Spatial Information Extraction from Radiology Reports

A
DISSERTATION

PRESENTED TO THE FACULTY OF
THE UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT HOUSTON
SCHOOL OF BIOMEDICAL INFORMATICS
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

BY

SURABHI DATTA, B.E., M.S.

THE UNIVERSITY OF TEXAS HEALTH SCIENCE CENTER AT HOUSTON

2022

Dissertation Committee:

Kirk Roberts, PhD[1], Advisor
Elmer Bernstam, MD, MSE[1]
Luca Giancardo, PhD[1]
Roy F. Riascos-Castaneda, MD[2]
Hua Xu, PhD[1]

[1]School of Biomedical Informatics,
The University of Texas Health Science Center at Houston
[2]McGovern Medical School,
The University of Texas Health Science Center at Houston

created in
XeLaTeX

Dedicated to my witty friend and loving husband Sarvesh – my *Buku*, my PhD companion, and my other biscuit in our sandwich cookie

This seemed possible because I met you,
This became possible because I lived with you
– in our flowery, quiet, and worky little world.

Acknowledgments

Many people need mention here whose direct and indirect influence led to this dissertation. First, I begin by thanking my advisor Dr. Kirk Roberts who has been a great mentor to me for the last several years. Before I started my PhD, I was told by two other well-known researchers in this field that I would be in good hands. It was more than true. Getting in touch with him and being his mentee is one of the few life events I will cherish forever. He guided me in designing and carrying out all the projects that resulted in this dissertation, always steered me in the right direction, and, more importantly, supported me on innumerable occasions. I learned a lot from his extensive editing and insightful feedback. He motivated me to pursue things I never thought I could and showed me ways to work productively. He helped me overcome the mental blocks that came my way while pursuing PhD. He was like a gardener who watered and nourished me time and again to keep my progress going. I am forever indebted to his guidance, support, and help.

Second, I thank my other committee members. Dr. Elmer Bernstam inspired me through his thought-provoking questions that helped to improve the quality of my research further and to think critically from the practical standpoint of clinical practice. Dr. Luca Giancardo provided inputs that went a long way in enhancing the clarity of the projects and also in improving the reachability of research presentation. Dr. Roy F. Riascos-Castaneda validated my research from a domain expert perspective and always motivated me. Dr. Hua Xu's expert input helped improve the applicability of my research.

Third, I thank the Biomedical Informatics, Genomics and Translational Cancer Research (BIG-TCR) predoctoral fellowship training program funded by the Cancer Prevention and Research Institute of Texas (CPRIT) for providing me with support and funding in the final year of my PhD.

I also thank Dr. Xiaoqian Jiang and Dr. W. Jim Zheng for the computing resources. Without their help, timely completion of some of the projects would have been challenging. I also extend my gratitude to the School of Biomedical Informatics for providing the infrastructure and support in pursuing my doctoral studies.

Fourth, I feel fortunate to share my PhD tenure with my husband, Sarvesh Soni, who works in the same lab as mine. His support made this journey a lot easier. Together we worked late nights, discussed ideas and research, and chased paper deadlines. I also acknowledge my former lab member, Yuqi Si, whose work inspired me during my early days as a PhD student. A special thanks to all the people (especially Jordan Godfrey-Stovall, Hio Cheng Lam, Atieh Pajouhi, and Shekhar Khanpara) who helped me curate data for my research.

Much credit goes to my parents for what I could achieve. Having gone through struggles themselves and been successful with their efforts, they have an immense interest in education and understand the true meaning of achieving something with hard work and determination. I learned a lot from both of them. My father (*Baba*, Mr. Binoy Krishna Datta), an engineer by profession, guided me in my education since childhood and inspired me through his attention to detail and problem-solving skills. My mother (*Ma*, Dr. Sabita Pal Datta), who herself pursued doctoral studies amidst many difficulties, helped shape me to work toward accomplishing my goals through her energetic and proactive nature, and her strong determination.

