RadGraph2: Modeling Disease Progression in Radiology Reports via Hierarchical Information Extraction

Sameer Khanna*   SAMEER_KHANNA@BERKELEY.EDU
Stanford University

Adam Dejl*   ADAMDEJL@MIT.EDU
Massachusetts Institute of Technology

Kibo Yoon   KIBOYOONMD@GMAIL.COM
Hanyang University

Steven QH Truong   BRAIN01@VINBRAIN.NET
Vinbrain

Hanh Duong   V.HANHDUONG@VINBRAIN.NET
Vinbrain

Agustina Saenz†   ASAENZ@BWH.HARVARD.EDU
Harvard Medical School

Pranav Rajpurkar†   PRANAV_RAJPURKAR@HMS.HARVARD.EDU
Harvard Medical School

Abstract

We present RadGraph2, a novel dataset for extracting information from radiology reports that focuses on capturing changes in disease state and device placement over time. We introduce a hierarchical schema that organizes entities based on their relationships and show that using this hierarchy during training improves the performance of an information extraction model. Specifically, we propose a modification to the DyGIE++ framework, resulting in our model HGIE, which outperforms previous models in entity and relation extraction tasks. We demonstrate that RadGraph2 enables models to capture a wider variety of findings and perform better at relation extraction compared to those trained on the original RadGraph dataset. Our work provides the foundation for developing automated systems that can track disease progression over time and develop information extraction models that leverage the natural hierarchy of labels in the medical domain.

---

*. These authors contributed equally.
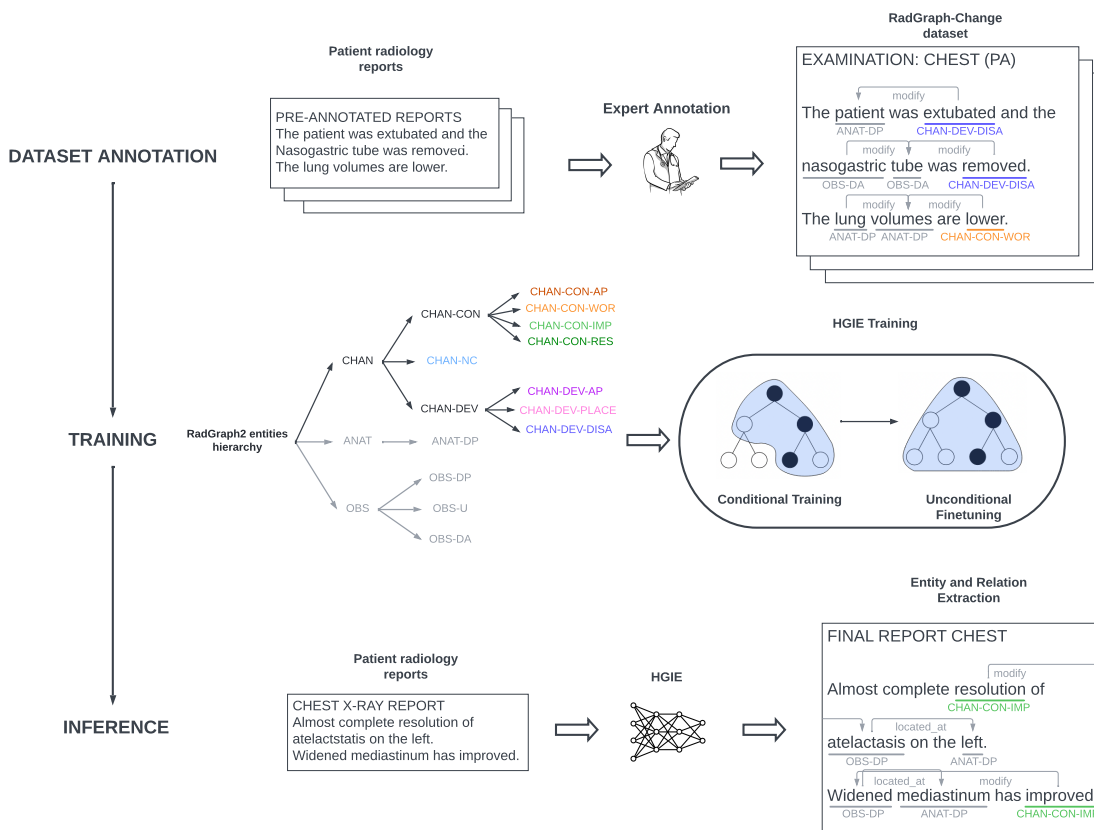
†. These authors contributed equally.

Figure 1: Overview of the HGIE training process and its application to a downstream task of change classification for radiology reports in the RadGraph2 dataset. Change entities, which have been added as part of our expanded information schema, are highlighted.

1. Introduction

Extracting and understanding information held within medical reports has become a central task in medical artificial intelligence, fueled by the creation of various medical report datasets. Recently, a push has been made away from conventional disease classification labels and labeling methods that can only mark the presence or absence of a narrow set of diseases (Irvin et al., 2019; McDermott et al., 2020; Smit et al., 2020). In its place, radiographs, graphs created from X-rays, are becoming more popular. They are capable of dynamically modeling rich contextual information by marking the level of presence and relations to other entities and modifications. They can also be used to support a wide range of clinical tasks. Recent work built upon the RadGraph dataset, the most widely-used medical report graph dataset, has shown that graphs built from medical reports are capable of facilitating a wide range of research in the fields of natural language processing, computer

vision, and multi-modal learning in various medical domains (Chiu et al., 2022; Li et al., 2022; Yang et al., 2022; McGrath et al., 2022; Wu et al., 2023; Stupp et al., 2022; Li et al., 2023; Yu et al., 2022; Jeong et al., 2023).

Nonetheless, RadGraph is still far from encapsulating all the information a patient's medical report contains; for example, it is missing key information regarding prior patient data. Like most past work on medical report interpretation, it is focused on detecting conditions in a single scan or instance. The schema utilized within the original dataset does not encapsulate the relations between various entities and only focuses on disease info rather than including information regarding disease progression and device placement or removal. One key component of clinical reports, such as those used in radiology, is comparison to prior patient data, which enables medical professionals to track disease progression over time. There has therefore been a strong push for AI tools that can not only interpret findings in a single study but can also integrate information from prior reports to provide a complete clinical picture (Acosta et al., 2022; Ramesh et al.; Moor et al., 2023; Rajpurkar and Lungren, 2023).

In this work, we construct RadGraph2, a new dataset of 800 expert-annotated chest X-ray reports and additional 220,913 reports annotated via inference. We are planning to make this data publicly available via the PhysioNet platform (Goldberger et al., 2000). RadGraph2 combines the schema from the previous version of RadGraph with additional entities focused on characterizing priors, like disease progression and device information, along with clinically relevant contextual information in radiology reports.

The new RadGraph2 schema introduces a taxonomy to the entities found within a chest X-ray report, taking advantage of their natural hierarchy. We show that capturing relationships between entities via a hierarchy can lead to increases in performance through our model HGIE, an information extraction model that modifies the DyGIE++ framework (Wadden et al., 2019). HGIE uses a two-phase training methodology in order to learn the entity hierarchy, achieving superior performance to DyGIE++ despite sharing the same model architecture.

In summary, our contributions are as follows:

- We devise a novel schema for capturing change along with clinically relevant contextual information in radiology reports.

- We introduce a hierarchy to the schema, organizing entities based on their natural relationships. We show that taking advantage of this inherent organization can lead to performance improvements.

- We extend the RadGraph dataset with additional expert-annotated reports, bringing the total number of X-ray reports densely annotated according to our schema to 800. This dataset is expanded via inference to include an additional 220,463 reports from MIMIC-CXR and 450 from CheXpert.

Generalizable Insights about Machine Learning in the Context of Healthcare

The automatic identification of changes in a patient's condition from clinical notes is a critical task in healthcare, as information about past disease progression is vital for medical

professionals. In this paper, we propose a hierarchical schema for modeling changes in radiology reports. We construct a new dataset for this schema and introduce a model trained to automatically identify changes based on the schema's hierarchy. Our approach enables automated systems to reason about the differences observed across multiple radiological studies, which may facilitate a variety of useful clinical applications, such as modeling patient healthcare trajectories and training large-scale medical imaging models. We also believe that our schema and general approach could be extended to other domains beyond radiology.

