–12917. [Online]. Available: https://aaai.org/proceeding/vol-34-no-07-aaai-20-technical-tracks-7/

[19] C. Y. Li, X. Liang, Z. Hu, and E. P. Xing, "Knowledge-driven encode, retrieve, paraphrase for medical image report generation," in *Proc. 33rd AAAI Conf. Artif. Intell. 31st Innov. Appl. Artif. Intell. Conf. 9th AAAI Symp. Educ. Adv. Artif. Intell.* Palo Alto, CA, USA: AAAI Press, 2019, pp. 6666–6673, doi: 10.1609/aaai.v33i01.33016666. [Online]. Available: https://aaai.org/proceeding/01-aaai-19-iaai-19-eaai-20/

[20] F. Nooralahzadeh, N. Perez Gonzalez, T. Frauenfelder, K. Fujimoto, and M. Krauthammer, "Progressive transformer-based generation of radiology reports," in *Proc. Findings Assoc. Comput. Linguistic*, M.-F. Moens, X. Huang, L. Specia, and S. W.-T. Yih, Eds. Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 2824–2832. [Online]. Available: https://aclanthology.org/2021.findings-emnlp.241

[21] S. Yan, W. K. Cheung, K. Chiu, T. M. Tong, K. C. Cheung, and S. See, "Attributed abnormality graph embedding for clinically accurate X-ray report generation," *IEEE Trans. Med. Imag.*, vol. 42, no. 8, pp. 2211–2222, Aug. 2023.

[22] Y. Xiong, J. Liu, K. Zaripova, S. Sharifzadeh, M. Keicher, and N. Navab, "Prior-RadGraphFormer: Prior-knowledge-enhanced transformer for generating radiology graphs from X-rays," in *Proc. 5th MICCAI Workshop, GRAIL 1st MICCAI Challenge*, Vancouver, BC, Canada. Berlin, Germany: Springer-Verlag, Sep. 23, 2023, pp. 54–63, doi: 10.1007/978-3-031-55088-1_5.

[23] Z. Wang, M. Tang, L. Wang, X. Li, and L. Zhou, "A medical semantic-assisted transformer for radiographic report generation," in *Medical Image Computing and Computer Assisted Intervention—MICCAI 2022*, L. Wang, Q. Dou, P. T. Fletcher, S. Speidel, and S. Li, Eds., Cham, Switzerland: Springer, 2022, pp. 655–664.

[24] S. Yang, X. Wu, S. Ge, S. K. Zhou, and L. Xiao, "Knowledge matters: Chest radiology report generation with general and specific knowledge," *Med. Image Anal.*, vol. 80, Aug. 2022, Art. no. 102510. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1361841522001578

[25] M. Li, B. Lin, Z. Chen, H. Lin, X. Liang, and X. Chang, "Dynamic graph enhanced contrastive learning for chest X-ray report generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 3334–3343.

[26] J. M. Steinkamp, C. Chambers, D. Lalevic, H. M. Zafar, and T. S. Cook, "Toward complete structured information extraction from radiology reports using machine learning," *J. Digit. Imag.*, vol. 32, no. 4, pp. 554–564, Aug. 2019, doi: 10.1007/s10278-019-00234-y.

[27] S. Datta, Y. Si, L. Rodriguez, S. E. Shooshan, D. Demner-Fushman, and K. Roberts, "Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest X-ray reports using deep learning," *J. Biomed. Informat.*, vol. 108, Aug. 2020, Art. no. 103473. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1532046420301027

[28] S. Datta and K. Roberts, "A dataset of chest X-ray reports annotated with spatial role labeling annotations," *Data Brief*, vol. 32, Oct. 2020, Art. no. 106056. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352340920309501

[29] C. P. Langlotz, "RadLex: A new method for indexing online educational materials," *RadioGraphics*, vol. 26, no. 6, pp. 1595–1597, Nov. 2006, doi: 10.1148/rg.266065168.

[30] D. Wadden, U. Wennberg, Y. Luan, and H. Hajishirzi, "Entity, relation, and event extraction with contextualized span representations," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong: Association for Computational Linguistics, 2019, pp. 5784–5789. [Online]. Available: https://aclanthology.org/D19-1585

[31] T. Santosh, P. Chakraborty, S. Dutta, D. K. Sanyal, and P. P. Das, "Joint entity and relation extraction from scientific documents: Role of linguistic information and entity types," in *Proc. 2nd Workshop Extraction Eval. Knowl. Entities Sci. Documents (EEKE)*, vol. 3004, 2021, pp. 15–19. [Online]. Available: https://ceur-ws.org/Vol-3004/

[32] Y. Shen, X. Ma, Y. Tang, and W. Lu, "A trigger-sense memory flow framework for joint entity and relation extraction," in *Proc. Web Conf.* New York, NY, USA: Association for Computing Machinery, Apr. 2021, pp. 1704–1715, doi: 10.1145/3442381.3449895.