I am also thankful to my childhood teacher and guide, Mr. Anil Baran Goswami, who still motivates me in every conversation I have with him. It is hard to imagine completing this journey without being in the shadow of his heartfelt blessings. There are many others whose inspiration and motivation touched me at different points in my life, including many friends (special thanks to Swetha Susan

George and Bhaveek Desai) and close relatives (especially my maternal uncle Dr. Nirmal Chandra Pal).

Last but not least, thanks to our little munchkin Shivik, whose arrival to this world, and our lives, poured a sweet and happy feeling to the culmination of my PhD journey.

Dissertation advisor: Kirk Roberts, PhD Surabhi Datta, M.S.

Spatial Information Extraction from Radiology Reports

Abstract

Radiology reports contain a radiologist's interpretations of images, and these images frequently describe spatial relations between radiological entities. Important radiographic findings are mostly documented in the reports in reference to an anatomical structure along with other clinically-relevant contextual information through spatial expressions. The spatially-grounded radiological entities mainly include clinical findings and medical devices. Although much work has focused on radiology information extraction, spatial language understanding in the radiology domain has remained less explored. The language used for representing spatial relations is complex and varied. Therefore, we aim to encode granular spatial information in the reports and automatically extract this important information using natural language processing (NLP) methods. Structured representation of this clinically significant spatial information has the potential to be used in a variety of downstream clinical applications. Such applications include fine-grained phenotyping for clinical trials and epidemiological studies, automated tracking of clinical findings and devices, and automatic image label generation. The three broad aims of this dissertation are to–1) build a robust spatial representation schema that can encode detailed spatial information of findings and devices, 2) develop state-of-the-art deep learning-based NLP methods to automatically extract the spatial information, and 3) develop clinical informatics applications using the spatial information extracted from reports. First, we define two spatial representation schemas, Rad-SpRL and Rad-SpatialNet, that are based on spatial role labeling and frame semantics, respectively. We construct manually annotated radiology report datasets following these schemas. We then propose transformer-based language models to

automatically identify the spatial information from these reports where we frame the extraction problem as both sequence labeling and question answering. To enable downstream applications, we also propose normalization methods to map the radiological entities in the reports to standard concepts in RadLex, a publicly available radiology lexicon. In addition to this, we also propose a weak supervision method to automatically create a large radiology training dataset for spatial information extraction without using any manual annotations. Further, we extend the Rad-SpatialNet schema to encode spatial language in a different domain, i.e., ophthalmology notes. Finally, we use the information extracted from radiology reports to develop an ischemic stroke phenotyping system and an automated radiology tracking system that aims to track the same radiological findings and medical devices across reports.

Vita

2014..... B.E., Computer Science, Birla Institute of Technology & Science, Pilani, India

2018..... M.S., Biomedical Informatics, The University of Texas Health Science Center at Houston

2016–2021 ..... Graduate Research Assistant, School of Biomedical Informatics, The University of Texas Health Science Center at Houston

2021–2022 ..... CPRIT BIG-TCR Predoctoral Fellow, School of Biomedical Informatics, The University of Texas Health Science Center at Houston

Publications

Datta, S., & Roberts, K. (2022). *Fine-Grained Spatial Information Extraction in Radiology as Two-turn Question Answering*. International Journal of Medical Informatics, 158, 104628.

Datta, S., Si, Y., Rodriguez, L., Shooshan, S. E., Demner-Fushman, D., & Roberts, K. (2020). *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.

Datta, S., Bernstam, E.V., & Roberts, K. (2019). *A frame semantic overview of NLP-based information extraction for cancer-related EHR notes*. Journal of Biomedical Informatics, 103301.

Datta, S., & Roberts, K. (2019). *A dataset of chest X-ray reports annotated with Spatial Role Labeling annotations*. Data in Brief, 32, 106056.