2. Related Work

In this work, we aim to extract key entities and their relations from radiology reports, building upon the work done by RadGraph. Various other natural language processing (NLP) approaches have been developed and used to extract information from medical reports and radiology reports (Irvin et al., 2019; Peng et al., 2018; McDermott et al., 2020; Smit et al., 2020; Jain et al., 2021b). The most popular approach uses automated radiology report labelers to label the reports within large-scale chest radiograph datasets like MIMIC-CXR and CheXpert (Johnson et al., 2019; Irvin et al., 2019) for a variety of common medical conditions. While these labels capture information about the presence of particular diseases and conditions, they do not capture more rich contextual information, such as the relations between the concepts mentioned in each report nor do they capture vital information regarding the presence of priors.

Information Extraction in the Medical Sphere RadGraph2 utilizes entity and relation schemas to extract detailed information from radiology reports. While there are approaches proposed that aim to achieve the same thing using entity extraction schemas (Hassanpour and Langlotz, 2016; Sugimoto et al., 2021) fact-focused schemas, (Steinkamp et al., 2019), or even spatial relations (Datta et al., 2020a,b), such approaches are limited due to their requirement for task-specific datasets that are densely annotated by domain experts.

We build upon RadGraph, which extracts medical information as disjoint radiograph representations via a combination of physician-annotated reports and entity and relation extraction model architecture (Jain et al., 2021a). This is in contrast to recent approaches like Chest ImaGenome, which structured chest X-ray reports as scene graphs using a joint rule-based approach that takes advantage of natural language processing (NLP) and an atlas-based bounding box detection pipeline (Wu et al., 2021).

Our newly proposed schema enables the classification of changes in radiology reports, an endeavor that has been considered by other approaches recently. Hassanpour et al. (2017) proposed an NLP system for characterizing several common types of changes, but their method largely relies on a set of pre-determined rules and does not allow for more fine-grained classification of changes and their context. Yuan et al. (2019) devised a deep learning architecture for the classification of changes in pulmonary nodular findings described in radiology reports, using a Siamese network to overcome data sparsity. However, their method conflates certain key change categories (e.g. new and indeterminate) and is also limited to sentence-level classification. Noto et al. (2021) applied a random forest classifier to detect changes in radiology reports of patients with high-grade gliomas. However, their

approach only considers two types of changes (stable and unstable) and operates on the report level, resulting in coarse-grained labels.

Studies on the Relationships Between Entities   Our modified information extraction model, HGIE, is built upon the DyGIE++ framework, which has shown great success in extracting information from medical reports (Wadden et al., 2019; Jain et al., 2021a). The DyGIE++ framework is a unified, multi-task framework that supports multiple information extraction tasks including named entity recognition and relation extraction. HGIE attempts to build upon this framework to improve performance on both tasks by taking advantage of the taxonomy that defines the relationships between various entities.

While HGIE takes advantage of a fixed hierarchical structure to extract entities and relations, other methodologies have considered relationships between entities for information extraction in different manners. Cotype (Ren et al., 2017) jointly embeds entity mentions, relation mentions, text features, and type labels into two low-dimensional spaces, where objects whose types are close will also have similar representations. Unlike HGIE, this approach autogenerates a noisy entity-type hierarchy for information extraction that is highly dependent on a large amount of labeled data. Fine-grained Entity Type Classification (Xu and Barbosa, 2018) adds a penalty term to the loss function that penalizes terms based on ancestor similarity. It is only capable of entity extraction, unlike HGIE which supports joint entity and relation extraction.

3. RadGraph2 Schema

We develop a novel hierarchical schema for entities and relations, which extends on the original RadGraph schema (Jain et al., 2021a), to capture detailed information about changes and their context described in radiology notes. In its original formulation, the schema was designed to maximize the coverage and retention of clinically relevant information contained in the reports while remaining sufficiently simple to enable quick and reliable labeling. Our extended schema adheres to these design principles while introducing additional entity types to represent various kinds of changes.

We utilized an iterative approach to schema development that ensures we create accurate and reliable models for capturing different change types. In each iteration, we identified entities that represent specific change types, and we labeled reports accordingly using the latest version of the schema. To ensure the schema is effective, we collected feedback from medical practitioners on areas where the schema may be ambiguous or unable to accurately capture information from annotated notes. We paid particular attention to these cases to refine and improve the schema. This iterative process was continued until we were satisfied with the schema's coverage, faithfulness, and reliability.

3.1. Entities

Entities are objects associating contiguous spans of tokens (i.e. words and punctuation marks) with their corresponding entity types. In our schema, we retain all the Anatomy and Observation entity types from the original RadGraph dataset (Jain et al., 2021a), adding several new entity types for describing changes. `CHAN-CON` entities (`CHAN-CON-AP` — condition appearance, `CHAN-CON-WOR` — condition worsening, `CHAN-CON-IMP` — condition

improvement, and `CHAN-CON-RES` — condition resolution) refer to changes in medical conditions, while `CHAN-DEV` (`CHAN-DEV-AP` — device appearance, `CHAN-DEV-PLACE` — change in device placement and `CHAN-DEV-DISA` — device disappearance) refer to changes related to supporting devices and tubes used by the patient. Finally, `CHAN-NC` indicates there have been no changes in comparison with the prior studies. We explain the different entity types in more detail in the following section.

3.1.1. Entity Types Descriptions

The Anatomy entities (`ANAT-DP`) mark mentions of anatomical locations or body parts within the report, as well as modifiers of these mentions. For example, "*left lower lung*" may be annotated as three `ANAT-DP` entities, one for each token.

The Observation entities (`OBS-DP`, `OBS-U` and `OBS-DA`) are used for tokens describing general impressions detailed in the given radiology note. These can include mentions of non-anatomical features visualized in the radiograph or descriptions of possible diagnoses. For instance, all separate tokens of the texts "*small pneumothorax*", "*support devices*", and "*within normal limits*" could be marked as Observation entities. Three different observation types are used for indicating the uncertainty level associated with the given entity: Definitely Present (`OBS-DP`), Uncertain (`OBS-U`), and Definitely Absent (`OBS-DA`). For example, in a report stating that there is "*no evidence of pneumothorax*", *pneumothorax* could be annotated as an `OBS-DA` entity due to the assertion of the absence of appreciable pneumothorax.

In addition to the Anatomy and Observation entities, our schema also defines a set of Change entities (`CHAN-NC`, `CHAN-CON-AP`, `CHAN-CON-WOR`, `CHAN-CON-IMP`, `CHAN-CON-RES`, `CHAN-DEV-AP`, `CHAN-DEV-PLACE`, `CHAN-DEV-DISA`) describing change or lack of change in comparison with the prior radiological studies.

- Change – No Change (`CHAN-NC`) entities denote that a certain condition or observation remained the same since a previously performed examination. For instance, in the texts "*no change in right pleural effusion*" and "*bibasilar atelectasis persists*", *change* and *persists* could be marked as `CHAN-NC`.

- Change in Condition (`CHAN-CON`) entities mark various kinds of changes in the observed medical conditions. We utilize four distinct sub-types for these entities: Condition Appearance (`CHAN-CON-AP`), Condition Worsening (`CHAN-CON-WOR`), Condition Improvement (`CHAN-CON-IMP`) and Condition Resolution (`CHAN-CON-RES`). `CHAN-CON-AP` indicates that a new adverse medical condition has been observed in the given patient. As an example, *new* in "*new pulmonary edema*" could be labeled with this entity type. `CHAN-CON-WOR` marks worsening in a certain aspect of the patient's clinical state. For example, a token *increased* in "*pleural effusion has increased*" would be appropriately marked as `CHAN-CON-WOR`. `CHAN-CON-IMP` indicates a general improvement, e.g. *diminished* in "*pleural effusion has diminished slightly*". Finally, `CHAN-CON-RES` signifies a complete resolution of a particular medical condition. Thus, in a note text such as "*opacity has completely cleared*", the word *cleared* might be assigned a `CHAN-CON-RES` label.

Our schema also defines three entity types for representing changes associated with medical devices: Device Appearance (CHAN-DEV-AP), Change in Device Placement (CHAN-DEV-PLACE), and Device Disappearance (CHAN-DEV-DISA). CHAN-DEV-AP entities denote that the patient has been fitted with a new medical device or tool; for example *intubated* in the report commentary "*patient has been inubated*". CHAN-DEV-PLACE entity type describes changes in the position of a medical device compared to previous studies. For instance, *migrated* in "*NG tube has migrated proximally*" could be marked as a CHAN-DEV-PLACE entity. Finally, CHAN-DEV-DISA indicates that a medical device or tool was detached or removed from the patient, as in "*nasogastric tube was removed*" where *removed* can be labeled with this entity type.