[33] Z. Zhong and D. Chen, "A frustratingly easy approach for entity and relation extraction," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, and Y. Zhou, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021, pp. 50–61. [Online]. Available: https://aclanthology.org/2021.naacl-main.5

[34] D. Ye, Y. Lin, P. Li, and M. Sun, "Packed levitated marker for entity and relation extraction," in *Proc. 60th Annu. Meeting Assoc. Comput. Linguistics*, S. Muresan, P. Nakov, and A. Villavicencio, Eds. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 4904–4917. [Online]. Available: https://aclanthology.org/2022.acl-long.337

[35] W. Wu, F. Wang, A. Yuan, F. Wu, and J. Li, "CorefQA: Coreference resolution as query-based span prediction," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020, pp. 6953–6963. [Online]. Available: https://aclanthology.org/2020.acl-main.622

[36] P. Stenetorp, S. Pyysalo, G. Topic, T. Ohta, S. Ananiadou, and J. Tsujii, "BRAT: A web-based tool for NLP-assisted text annotation," in *Proc. 13th Conf. Eur. Chapter Assoc. Comput. Linguistics*, F. Segond, Ed., Avignon, France: Association for Computational Linguistics, Apr. 2012, pp. 102–107. [Online]. Available: https://aclanthology.org/E12-2021

[37] C. Grouin, S. Rosset, P. Zweigenbaum, K. Fort, O. Galibert, and L. Quintard, "Proposal for an extension of traditional named entities: From guidelines to evaluation, an overview," in *Proc. 5th Linguistic Annotation Workshop*, N. Ide, A. Meyers, S. Pradhan, and K. Tomanek, Eds. Portland, OR, USA: Association for Computational Linguistics, Jun. 2011, pp. 92–100. [Online]. Available: https://aclanthology.org/W11-0411

[38] C. Walker, S. Strassel, J. Medero, and K. Maeda. (2005). *Ace 2005 Multilingual Training Corpus*. Linguistic Data Consortium, Philadelphia, PA, USA. Accessed: Feb. 29, 2024. [Online]. Available: https://catalog.ldc.upenn.edu/LDC2006T06

[39] Y. Luan, L. He, M. Ostendorf, and H. Hajishirzi, "Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 3219–3232. [Online]. Available: https://aclanthology.org/D18-1360

[40] B. Taillé, V. Guigue, G. Scoutheeten, and P. Gallinari, "Let's stop incorrect comparisons in end-to-end relation extraction!" in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Stroudsburg, PA, USA: Association for Computational Linguistics, Nov. 2020, pp. 3689–3701. [Online]. Available: https://aclanthology.org/2020.emnlp-main.301

[41] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, vol. 1, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

[42] I. Beltagy, K. Lo, and A. Cohan, "SciBERT: A pretrained language model for scientific text," in *Proc. Conf. Empirical Methods Natural Lang. Process. 9th Int. Joint Conf. Natural Lang. Process. (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong: Association for Computational Linguistics, 2019, pp. 3615–3620. [Online]. Available: https://aclanthology.org/D19-1371

[43] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. for Healthcare*, vol. 3, no. 1, pp. 1–23, Oct. 2021, doi: 10.1145/3458754.

[44] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020, doi: 10.1093/bioinformatics/btz682.

[45] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, and M. McDermott, "Publicly available clinical BERT embeddings," in *Proc. 2nd Clin. Natural Lang. Process. Workshop*, A. Rumshisky, K. Roberts, S. Bethard, and T. Naumann, Eds. Minneapolis, MI, USA: Association for Computational Linguistics, Jun. 2019, pp. 72–78. [Online]. Available: https://aclanthology.org/W19-1909

[46] Y. Peng, S. Yan, and Z. Lu, "Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets," in *Proc. 18th BioNLP Workshop Shared Task*, D. Demner-Fushman, K. B. Cohen, S. Ananiadou, and J. Tsujii, Eds. Florence, Italy: Association for Computational Linguistics, 2019, pp. 58–65. [Online]. Available: https://aclanthology.org/W19-5006

[47] Y. Liao, H. Xiang, H. Liu, and I. Spasic, "Cxrgraph: Using information extraction to normalize the training data for automatic radiology report generation," PhysioNet, MIT Lab. Comput. Physiol., Cambridge, MA, USA, Rep., 2024. [Online]. Available: https://physionet.org/about/

[48] Y. Liao. (2024). *Code for Cxrgraph*. Github. Accessed: Mar. 3, 2024. [Online]. Available: https://github.com/yxliao95/cxrgraph

YUXIANG LIAO received the M.S. degree in advanced computer science from the University of Leeds, U.K., in 2020. He is currently pursuing the Ph.D. degree with Cardiff University, U.K., with a focus on deep learning technology for automatic radiology report generation.

In particular, he is working on exploring AI-assisted radiology reporting system to reduce the burden on radiologists. His research interests include the cross-disciplinary application of deep learning methods in healthcare and clinical practice.

HAISHAN XIANG received the M.P.H. degree in epidemiology from Sun Yat-sen University, China, in 2019.

He was with Shenzhen Maternal and Child Healthcare Hospital. He is currently the Deputy Director of Baoan District AIDS and STD Professional Committee, Baoan Center for Disease Control and Prevention, Shenzhen, China. His research interests include artificial intelligence in radiology diagnosis, impact of environmental factors on male reproductive health and birth defects, and maternal and child healthcare. He is a member of Guangdong Provincial Health Statistics Society.

HANTAO LIU received the Ph.D. degree from Delft University of Technology, Delft, The Netherlands, in 2011.

He is currently a Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. His research interests include image processing, machine learning, computer vision, applied perception, and medical imaging.

IRENA SPASIĆ received the Ph.D. degree in computer science from the University of Salford, U.K., in 2004.

She worked in various universities, such as the University of Belgrade, the University of Salford, and The University of Manchester. In 2010, she joined the School of Computer Science and Informatics, Cardiff University, where she became a Full Professor, in 2016. She is also the Co-Founder of U.K. Healthcare Text Analytics Research Network. Her research interests include text mining, knowledge representation, machine learning, and information management with applications in healthcare, life sciences, and social sciences. In 2020, she was elected as a fellow of the Learned Society of Wales.