Datta, S., Lam, H. C., Pajouhi, A., Mogalla, S., & Roberts, K. (2022). *A Cross-document Coreference Dataset for Longitudinal Tracking across Radiology Reports*. Proceedings of the 13th Conference on Language Resources and Evaluation, 3686–3695.

Datta, S., Khanpara, S., Riascos-Castaneda, R. F., & Roberts, K. (2021). *Leveraging Spatial Information in Radiology Reports for Ischemic Stroke Phenotyping*. AMIA Summits on Translational Science Proceedings, 170-179.

Datta, S., Godfrey-Stovall, J., & Roberts, K. (2020). *RadLex Normalization in Radiology Reports*. AMIA Annual Symposium Proceedings, 338-347.

Datta, S., Ulinski, M., Godfrey-Stovall, J., Khanpara, S., Riascos-Castaneda, R. F., & Roberts, K. (2020). *Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports*. Proceedings of the 12th Language Resources and Evaluation Conference, 2251-2260.

Datta, S., & Roberts, K. (2020). *Spatial Relation Extraction from Radiology Reports using Syntax-Aware Word Representations*. AMIA Summits on Translational Science Proceedings, 116-125.

Datta, S., & Roberts, K. (2020). *A Hybrid Deep Learning Approach for Spatial Trigger Extraction from Radiology Reports*. Proceedings of Workshop on Spatial Language Understanding, 50-55.

Wu, S., Roberts, K., Datta, S., ... & Xu, H. (2019). *Deep Learning in Clinical Natural Language Processing: A Methodical Review*. Journal of the American Medical Informatics Association, 27(3), 457–470.

Grants and Awards

CPRIT BIG-TCR Predoctoral Fellowhip at School of Biomedical Informatics UTHealth, 2022.

Best Presentation at AMIA 2021 NLP-WG Symposium Doctoral Consortium, 2021.

Dr. Noriaki Aoki Scholarship at School of Biomedical Informatics UTHealth, 2021.

Second Prize in Student Paper Competition at AMIA 2021 Informatics Summit, 2021.

James T. Willerson Endowed Scholarship at School of Biomedical Informatics UTHealth, 2020.

School of Biomedical Informatics Excellence Scholarship, 2020.

Dr. Noriaki Aoki Fund Travel Award at School of Biomedical Informatics UTHealth, 2019.

Field of Study

Biomedical Informatics

Contents

Listing of Tables

Listing of Figures

1

Introduction

Radiology is a field of medicine that uses medical images to diagnose diseases and guide their treatments. Medical imaging is used for disease prevention through screening, identifying abnormalities, disease staging, facilitating decision support, evaluating patient's progress during treatment, and prognosis (Brady et al., 2021). The market size of the medical imaging industry is more than 100 billion dollars, and there is a growing number of imaging procedures conducted annually, which is close to 700 million in 2021 (Levin & Janiga, 2021). As a result, a large volume of radiology reports are generated per year.

Radiology reports are one of the most important sources of medical information about a patient and are thus one of the most-targeted data sources for natural language processing (NLP) in medicine (Pons et al., 2016). These reports describe a radiologist's interpretation of one or more two-, three- or four-dimensional images (e.g., *X-ray*, *computed tomography*, *magnetic resonance*

imaging, *ultrasound*, *positron emission tomography*). As a consequence, these reports are filled with spatial relationships between medical findings (e.g., *tumor*, *pneumonia*, *inflammation*) or devices (e.g., *tube*, *stent*, *pacemaker*) and anatomical location (e.g., *right ventricle*, *chest cavity*, *T4*, *femur*). Besides radiology-specific knowledge and experience, interpreting spatial relations from radiological images requires good spatial ability skills on the part of radiologists as it often involves mental visualization of complex 3D anatomical structures to describe the locations of radiographic findings. In this context, a few studies have highlighted the possible requirement of these skills in prospective radiologists to perceive and understand the spatial relationships between different objects in radiology practice (Birchall, 2015; Corry, 2011).