It is important to note the natural hierarchy within the entity types in our annotation schema. Anatomy (ANAT), Observation (OBS) and Change (CHAN) are more general entity types. These can be further subdivided into even more fine-grained categories. For example, a change (CHAN) can be further described as an improvement in condition (CHAN-CON-IMP). We utilize this structure of the entity labels in our hierarchical model, described in section 5. The full taxonomy of our entity types is visualized in Figure 1.

3.2. Relations

Relations in our schema are defined as directed edges between entities. Similarly to entities, each relation is associated with its corresponding label. In our work, we utilize the same set of relation labels as in Jain et al. (2021a) with slight modifications to their definitions. There are three types of relations overall: Modify (modify), Located At (located_at), and Suggestive Of (suggestive_of).

3.2.1. Relation Types Descriptions

The modify relation is used for associating Observation and Anatomy modifiers with their main entity, as well as for connecting Change entities to Observation and Anatomy entities the change relates to. For example, in the report text samples "*left lower lung*" and "*nasogastric tube was removed*" that we introduced earlier, the entity pairs (*left*, *lung*), (*lower*, *lung*), (*nasogastric*, *tube*) and (*removed*, *tube*) would be connected by a modify relationship.

The located_at relationship links Observation and Anatomy entities and indicates that the source Observation is related to the target Anatomy. While it commonly describes the location at which a certain observation was noted (hence the name of this relation type), it can also be used to describe other kinds of relationships between Observation and Anatomy entities. For instance, in the note text "*cardiac silhouette is not enlarged*", the OBS-DA entity *enlarged* and the ANAT-DP entity *silhouette* could be connected via a located_at relation.

Finally, the suggestive_of relation represents cases in which a presence of a certain Change or Observation is derived from another Change or Observation. For example, in the report sentence "*The lungs are hyperinflated suggestive of COPD.*", the entity pair (*hyperinflated*, *COPD*) could be in a suggestive_of relationship.

4. RadGraph2 Dataset

The RadGraph2 dataset is composed of 800 chest X-ray reports annotated according to a hierarchical schema. It adds fine-grained information about changes in comparison to priors while also adding 200 additional labeled reports, considerably increasing the sizes of the training and test sets.

4.1. Annotation Process

For the re-annotation of the reports from the original RadGraph dataset (Jain et al., 2021a) according to our schema, we liaised with a team of four board-certified radiologists and one academic hospitalist. The 600 reports from the RadGraph development and test datasets were imported into a specialized text labeling platform Datasaur (Datasaur) and split among the available annotators. The annotators were instructed to focus on correctly marking the mentions of change and the relations associated with these mentions while making minimal modifications to the entities and relations not associated with any changes, as we considered those to have been labeled with sufficient quality. Nevertheless, the annotators were free to correct blatant mistakes in any aspect of the annotations. Annotation instructions can be found in the supplementary material.

We conducted a pilot labeling task where they were provided with a set of 15 reports to label according to our instructions. During the pilot, we evaluated the agreement between the annotators as well as the general reliability of the labeling. We also took this opportunity to provide personalized feedback to each of the annotators so as to rectify possible misunderstandings of the labeling instructions. As the pilot study was used for training purposes, we discarded the labels obtained during the pilot task in order to maintain the integrity of the dataset.

In addition to the 600 reports from the original RadGraph development and test sets, we also extended RadGraph2 with 200 additional reports randomly sampled from the MIMIC-CXR portion of the RadGraph inference dataset. To speed up and simplify the task, we based the initial labels for these reports on the output of the RadGraph Benchmark model included in the inference set. The annotators were instructed to label the entities and relations associated with changes, as well as to correct any possible mistakes or deficiencies in the entities and relations identified by the benchmark model.

We achieved high agreement during the annotation process. The minimum pairwise agreement, computed using Cohen's Kappa, was 0.9943, and the median was 0.9963.

4.2. Data Overview

The resulting RadGraph2 dataset is comprised of 800 densely annotated reports with 23457 entities and 17373 relations. It is split into three partitions: a train partition with 575 annotated reports originating from MIMIC-CXR, a validation partition with 75 MIMIC-CXR reports, and a test partition with 100 reports from MIMIC-CXR and 50 reports from CheXpert. We give the detailed label statistics for the different dataset partitions in Table 1.

Table 1: Label statistics for the RadGraph2 datasets

	Train (%)	Validation (%)	MIMIC-CXR Test (%)	CheXpert Test (%)
ANAT	7081 (41.9)	987 (43.7)	1186 (42.6)	641 (42.7)
OBS-DP	5822 (34.4)	731 (32.3)	901 (32.4)	543 (36.2)
OBS-U	659 (3.9)	87 (3.8)	114 (4.1)	47 (3.1)
OBS-DA	1957 (11.6)	276 (12.2)	424 (15.2)	167 (11.1)
CHAN-NC	814 (4.8)	102 (4.5)	99 (3.6)	69 (4.6)
CHAN-CON-AP	70 (0.4)	5 (0.2)	4 (0.1)	5 (0.3)
CHAN-CON-WOR	186 (1.1)	35 (1.5)	14 (0.5)	17 (1.1)
CHAN-CON-IMP	139 (0.8)	18 (0.8)	17 (0.6)	5 (0.3)
CHAN-CON-RES	23 (0.1)	4 (0.2)	1 (0.0)	0 (0.0)
CHAN-DEV-AP	21 (0.1)	3 (0.1)	4 (0.1)	5 (0.3)
CHAN-DEV-PLACE	19 (0.1)	1 (0.0)	1 (0.0)	1 (0.1)
CHAN-DEV-DISA	44 (0.3)	7 (0.3)	3 (0.1)	5 (0.3)
Total Entities	16913 (100.0)	2260 (100.0)	2783 (100.0)	1501 (100.0)
modify	7945 (63.4)	1085 (64.1)	1209 (61.0)	765 (65.6)
located_at	4088 (32.6)	536 (31.7)	694 (35.0)	336 (28.8)
suggestive_of	500 (4.0)	71 (4.2)	78 (3.9)	66 (5.7)
Total Relations	12533 (100.0)	1692 (100.0)	1981 (100.0)	1167 (100.0)

5. HGIE

Unlike traditional text-to-graph models like DyGIE++ where entities are not considered to have a type hierarchy, our proposed model HGIE aims to take advantage of the inherently structured organization of our labels to improve information extraction performance.

Hierarchical recognition   Our hierarchical recognition (HR) system used in HGIE utilizes an entity taxonomy. We note that there are inherent relationships between the various entities used to label our graphs. For example, CHAN-CON-WOR and CHAN-CON-AP are both entities that refer to changes in a patient's condition. Taking advantage of these relationships between the various entities, we construct the entity taxonomy shown in Figure 1.

We use a BERT-based model as our backbone for the HR task, with the objective of extracting 12 scalar outputs. Each output is assumed to represent the conditional probability of an entity being true given that its parent in the entity hierarchy is true. During the inference phase, however, all the entities should be unconditionally predicted. Note that under such a training regimen, a trained model's unconditional probabilities can be calculated from the output using the probability chain rule via a simple application of the Bayes rule.

For example, the probability for a given leaf node can be computed as shown in Equation 1. Here, N_n denotes the leaf node. N_0 denotes the root node. The root node is defined as a dummy node that all entity types inherit from. Thus, no matter what the given true entity type is, the root node is always a correct ancestor; in other words $P(N_0) = 1$. N_k for k from 0 to the tree's max depth are defined such that N_k is a child of N_{k-1}.

$$P(N_n) = P(N_0) * \prod_{k=1}^{n} P(N_k|N_{k-1}) \tag{1}$$

Here, we condition the probability of a given entity type being correct on the probability that all of its ancestors are correct as well. This ensures that the predicted unconditional probability of a parent entity is guaranteed to be greater than or equal to its children's
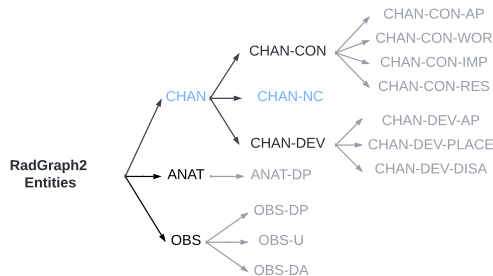
9

Figure 2: Visualization of Conditional Training. Here, the correct label is `CHAN-NC`. Black refers to negative labels, blue refers to positive labels. Under conditional training, we only penalize at the level of difference. Greyed-out portions of the tree are not part of the activation area in the loss function. In this example, the correct type is CHAN-NC, so we would penalize at the ANAT/OBS level, not at the ANAT-DP/OBS-DP/OBS-U/OBS-DA level. This improves computational complexity while also not hurting performance, as the probability of ANAT and OBS encompasses the probabilities of its children nodes. Additionally, we don't penalize or consider the nodes at levels beyond the correct classification of the node, in this example the CHAN-CON-AP/CHAN-CON-WOR/CHAN-CON-IMP/CHAN-CON-RES/CHAN-DEV-AP/CHAN-DEV-PLACE/CHAN-DEV-DISA. This is because their probabilities are already handled by comparing against CHAN-CON and CHAN-DEV.

entities. Our model is trained in two phases, first conditionally via a fine-to-coarse tree-based loss function before being finetuned unconditionally.