The spatial relations encountered in radiology text provide sufficient contextual information related to the findings and devices. Moreover, some of these spatially-grounded findings demand immediate action by the physician ordering the imaging examination. Therefore, it is important to understand the spatial meanings from the unstructured reports and generate structured representations of the spatial relations for various downstream clinical applications. Such applications include easy visualization of the important actionable findings, predictive modeling, cohort retrieval, automated tracking of radiological findings, and automatic generation of more complete annotations for associated images containing spatial and diagnosis-related information of findings. Although numerous work has focused on information extraction (IE) from radiology reports (Cornegruta et al., 2016; Hassanpour & Langlotz, 2016; Annarumma et al., 2019; Wang et al., 2017), spatial language understanding in the radiology domain has still remained less explored and forms the focus of this dissertation. For this, we aim to first define spatial representation schemas that can capture various spatial and contextual information from the report text and later use these schemas to build natural language processing (NLP) systems for automatically extracting such spatial information.

Table 1.1 shows an example radiology report. The example demonstrates the large number of findings and the relationships with various anatomical entities found within radiology reports.

Table 1.1: Example radiology report containing spatial language. This report is publicly available without restriction from openi.nlm.nih.gov (image ID: CXR1000_IM-0003-1001). Findings are in green, anatomical locations are in blue, while the spatial expressions are in cyan. *Note*: XXXX corresponds to phrases stripped by the automatic de-identifier.

Comparison: XXXX PA and lateral chest radiographs Indication: XXXX-year-old male, XXXX. Findings: There is XXXX increased opacity within the right upper lobe with possible mass and associated area of atelectasis or focal consolidation. The cardiac silhouette is within normal limits. XXXX opacity in the left midlung overlying the posterior left 5th rib may represent focal airspace disease. No pleural effusion or pneumothorax. No acute bone abnormality. Impression: 1. Increased opacity in the right upper lobe with XXXX associated atelectasis may represent focal consolidation or mass lesion with atelectasis. Recommend chest CT for further evaluation. 2. XXXX opacity overlying the left 5th rib may represent focal airspace disease.

In this dissertation, we propose two spatial representation schemas to encode spatial language in radiology text. The first is based on the Spatial Role Labeling (SpRL) scheme, which we refer to as Rad-SpRL. In Rad-SpRL, common radiological entities tied to spatial relations are encoded through four spatial roles: TRAJECTOR, LANDMARK, DIAGNOSIS, and HEDGE, all identified in relation to a spatial preposition (or SPATIAL INDICATOR). The second schema is based on the principle of frame semantics. Frame semantics provide a useful way to represent information in text and has been utilized in constructing semantic frames to encode spatial relations (Petruck & Ellsworth, 2018). Specifically, we extend the SpatialNet framework in the general domain (Ulinski et al., 2019) that utilizes FrameNet (Baker, 2014) with the aim to generate more accurate representations of spatial language used by radiologists. We refer to this schema as Rad-SpatialNet. Thus, Rad-

SpRL is a basic schema and captures four main spatial roles that different radiological entities play in a sentence, whereas Rad-SpatialNet is an advanced representation that incorporates more spatial details in the text.

As described above, there exists spatial relations between findings/devices and anatomical locations in the reports. There are also mentions of other clinically relevant contextual details associated to the spatial relations such as potential diagnoses and a device's distance from the anatomical structure. Moreover, there are spatial and other descriptors describing a radiological entity (e.g., finding, anatomy) that enhance the richness of the labels for the corresponding medical images. The spatial descriptors represent both spatial (e.g., laterality, size, morphology) and other properties of an imaging observation (e.g., composition, distribution pattern, density) described in reports. Other descriptors include status, quantity, temporality, and negation. Our second schema–Rad-SpatialNet organizes all these clinically important information following frame semantics. In frame semantics, a lexical unit (LU) is the word or phrase that invokes a frame and the participants of a frame constitute the frame elements (FEs). In the context of spatial relation frames, an LU is either a spatial preposition/verb (which we refer to as a "trigger") or a radiological entity, and all the spatial roles and descriptors linked to the LU form the FEs. We refer to the spatial roles (connected to a spatial trigger) and the spatial descriptors (connected to a radiological entity) as "Spatial Frame Elements" (SFEs). The other entity-specific descriptors are referred to as "Descriptive Frame Elements" (DFEs). Some of these frame elements are illustrated in Figure 1.1.

The second broad focus of this dissertation is to develop advanced NLP methods that can automatically identify the spatial information from the reports with high performance. We propose two methods for IE - the first is based on sequence labeling approach and the second is based on framing the extraction task as question answering (QA). Owing to its promising performance both for sequence labeling and QA, we predominantly use the transformer-based pre-trained language model BERT (Devlin et al., 2019) in our proposed methods for spatial IE.

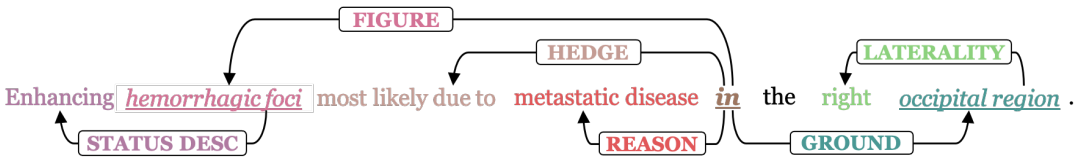

Figure 1.1: Spatial and descriptive frame elements in radiology text. Figure, Ground, Hedge, and Reason are the spatial frame elements of the frame instantiated by the spatial trigger *in*. *Enhancing* denotes a status descriptor and is a descriptive element of the frame evoked by the finding entity *hemorrhagic foci* whereas *right* is the laterality and is a spatial frame element of the frame evoked by the anatomical entity *occipital region*. The *underlined and italicized* texts indicate the lexical units of the frames.

More specifically, for the sequence labeling approach, we perform spatial IE in two steps. In the first step, we identify the spatial indicators (in Rad-SpRL) or spatial triggers (in Rad-SpatialNet) from a report sentence. In the second step, we identify the corresponding spatial roles (in Rad-SpRL) or frame elements (in Rad-SpatialNet) given an indicator/trigger in that sentence. Both these steps are formulated as sequence labeling task where each sentence is fed into the BERT model as input and the contextual representations from the BERT encoder output is used to get the final predicted labels for each word in a sentence. For the second approach, that is, IE as QA, we propose a multi-turn QA method (specifically, two-turn) to identify the spatial information. This is inspired by the recent studies (Levy et al., 2017; Li et al., 2020b; Liu et al., 2020) that demonstrated the advantages of employing a QA framework over other traditional methods for IE problems as well as by studies (Li et al., 2019, 2020a; Wang et al., 2020) that highlighted the advantages that multi-turn QA provides. Our two-turn QA approach identifies spatial and descriptor information by answering queries given a radiology report text. We frame the extraction problem such that all the main radiology entities (e.g., finding, device, anatomy) and the spatial trigger terms (denoting the presence of a spatial relation between finding/device and anatomical location) are identified

in the first turn. In the subsequent turn, various other contextual information that acts as spatial roles with respect to a spatial trigger term are extracted along with identifying the spatial and other descriptor terms qualifying a radiological entity. The queries are constructed using separate templates for the two turns and we employ two query variations in the second turn.

In order to enable the use of the extracted radiological entities in downstream clinical applications that can work across multi-institutional reports, it is crucial to map the entities to concepts in a standardized vocabulary of radiology terms. This process of mapping the entity spans in text to standard concepts in a vocabulary is known as concept normalization. There is limited research in this direction, and, therefore, in this dissertation, we attempt to to normalize a diverse set of radiological entities to RadLex (Langlotz, 2006) terms. For this, we first manually construct a normalization corpus by annotating entities from three types of radiology reports. This corpus contains a total of 1706 entity mentions. We then propose two BERT-based methods for automatic normalization.