First Phase: Initialization under Conditional Training In the first phase of entity training, the HR system is trained on data under the condition that its parent class is positive. This follows what has been done in other work on hierarchical classification (Redmon and Farhadi, 2017; Roy et al., 2020; Yan et al., 2015; Chen et al., 2019; Pham et al., 2021). The intention behind this training regime is that it directly models the conditional probabilities of the entities by learning the dependent relationships between parent and child entities and concentrating on distinguishing lower-level labels, in particular the leaf entities.

For each entity, the losses are only calculated on entities whose parent entity is also positive. Figure 2 shows an example of how this works using the RadGraph2 entity hierarchy.

Training under a conditional probability regimen could improve the model's initialization as it attempts to focus the model on identifying entities under the same parent entity, rather than having to discriminate the entity across all other possible labels. This can help to ease confounding and convergence factors.

The idea behind the regimen is that it can help alleviate the issue of low label prevalence due to lower reliance on negative samples that occurs within the medical report problem space.

Fine-to-Coarse Depth-Dependant Hierarchical Loss Function   For conditional training, we compute a sum of entity recognition losses over each level within the entity tree. The entity tree is defined via (N, E), where $N = \{n_1, n_2, n_3, .., n_n\}$ are the set of nodes, and where $E \subseteq N \times N$ is the set of edges in the tree such that $(n_i, n_j) \in E$ if and only if n_i is the parent node of n_j. Without loss of generalizability, we assume that the first k nodes correspond to the leaves of the tree, such that $\nexists n_j \in N$ where $(n_i, n_j) \in E$ for $i = 1, ..., k$. These nodes will refer to the finest-scaled entities in our system. This also means that if $\exists n_j \in N$ where $(n_i, n_j) \in E$, then n_i is not a leaf node and is instead a more general entity definition of n_j. The label l used for HR will be at the leaf level, that is $l \in \{1, ..., k\}$.

The depth of node n_i under this definition of a tree is denoted via $\text{depth}(n_i)$, which refers to the number of edges between node n_i and the root node. Thus, the max depth (d_{max}) is determined by the largest number of edges between any node and the root; in other words $d_{max} = \max_i \text{depth}(n_i)$.

The HR model outputs one logit, which we denote via x_i per leaf node n_i. Thus, the probability p associated with leaf node n_i can be computed by applying the softmax function σ on the logit x_i. Thus, p_i denotes the i-th entry of the vector $\sigma(x)$. As discussed above, the probability of a parent node is defined to be the sum of the probabilities of its leaf nodes, leading the probability computation to be fine-to-coarse. Combining these principles together provides us with the formula for computing the probability for any node in the tree as shown in Equation 2.

$$p_i = \begin{cases} \sigma(x_i) & \text{IF } 1 \leq i \leq k \\ \displaystyle\sum_{(n_i, n_j) \in E} p_j & \text{Otherwise} \end{cases} \tag{2}$$

As we take the softmax of the logits, and all parent nodes probabilities are defined to be the summation of a subset of leaf nodes, $0 \leq p_i \leq 1$ for any node n_i, making this definition of the conditional probability for any node valid.

The correct label index c provided to the HR model will be such that corresponds to the index of a leaf node ($c \in \{1, ..., k\}$), enabling easy loss computation for leaf nodes. In order to compute losses at all depths, we need to determine the correct node at depth d which we denote n_{cd}. This computation is shown in Equation 3.

$$n_{cd} = \begin{cases} n_c & \text{IF } \text{depth}(n_c) = d \\ n_i & \text{IF } d < \text{depth}(n_c) \text{ AND } d = \text{depth}(n_i) \text{ AND } \text{anc}(n_i, n_c) \\ \text{N/A} & \text{Otherwise} \end{cases} \tag{3}$$

Here, $\text{anc}(n_i, n_c)$ is true if and only if node n_i is an ancestor of node n_c; that is node n_i represents a superclass of node n_c. Additionally, as mentioned in Section 5, losses for conditional training are not computed for all nodes. Those nodes where the loss is not computed do not have a correct node, thus n_{cd} is defined as N/A in those scenarios.

Putting these definitions together, the loss at depth d is computed using the negative log loss of the probability at the given depth. The hierarchical loss L is the summation of these losses at all depths in the tree as shown in Equation 4.
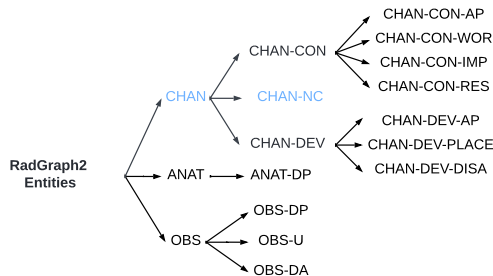
Figure 3: Visualization of Unconditional Fine-tuning. Here, the correct label is `CHAN-NC`. Black refers to negative labels, blue refers to positive labels. As the whole tree is part of the activation area in this phase, there are no greyed-out parts.

$$L = -\sum_{d=0}^{\mathrm{d}_{max}} \log(p_{n_{cd}}) \tag{4}$$

Second Phase: Fine-tuning under Unconditional Training This stage aims at improving the accuracy of unconditional probability predictions, which is used during inference and is thus critical to classification performance. To achieve this, we finetune our hierarchically trained network on the full dataset using a standard categorical cross-entropy loss function and smaller learning rates. This training stage aims at improving the capacity of the network in predicting parent-level labels, which could be either positive or negative. Implementation details used for training can be found in section 6.1.3.

6. Experiments

We propose an entity and relation extraction task for radiology reports that can be developed using our development dataset and tested using our test dataset. Like the original RadGraph dataset, our dataset processes each radiology report into a sequence of space-delimited tokens, where punctuation like commas and semicolons have been separated from words to support entity recognition. For each report, we provide annotations identifying the type and span of each entity as well as relations between entities.

6.1. Evaluation

6.1.1. EVALUATION MODELS

We compare two entity and relation extraction model baselines against HGIE for entity and relation extraction tasks on RadGraph2. The first baseline uses the DyGIE++ framework (Wadden et al., 2019), which was able to achieve state-of-the-art performance on the original RadGraph by jointly extracting entities and relations. Our second baseline uses the Princeton University Relation Extraction model (PURE) by Zhong et al. (Zhong and Chen,

2021), which achieved state-of-the-art at the time of relation extraction using a pipeline approach that decomposes entity and relation extraction into different subtasks. We pit these baselines against HGIE which, as detailed above, builds upon the DyGIE++ framework to take advantage of the relationship between entities.

6.1.2. EVALUATION METHOD

We evaluate our models on the entity and relation extraction tasks for radiology reports, using the training and validation partitions for training and the test partition for the final evaluation. Each of the input radiology reports is transformed into a sequence of space-delimited tokens with punctuation marks being considered as separate tokens. Every report is also associated with annotations that provide information about the spans and types for all entities and relations occurring in the report.

In order to assess HGIE's performance in the medical sphere we evaluate all models on both RadGraph2 and the original RadGraph. With different definitions for relations, the introduction of new entities, and a vastly different entity hierarchy, both datasets are considerably different and thus allow for a comprehensive evaluation of our approach.

It is also important to determine how effective RadGraph2 is at performing tasks the original RadGraph dataset was capable of. We evaluate our benchmark model on the intersection of the test partitions of both datasets; that is, on the entities and relations that are the same between the two datasets. The benchmark model trained on RadGraph2 should ideally perform similarly to the one trained on RadGraph set when evaluated on the special set.

We report results on the MIMIC-CXR and CheXpert test partitions separately across all experiments. For more details on these partitions, refer to Section 4.2.

6.1.3. IMPLEMENTATION DETAILS

We use BERT (Devlin et al., 2018) for all approaches. For DyGIE++, we use a learning rate of 5e-5 with a batch size of 1 for BERT and a learning rate of 1e-3 for task-specific layers, consistent with the approach used by Wadden et al. These same parameters are used for the first training phase of HGIE, while the fine-tuning phase uses a learning rate of 8e-6 with a batch size of 1 for BERT and a learning rate of 2e-4 for task-specific layers, which is consistent with the finetuning-approach used by (Xu et al., 2021).