To examine the generalizability of Rad-SpatialNet, we extend this schema to a different domain, that is, ophthalmology, to represent spatial language in ophthalmology notes. We update the Rad-SpatialNet schema with additional frame elements and this resulted in a new schema for ophthalmology– Eye-SpatialNet. We annotate 600 ophthalmology notes with detailed spatial and contextual information of ophthalmic entities and apply our previously described two-turn QA approach to automatically extract spatial information from the notes.

Supervised deep learning methods rely on a large amount of labeled data to provide satisfactory performance. However, manual annotations are expensive, time-consuming, and requires domain expertise. To this end, weak supervision approaches provide ways to automatically create large labeled data without any manual involvement (Ratner et al., 2020; Shang et al., 2018; Safranchik et al., 2020). More specifically, we propose a data programming-based weak supervision method to automatically create a large labeled dataset of radiology reports for spatial IE.

The final aspect of this dissertation is to use the important clinical information including spatial information extracted from the radiology reports to develop useful clinical informatics applications. We particularly develop two applications–phenotyping and automated tracking. For the phenotyping application, we focus on ischemic stroke phenotypes with location-specific information: brain region affected, laterality, stroke stage, and lacunarity. We first use our BERT-based Rad-SpatialNet NLP system to identify the clinically important spatial information and then apply simple domain rules on top of the extracted information to classify the stroke phenotypes.

Our automated tracking problem is framed as cross-document coreference resolution (CDCR) task. For this, we propose a new CDCR dataset to identify the co-referring radiological findings and medical devices across a patient's radiology reports. Our annotated dataset contains 5872 mentions (findings and devices) spanning 638 MIMIC-III radiology reports across 60 patients, covering multiple imaging modalities and anatomies. We propose two methods–string matching (used as the baseline method) and a BERT-based method to identify the cross-report coreferences.

The following chapters in this dissertation are organized as follows. Chapter 2 provides a detailed description of previous work on information extraction in the radiology domain, normalization and weak supervision in the medical domain, as well as applications that use radiology report information. Chapter 3 discusses our two proposed spatial representation frameworks (Rad-SpRL and Rad-SpatialNet) designed to encode spatial information in radiology text along with descriptions of the annotated datasets and the annotation processes. This chapter largely corresponds to work published in Datta et al. (2020a,b). Chapter 4 describes our proposed deep learning-based NLP methods to extract spatial information from report text. This chapter also includes the evaluation setup and the performance (average of 10-fold cross validation in most cases) of our proposed methods and corresponds to three published studies (Datta et al., 2020a,b; Datta & Roberts, 2022). Chapter 5 , that corresponds to work published in Datta et al. (2021a), describes our radiology normalization corpus, our proposed BERT-based systems developed for automatic normalization