For PURE entity extraction, we use a learning rate of 1e-5 with a batch size of 16 for BERT, a span length of 3, and a learning rate of 5e-5 for task-specific layers. For PURE relation extraction, we use a learning rate of 2e-5, a context window of 50, and a batch size of 16 for BERT. These values are derived from Jain et al. (2021a).

For each of our approaches, in addition to using BERT weight initializations, we use weight initializations from five different biomedical pre-trained models: BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), PubMedBERT (Gu et al., 2021), BlueBERT (Peng et al., 2019), and RadBERT (Yan et al., 2022).

6.1.4. EVALUATION METRICS

We report both micro and macro F1 for joint entity recognition and relation extraction for all models. A predicted entity is considered correct if both the predicted span boundaries

Table 2: Relation extraction performance on RadGraph2. Best-performing values are bolded.

Model	MIMIC-CXR Relations		CheXpert Relations	
	F1 micro	F1 macro	F1 micro	F1 macro
DyGIE++				
BERT Base	0.848	0.823	0.703	0.673
BioBERT	0.850	0.815	0.714	0.689
ClinicalBERT	0.847	0.783	0.703	0.642
PubMedBERT	0.853	0.824	0.719	0.692
BlueBERT	0.820	0.783	0.651	0.642
RadBERT	0.851	0.832	0.723	0.693
PURE				
BERT Base	0.851	0.833	0.709	0.676
BioBERT	0.850	0.812	0.715	0.691
ClinicalBERT	0.841	0.785	0.708	0.649
PubMedBERT	0.854	0.841	0.718	0.693
BlueBERT	0.817	0.776	0.646	0.639
RadBERT	0.855	0.843	0.718	0.698
HGIE (Ours)				
BERT Base	0.841	0.801	0.701	0.672
BioBERT	0.853	0.838	0.727	**0.718**
ClinicalBERT	0.841	0.815	0.715	0.691
PubMedBERT	0.874	0.857	0.732	0.707
BlueBERT	0.828	0.785	0.675	0.654
RadBERT	**0.879**	**0.862**	**0.739**	0.714

and predicted entity type are correct. A predicted relation is considered correct if the pair of entities are correct, including both span boundaries and entity type, and the relation type is correct. As the performance on relation extraction also depends on the performance on the entity recognition (as the predicted relations cannot possibly be correct if the underlying entities are wrong), we focus our evaluation on this metric, as it provides a more holistic view of the performance of the model. Results on entity recognition can be found in Appendix A. We separately report the results on the test data originating from the MIMIC-CXR and CheXpert.

7. Results

Taking advantage of hierarchy leads to improvements in performance. We evaluate HGIE against our baseline models on both the MIMIC-CXR and CheXpert test sets. When comparing approaches, we use the strict relation extraction metric defined above as the primary end-to-end approach metric, as it uses both predicted entities and relations in its computation. Results are shown in Table 2. HGIE is the best-performing model on this dataset, with the model outperforming its counterparts on both test sets. For most initializations, HGIE beats DyGIE++ and PURE in both F1 micro and macro on both test

Table 3: Comparison of the entity extraction performance on the best performing Rad-Graph2 and RadGraph models on a common set of entities. Best-performing values are bolded.

Best Model	Anatomy	Observation: Definitely Present	Observation: Uncertain	Observation: Definitely Absent	F1 micro	F1 macro
			MIMIC-CXR			
RadGraph	0.968	0.922	0.700	**0.952**	0.940	0.886
RadGraph2	**0.983**	**0.963**	**0.798**	0.813	**0.943**	**0.891**
			CheXpert			
RadGraph	**0.941**	0.884	0.714	**0.910**	**0.905**	0.862
RadGraph2	0.880	**0.939**	**0.812**	0.823	0.894	**0.864**

Table 4: Comparison of relation extraction performance on the best performing RadGraph2 and RadGraph models on a common set of entities and relations. Best-performing values are bolded.

Best Model	Modify	Located At	Suggestive Of	F1 micro	F1 macro
			MIMIC-CXR		
RadGraph	0.804	**0.861**	0.685	0.823	0.783
RadGraph2	**0.906**	0.826	**0.838**	**0.874**	**0.857**
			CheXpert		
RadGraph	0.709	**0.779**	0.588	0.725	0.692
RadGraph2	**0.777**	0.712	**0.730**	**0.746**	**0.740**

sets. Across all initializations, HGIE outperforms DyGIE++ on both test sets. HGIE and DyGIE++ possess the same model architectures, only differing in how they are trained, indicating that utilizing the entity taxonomy during training leads to better performance on RadGraph2. Appendix A shows similar results for the original RadGraph dataset.

RadGraph2 models perform at least as well as original RadGraph models on the same tasks. We specifically evaluate the performance of the best-performing RadGraph2 model alongside the best-performing original RadGraph model on both entity recognition and relation extraction. These models are cross-compared between datasets by creating an intersection dataset. This is created by pruning entities and relations that are unique to RadGraph2. Results can be found in Tables 3 and 4.

Both benchmark models perform similarly on the entity extraction task, with the RadGraph2 model doing better for 2 out of 4 entity types on both MIMIC-CXR and CheXpert. The RadGraph2 model has higher performance on the more difficult relation extraction task, doing better for 3 out of 4 relation types on both MIMIC-CXR and CheXpert.

Table 5: Effect of Entity Taxonomy Tree Depth on HGIE Performance. Best values are bolded.

Model	MIMIC-CXR Relations		CheXpert Relations	
	F1 micro	F1 macro	F1 micro	F1 macro
Depth 3				
BERT Base	0.841	0.801	0.701	0.672
BioBERT	0.853	0.838	0.727	**0.718**
ClinicalBERT	0.841	0.815	0.715	0.691
PubMedBERT	0.874	0.857	0.732	0.707
BlueBERT	0.828	0.785	0.675	0.654
RadBERT	0.879	0.862	0.739	0.714
Depth 2				
BERT Base	0.841	0.801	0.702	0.660
BioBERT	0.856	0.840	0.726	0.711
ClinicalBERT	0.852	0.833	0.722	0.717
PubMedBERT	0.881	**0.862**	0.741	0.716
BlueBERT	0.824	0.781	0.685	0.658
RadBERT	**0.883**	0.861	**0.742**	0.717

Modifying taxonomy structure can lead to improvements in performance. So far, we have treated the structure of the entity taxonomy to be fixed. However, there are two different ways to categorize the entities, which can be defined based on the maximum depth of the taxonomy tree. Here, we seek to determine the effects of depth on the final model performance. We pit the depth 3 tree shown in Figures 1, 2, and 3 against a depth 2 tree where all change-related entities are children of the generic change node. Both models have similar performance to each other across model weight initializations. Nonetheless, the DGIE model trained under a taxonomy with a maximum depth of 2 tends to outperform the model with a taxonomy depth of 3. This is likely due to the small number of labeled entities for certain leaf nodes, like `CHAN-CON-RES` which only has a single example in the joint test set, adding noise during the conditional training phase. These results indicate that the structure of the tree in low data label environments should be considered akin to a tunable hyperparameter, with various depths evaluated in order to obtain optimal performance.

8. Limitations

Our study has several notable limitations. First, we only evaluate the HGIE model on radiology report datasets despite it being designed to generalize to any information extraction task where entities form a taxonomy. We do expect that the model training methodology proposed will work for other domains to solve similar problems, and strongly encourage research into further applications. Second, while care has been taken to make our RadGraph2 information schema as clear as possible, there are still certain cases in which the labeling can be ambiguous. For example, a radiologist may describe a change with a certain degree of uncertainty, which cannot be directly modeled by our schema. Third, the reports in our

data are limited to chest X-ray radiology notes from MIMIC-CXR and CheXpert, though it is important to note that the RadGraph2 schema is designed such that it can annotate radiology reports in general. Fourth, the reports for both the MIMIC-CXR and CheXpert datasets are collected only from two United States hospitals (Beth Israel Deaconess Medical Center in Boston, MA, and Stanford Hospital in Stanford, CA, respectively). The distribution of the data may thus be substantially different from other hospitals and countries and the users planning to make use of the RadGraph2 dataset are urged to seriously consider the risks associated with domain shift.

9. Ethical Considerations

Some ethical considerations associated with this work stem from our usage of medical reports from MIMIC-CXR and CheXpert. As in any situation in which real-world patient medical information is used in research, great care must be exercised in order not to compromise the privacy and rights of the human subjects from which the data has been collected. In our case, the risks associated with the usage of health information are greatly mitigated by the fact that the data has been carefully de-identified. Due to this fact, our work does not constitute human subject research and is exempt from a review by an institutional review board (IRB). Nevertheless, we still recognize that the data is potentially sensitive and thus take great care to use it responsibly and in accordance with its license.

10. Conclusion

We introduce RadGraph2, a new dataset of 800 chest X-ray expert-annotated reports and additional 220,913 reports annotated via inference. RadGraph2 combines the schema from the previous version of RadGraph with additional entities focused on characterizing priors along with clinically relevant contextual information in radiology reports. We characterize the entities via a hierarchy that follows the natural relationships between various types of entities. Via a modification of an information extraction model that has shown great promise in the medical sphere, we show that taking advantage of this hierarchy during training can lead to an improved joint entity and relation extraction performance.

We believe that our work paves the way for the development of systems capable of automatically characterizing disease progression in patients over time, as well as novel information extraction models leveraging the natural hierarchy of the labels in their target domain.

Acknowledgements

References

Julián N Acosta, Guido J Falcone, and Pranav Rajpurkar. The need for medical artificial intelligence that incorporates prior images. *Radiology*, page 212830, 2022.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*, 2019.

Haomin Chen, Shun Miao, Daguang Xu, Gregory D Hager, and Adam P Harrison. Deep hierarchical multi-label classification of chest x-ray images. In *International conference on medical imaging with deep learning*, pages 109–120. PMLR, 2019.

Hwa-Yen Chiu, Heng-Sheng Chao, and Yuh-Min Chen. Application of artificial intelligence in lung cancer. *Cancers*, 14(6):1370, 2022.

Datasaur. Datasaur.ai. https://datasaur.ai/.

Surabhi Datta, Yuqi Si, Laritza Rodriguez, Sonya E Shooshan, Dina Demner-Fushman, and Kirk Roberts. Understanding spatial language in radiology: Representation framework, annotation, and spatial relation extraction from chest x-ray reports using deep learning. *Journal of biomedical informatics*, 108:103473, 2020a.

Surabhi Datta, Morgan Ulinski, Jordan Godfrey-Stovall, Shekhar Khanpara, Roy F Riascos-Castaneda, and Kirk Roberts. Rad-spatialnet: a frame-based resource for fine-grained spatial relations in radiology reports. In *LREC... International Conference on Language Resources & Evaluation:[proceedings]. International Conference on Language Resources and Evaluation*, volume 2020, page 2251. NIH Public Access, 2020b.

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

A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. Ch. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley. PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000. Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Saeed Hassanpour and Curtis P Langlotz. Information extraction from multi-institutional radiology reports. *Artificial intelligence in medicine*, 66:29–39, 2016.

Saeed Hassanpour, Graham Bay, and Curtis P Langlotz. Characterization of change and significance for clinical findings in radiology reports through natural language processing. *Journal of digital imaging*, 30(3):314–322, 2017.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. 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, 2019.

Saahil Jain, Ashwin Agrawal, Adriel Saporta, Steven Truong, Du Nguyen Duong, Tan Bui, Pierre Chambon, Yuhao Zhang, Matthew Lungren, Andrew Ng, Curtis Langlotz, Pranav Rajpurkar, and Pranav Rajpurkar. Radgraph: Extracting clinical entities and relations from radiology reports. In J. Vanschoren and S. Yeung, editors, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021a.

Saahil Jain, Akshay Smit, Steven QH Truong, Chanh DT Nguyen, Minh-Thanh Huynh, Mudit Jain, Victoria A Young, Andrew Y Ng, Matthew P Lungren, and Pranav Rajpurkar. Visualchexbert: addressing the discrepancy between radiology report labels and image labels. In *Proceedings of the Conference on Health, Inference, and Learning*, pages 105–115, 2021b.

Jaehwan Jeong, Katherine Tian, Andrew Li, Sina Hartung, Fardad Behzadi, Juan Calle, David Osayande, Michael Pohlen, Subathra Adithan, and Pranav Rajpurkar. Multi-modal image-text matching improves retrieval-based chest x-ray report generation. *arXiv preprint arXiv:2303.17579*, 2023.

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

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

Mingjie Li, Wenjia Cai, Karin Verspoor, Shirui Pan, Xiaodan Liang, and Xiaojun Chang. Cross-modal clinical graph transformer for ophthalmic report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20656–20665, 2022.

Mingjie Li, Bingqian Lin, Zicong Chen, Haokun Lin, Xiaodan Liang, and Xiaojun Chang. Dynamic graph enhanced contrastive learning for chest x-ray report generation. *arXiv preprint arXiv:2303.10323*, 2023.

Matthew BA McDermott, Tzu Ming Harry Hsu, Wei-Hung Weng, Marzyeh Ghassemi, and Peter Szolovits. Chexpert++: Approximating the chexpert labeler for speed, differentiability, and probabilistic output. In *Machine Learning for Healthcare Conference*, pages 913–927. PMLR, 2020.

Anika L McGrath, Geraldine McGinty, Wendie A Berg, Ellen B Mendelson, Michele B Drotman, Richard L Ellis, and Curtis P Langlotz. Optimizing the breast imaging report for today and tomorrow, 2022.

Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

Tommaso Di Noto, Chirine Atat, Eduardo Gamito Teiga, Monika Hegi, Andreas Hottinger, Meritxell Bach Cuadra, Patric Hagmann, and Jonas Richiardi. Diagnostic surveillance of high-grade gliomas: towards automated change detection using radiology report classification. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 423–436. Springer, 2021.

Yifan Peng, Xiaosong Wang, Le Lu, Mohammadhadi Bagheri, Ronald Summers, and Zhiyong Lu. Negbio: a high-performance tool for negation and uncertainty detection in radiology reports. *AMIA Summits on Translational Science Proceedings*, 2018:188, 2018.

Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: an evaluation of bert and elmo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474*, 2019.

Hieu H Pham, Tung T Le, Dat Q Tran, Dat T Ngo, and Ha Q Nguyen. Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. *Neurocomputing*, 437:186–194, 2021.

Pranav Rajpurkar and Matthew P Lungren. The current and future state of ai interpretation of medical images. *New England Journal of Medicine*, 388(21):1981–1990, 2023.

Vignav Ramesh, Nathan Chi, and Pranav Rajpurkar. Cxr-pro: Mimic-cxr with prior references omitted.

Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

Xiang Ren, Zeqiu Wu, Wenqi He, Meng Qu, Clare R Voss, Heng Ji, Tarek F Abdelzaher, and Jiawei Han. Cotype: Joint extraction of typed entities and relations with knowledge bases. In *Proceedings of the 26th international conference on world wide web*, pages 1015–1024, 2017.

Deboleena Roy, Priyadarshini Panda, and Kaushik Roy. Tree-cnn: a hierarchical deep convolutional neural network for incremental learning. *Neural Networks*, 121:148–160, 2020.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. Chexbert: combining automatic labelers and expert annotations for accurate radiology report labeling using bert. *arXiv preprint arXiv:2004.09167*, 2020.

Jackson M Steinkamp, Charles Chambers, Darco Lalevic, Hanna M Zafar, and Tessa S Cook. Toward complete structured information extraction from radiology reports using machine learning. *Journal of digital imaging*, 32(4):554–564, 2019.

Doron Stupp, Ronnie Barequet, I-Ching Lee, Eyal Oren, Amir Feder, Ayelet Benjamini, Avinatan Hassidim, Yossi Matias, Eran Ofek, and Alvin Rajkomar. Structured understanding of assessment and plans in clinical documentation. *medRxiv*, pages 2022–04, 2022.

Kento Sugimoto, Toshihiro Takeda, Jong-Hoon Oh, Shoya Wada, Shozo Konishi, Asuka Yamahata, Shiro Manabe, Noriyuki Tomiyama, Takashi Matsunaga, Katsuyuki Nakanishi, et al. Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116:103729, 2021.

David Wadden, Ulme Wennberg, Yi Luan, and Hannaneh Hajishirzi. Entity, relation, and event extraction with contextualized span representations. *arXiv preprint arXiv:1909.03546*, 2019.

Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Medklip: Medical knowledge enhanced language-image pre-training. *medRxiv*, pages 2023–01, 2023.

Joy T Wu, Nkechinyere N Agu, Ismini Lourentzou, Arjun Sharma, Joseph A Paguio, Jasper S Yao, Edward C Dee, William Mitchell, Satyananda Kashyap, Andrea Giovannini, et al. Chest imagenome dataset for clinical reasoning. *arXiv preprint arXiv:2108.00316*, 2021.

Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. Gradual fine-tuning for low-resource domain adaptation. In *Proceedings of the Second Workshop on Domain Adaptation for NLP*, pages 214–221, Kyiv, Ukraine, April 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.adaptnlp-1.22.

Peng Xu and Denilson Barbosa. Neural fine-grained entity type classification with hierarchy-aware loss. *arXiv preprint arXiv:1803.03378*, 2018.

An Yan, Julian McAuley, Xing Lu, Jiang Du, Eric Y Chang, Amilcare Gentili, and Chun-Nan Hsu. Radbert: Adapting transformer-based language models to radiology. *Radiology: Artificial Intelligence*, 4(4):e210258, 2022.

Zhicheng Yan, Hao Zhang, Robinson Piramuthu, Vignesh Jagadeesh, Dennis DeCoste, Wei Di, and Yizhou Yu. Hd-cnn: hierarchical deep convolutional neural networks for large scale visual recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 2740–2748, 2015.

Shuxin Yang, Xian Wu, Shen Ge, S Kevin Zhou, and Li Xiao. Knowledge matters: Chest radiology report generation with general and specific knowledge. *Medical Image Analysis*, 80:102510, 2022.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *medRxiv*, 2022. doi: 10.1101/2022.08.30.22279318. URL https://www.medrxiv.org/content/early/2022/08/31/2022.08.30.22279318.

Jianbo Yuan, Henghui Zhu, and Amir Tahmasebi. Classification of pulmonary nodular findings based on characterization of change using radiology reports. *AMIA Summits on Translational Science Proceedings*, 2019:285, 2019.

Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *NAACL*, 2021.

Appendix A. Performance on Original RadGraph

As discussed in the main manuscript, we wish to evaluate how well RadGraph2 encompasses the original dataset. To this end, we perform a head-to-head evaluation of the best models trained on each dataset and evaluate their performance on the intersection between their test sets.

To do this, we evaluate our models using the relation extraction task on the original RadGraph dataset. We constructed an entity taxonomy tree where observation-based entities were children of a generalized observation parent node and had a depth of 2. As `ANAT-DP` is the only type of anatomy node, we set it to be a leaf node at depth 1.

Except for F1-macro on the MIMI-CXR test set with a PubMedBERT initialization, HGIE outperforms both baselines across all initializations. The best-performing model on this dataset was HGIE using a PubMedBERT initialization. Like on the RadGraph2 dataset, model weight initialization has a significant effect on final model performance but the deltas are slightly less pronounced on this dataset. F1 deltas for different initializations are as high as 0.040 for HGIE compared to 0.026 for PURE and 0.031 for DyGIE++.

DyGIE++ achieves an F1-micro of 0.826 and F1-macro of 0.787 for MIMIC-CXR Relations and achieves an F1-micro of 0.726 and F1-macro of 0.694 for CheXpert Relations. HGIE achieves an F1-micro of 0.852 and F1-macro of 0.791 for MIMIC-CXR Relations and achieves an F1-micro of 0.744 and F1-macro of 0.723 for CheXpert Relations.

Table 6: Relation extraction performance on RadGraph 1.0. Best-performing values are bolded. As was the case for RadGraph2, the best combination is HGIE with a RadBERT initialization.

Model	MIMIC-CXR Relations		CheXpert Relations	
	F1 micro	F1 macro	F1 micro	F1 macro
DyGIE++				
BERT Base	0.805	0.752	0.712	0.688
BioBERT	0.801	0.731	0.701	0.668
ClinicalBERT	0.806	0.739	0.701	0.672
PubMedBERT	0.823	0.783	0.725	0.692
BlueBERT	0.803	0.712	0.705	0.664
RadBERT	0.826	**0.787**	0.726	0.694
PURE				
BERT Base	0.805	0.731	0.722	0.648
BioBERT	0.806	0.757	0.721	0.654
ClinicalBERT	0.809	0.746	0.728	0.664
PubMedBERT	0.812	0.745	0.729	0.679
BlueBERT	0.818	0.738	0.699	0.655
RadBERT	0.814	0.746	0.723	0.675
HGIE (Ours)				
BERT Base	0.809	0.752	0.719	0.699
BioBERT	0.841	0.773	0.739	**0.730**
ClinicalBERT	0.842	0.774	0.745	0.729
PubMedBERT	0.849	0.779	**0.752**	0.726
BlueBERT	0.838	0.740	0.730	0.699
RadBERT	**0.852**	0.791	0.744	0.723

Appendix B. Annotation Instructions

In this section, we provide the annotation instructions that were given to all annotators on the project.

B.1. Datasaur Platform Instructions

For this task, please label the reports assigned to you on datasaur.ai (a labeling platform). In addition to features shown in the video (some of which are not relevant, e.g. the ML-assisted labeling), you can also use the list of files under extensions on the right to navigate between the different batches of reports assigned to you and to mark these as complete once you have finished checking and modifying them.

Before you start labeling, please make sure that you select the correct set of labels for the arrow relations. You can do this by clicking on the dropdown menu marked with an arrow (in the top left part of the labeling interface) and selecting the second option from the top.

Please make sure that you use the dropdown on the top left rather than the one on the right side of the interface. You can check that you applied the correct label set by clicking on any arrow within a report — if the label set is correct, the available options should be located_at, modify, and suggestive_of.

B.2. General Task Information

The labeling task is to correct any mistakes or deficiencies in the existing RadGraph labels. The labels consist of entities (Change, Anatomy, and Observation) and relations between them (suggestive_of, located_at and modify) marked in each radiology report. Most of these entities and relations are already in place and are shown in the Datasaur interface, but some of them may be missing, mislabelled or improperly selected. Your main task is to look for any comparisons to the prior radiology examinations mentioned in the reports and ensure that they are correctly marked using the appropriate Change entities (described below) and relations connected to them. Thus, you may need to change the types of certain entities or relations, add new ones, or (in some cases) remove existing ones. The Anatomy and Observation entities and the relations between them should be substantially more complete and reliable, so you are not expected to actively check them (provided that they do not mark comparison to a prior, in which case they should be converted into a Change entity), but please feel free to amend or modify them if you spot any obvious mistakes.

Update: Please focus on correctly marking the change entities and the relations associated with them and try not to make extensive changes to the entities and relations not related to changes. Please see the "Datasaur annotator feedback" document for details. (Note: The referred-to document includes names and provides personalized feedback to the annotators. In order to not break anonymity, these details will not be included.)

B.3. Entities

- Change (CHAN): an expression explicitly describing change or lack of change compared to prior or indicating that comparison to prior has occurred. Examples of

change entities include "stable", "change", "unchanged", "new", "larger", "comparison". There are several different types of change entities, which we describe below along with examples. Each example shows the text span(s) marked with the given entity type, highlighted in bold. We also give the outgoing relations from the change entity (see below for information regarding relations).

– CHAN-NC indicates a lack of change since the prior study. Note that this does not cover cases in which a finding is reported without explicit information of whether it is different from the prior or not — in these cases, the change entity types should not be used at all (instead, the appropriate OBS- or ANAT- entities should be marked).

   * In comparison with the earlier study of this date, there is essentially no change in the appearance of the enteric tube.
      · change modifies tube
   * Moderately severe bibasilar atelectasis persists.
      · persists modifies atelectasis
   * Cardiomediastinal silhouette is unchanged.
      · unchanged modifies silhouette

– CHAN-IMP indicates an improvement in a certain aspect of the patient's clinical state compared to the prior. If a certain adverse medical condition of the patient was described as being completely resolved, use CHAN-CON-RES instead.

   * Left basilar opacity has nearly resolved in the right lower lobe opacity has improved.
      · resolved modifies opacity (first occurrence) — note that the condition has not completely resolved, so the change is only considered to be an improvement
      · improved modifies opacity (second occurrence)
   * Compared to the most recent study, there is improvement in the mild pulmonary edema and a decrease in the small left pleural effusion.
      · Improvement modifies edema
      · decrease modifies effusion

– CHAN-WOR indicates a worsening in a certain aspect of the patient's clinical state compared to the prior. If a certain adverse medical condition of the patient was described as being completely new, use CHAN-CON-AP instead.

   * Mild - to - moderate diffuse pulmonary edema is slightly worse.
      · worse modifies edema
   * Moderate right pleural effusion has increased since...
      · increased modifies effusion

– CHAN-CON-AP indicates that, compared to the previous report(s), a new adverse medical condition has been observed in the given patient.

   * There is also a new left basilar opacity blunting the lateral costophrenic angle (...)

      · new modifies opacity

– CHAN-CON-RES indicates that compared to the previous report(s), a certain medical condition previously observed in the patient has completely resolved

      ∗ Indistinct superior segment left lower lobe opacities have resolved.

        · resolved modifies opacities

– CHAN-DEV-AP indicates that, compared to the previous report(s), the patient has been fitted with a new medical device or tool (e.g. intubated, catheterized, . . .)

      ∗ The patient has received the new feeding tube.

        · new modifies tube

– CHAN-DEV-PLACE indicates that the position of a medical device in the body of a patient changed compared to prior studies.

      ∗ Left pleural drain has been pulled in a slightly higher position.

        · higher modifies position

      ∗ Left pleural drain has been advanced to the left apex.

        · Advanced modifies drain

– CHAN-DEV-DISA indicates that, compared to the previous report(s), a medical device or tool was detached or removed from the patient.

      ∗ In the interval, the patient has been extubated (...)

        · Extubated modifies patient

      ∗ The nasogastric tube has been removed.

        · removed modifies tube

– The change entity types are used solely for spans of tokens indicative of change (or no change) in the report texts, not for entities representing medical conditions, devices, etc. related to the change. Instead, the entity denoting change will typically be attached to the other relevant entities in the report via a relation. This is demonstrated in the above examples.

– Anatomy (ANAT-DP): an anatomical body part that occurs in the radiology report. Examples of anatomy entities include "lung", "left lower lobe of the lung" (multiple entities), or "aortic arch" (multiple entities).

      ∗ Anatomy vs. Anatomy Modifier: in this schema, all anatomy modifiers are annotated as anatomy entities. In the case that anatomy modifies the scope or degree of a second anatomy, a "modify" relation is added to denote the modification relationship between the two entities. See below for a definition and example of the "modify" relation. So, "left lung" would be two anatomy entities, where "left" modifies "lung".

– Observation (OBS): an observation made from the images and associated with visual features, identifiable pathophysiologic processes or diagnostic disease classifications. Examples of multiple observation entities include "airspace opacity", "mass", "bilateral pleural effusion" or "pneumonia".

∗ Observation vs. Observation Modifier: in this schema, all observation modifiers are annotated as observation entities. In the case that an observation modifies the scope or degree of a second observation, a "modify" relation is added to denote the modification relationship between the two entities. See below for a definition and example of the "modify" relation.

– Note that each observation entity is associated with an uncertainty attribute measuring the uncertainty level of an observation entity or anatomy entity. Each uncertainty attribute can have one of three values: definitely present (OBS-DP), uncertain (OBS-U), or definitely absent (OBS-DA).

∗ When you label an observation, you will pick from the following options: Observation::definitely present (OBS-DP), Observation::uncertain (OBS-U), or Observation::definitely absent (OBS-DA).

∗ In some cases text spans for Observations are not continuous. In that case, each span should be labeled as observation with a "modify" relation between them.

B.4. Relations

• Relations are directed arrows from one entity to another that are used to describe a relationship between two entities. We define the following three types of relations, in the form of "relation_type (entity _type, entity_type)":

– suggestive_of (observation, observation) or (change, observation) or (observation, change): a relation between two observation entities indicating that the status of the second observation is inferred from that of the first observation or indicating that change is derived from an observation or indicating that an observation is derived from a change.

– located_at (observation, anatomy): a relation between an observation and an anatomy entity indicating that the observation is related to the anatomy. While located_at often refers to location, it can also be used to describe other relations between an observation and an anatomy. For example, in the sentence, "heart is normal", "normal" is an observation, "heart" is an anatomy, and "located_at" is a relation from "normal" to "heart."

– modify (observation, observation) or (anatomy, anatomy) or (change, any) or (observation, change): a relation between two observation entities indicating that the first observation modifies the scope of or quantifies the degree of the second observation. This relation is often added when the first observation entity is an "observation modifier" that modifies the second observation. The same logic applies to an anatomy entity modifying another anatomy entity. Additionally, the modify relationship is commonly used to connect change entities to other entities, as demonstrated in the examples above. In some cases, the modify relation can also be used to connect an observation entity to a change entity, as shown in some of the examples in this document.

B.5. Change Entities Annotation Conventions

- In some cases, there might be multiple ways in which to mark the change entities in the given report. In order to improve the consistency of the data, we adopted several conventions:

  - Sometimes, there might conceivably be multiple spans of words indicating a certain change category. We generally strive to select the minimum possible span related to the description of the change type in the note text. For example, if the note says "Biapical scarring is again seen", we mark the entity "again" as CHAN-NC rather than using an entity with a longer span "again seen". Similarly, if the note says "In comparison with the earlier study of this date, there is essentially no change in the appearance of the enteric tube", we only mark the entity "change" as CHAN-NC instead of marking an entity "no change" (the model using the annotations will still consider the wider context, so "no" does not need to be annotated at all).

  - We almost exclusively use the "modifies" relation when connecting the change entities with the rest of the graph. If the change entity is connected to the rest of the graph using a "located_at" relation, please consider changing it to "modifies". Additionally, the change entity should only be connected to the main entity/entities that is/are related to the change.

  - If the report indicates uncertainty with regards to the presence of a change, we will usually attempt to ascertain whether the presence of the change was perceived to be likely by the author of the note. If yes, we mark the change, otherwise, we annotate these entities using the "OBS-" markers from the origin=al RadGraph (version 1)

  - We mark all mentions of a change in a report, even if some of the mentions are duplicate (i.e. they repeat information about an identical change). Additionally, the reports will often contain multiple changes of different types (e.g. if certain aspects of the patient's medical state improved and others worsened).

B.6. Suggested Annotation Process

- (Note: These original instructions for annotating the reports are included for information only, you are only expected to verify and correct the existing labels)

  - Here is a suggested process for each sentence or sentence group:
    * Find the main anatomy (if present), and label as anatomy
    * Find any anatomic modifiers. Label these modifiers as anatomy and link each to main anatomy using "modify".
    * Find all observations and label them as observations
    * Link all observations to the appropriate anatomy using "located_at" link.
    * Find any observation modifiers and label as observations with modifier link to observations they modify

∗ Take note of any mentions of change (or no change) and label them appropriately
∗ Link the change entities to the rest of the graph as appropriate

– Here is how this process would be applied to the following example: "The heart is top normal in size, though this is stable".

∗ Main anatomy is heart.
∗ size is an attribute of heart, so it is also labeled as anatomy with "modify" relation to heart.
∗ normal is an observations with "located_at" relation to the main anatomy, heart. Note that top is an observation with "modify" relation to normal.
∗ stable is a change (in particular, CHAN-NC) with "modifies" relation to the main aspect which is stable, i.e. normal

B.7. Other Specific Annotation Rules

- Please follow these additional rules.

  – The main anatomy may be an adjective or may be part of a compound term. For pleural effusion, pleural is anatomy and effusion is observation. A common discrepancy was labeling the entire phrase as the observation.

  – Words such as size, volume, wall of, silhouette, structure, length, near, below, are typically attributes or modifiers of anatomy: e.g., lung volume, heart size, cardiac silhouette, osseous structure, below the diaphragm, near the apex.

  – By convention, compound words, such as cardiomegaly, which means an enlarged heart and includes both anatomy and observation semantics, should be labeled as an observation. In the same way, pneumothorax or cirrhosis are observations with the anatomic location only implied by the term itself.

  – Terms like left, right, and bilateral should be considered anatomy modifiers and labeled as anatomy. For example, in left kidney, both left and kidney are anatomy. For bilateral pleural effusions, bilateral is a modifier for the primary anatomy (pleural).

  – When an anatomy and the modifier(s) are next to each other, identify a primary term and the modifier(s) separately for consistency even if they are adjacent. For example, bilateral pleural, should be labeled as two observations, with bilateral modifying pleural. This creates consistency with the case where the two tokens are distant in text, such as bilateral small effusions or large effusions which are now bilateral.

  – A word in a sentence can sometimes modify a word in a previous sentence. For example, in the sentences, "The tube extends to the stomach. The tip is near the GE junction", the word "tip" should have a modifier link to the word "tube" in the first sentence.

  – Whenever there is any degree of uncertainty, mark the anatomy or observation as uncertain. For the phrases "No signs of pneumonia" and "No evidence of

pneumonia", pneumonia is definitely absent. For the phrases "No definite signs of pneumonia," "No clear signs of pneumonia," and "No obvious signs of pneumonia," pneumonia is uncertain due to the additional qualifier.