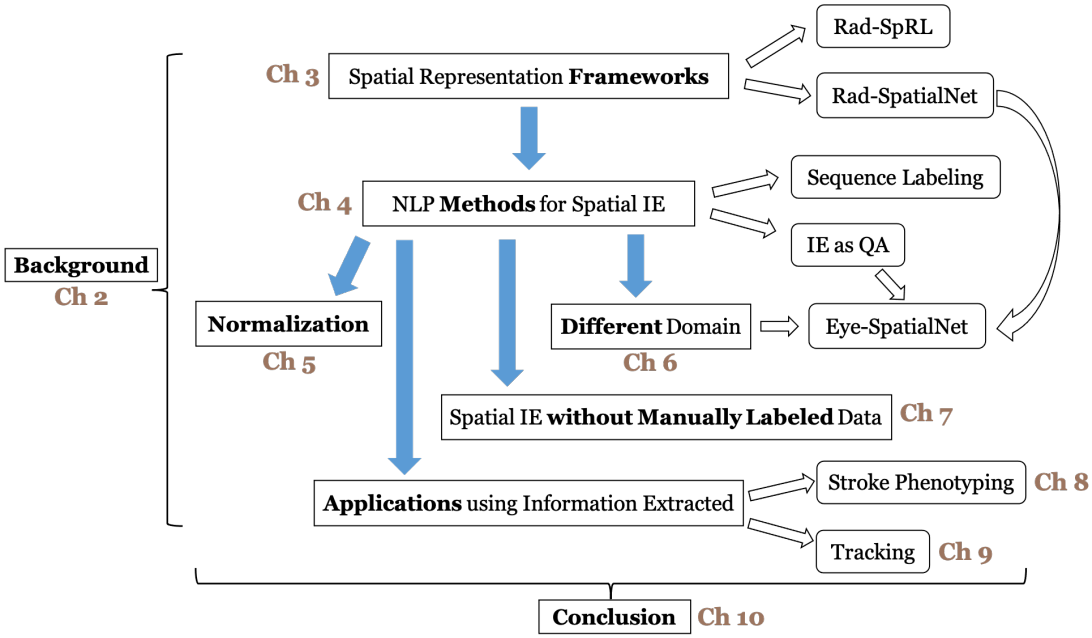

Figure 1.2: All the chapters following the Introduction. Ch - Chapter. NLP - Natural Language Processing. IE - Information Extraction. QA - Question Answering.

of radiological entities to RadLex concepts as well as the system performance. Chapter 6 describes the Eye-SpatialNet schema and the results of our two-turn QA method on ophthalmology notes. Chapter 7 provides a detailed description of our proposed weak supervision approach for radiology spatial IE. Chapters 8 and 9 correspond to work in Datta et al. (2021b) and Datta et al. (2022), respectively, where the former discusses the ischemic stroke phenotyping application and the latter discusses the automated tracking system. The organization of the chapters in this dissertation is shown in Figure 1.2.

2

Background

This chapter summarizes the prior work related to natural language processing (NLP) in the field of radiology with specific focus on spatial information extraction. The following sections describe existing studies on spatial representation frameworks to encode spatial language in text, NLP methods for information extraction, medical entity normalization, weak supervision for information extraction, and a few informatics applications that leverage the important clinical information from radiology reports. Each section also highlights the research gaps and limitations in existing work that serve as the motivation behind the proposed work for the subsequent chapters.

2.1 Spatial Representation Framework for Text

Different representation frameworks have been proposed to encode spatial knowledge in textual data for various use cases. Among the early works, Hayward & Tarr (1995) investigated the structural similarities between visual and linguistic representations of space. Mani et al. (2010) proposed SpatialML to represent geographical location information including geo-coordinates and orientation and annotated ACE English documents as per SpatialML. This representation encodes the spatially-related entities through contextual information such as direction and distance as well as the actual physical connection between the related entities (using the Region Connection Calculus). However, this representation is specific to the geographical aspects of the spatial language. At the same time, Kordjamshidi et al. (2010) proposed Spatial Role Labeling (SpRL) that involves extracting spatial arguments of the spatial relations in a sentence. This framework is an improvement over representations such as SpatialML and STM spatio-temporal markup (Pustejovsky & Moszkowicz, 2008) as this is more generalizable in terms of spatial language expressiveness and handles a greater number of spatial concepts (both static and dynamic). This has also been utilized on biomedical (Kordjamshidi et al., 2015) and consumer health data (Roberts et al., 2015). Later, Guadarrama et al. (2013) proposed a system where users can interact with robots by issuing commands or asking queries. Here, the system learns to both recognize the objects in a shared environment and identify the spatial relationships between the objects. Fasola & Mataric (2013) also devised methods to represent dynamic spatial relations for facilitating interactive instruction of robots. For text-to-scene generation, Coyne & Sproat (2001) and Coyne et al. (2010) developed and improved the WordsEye system that automatically converts natural language text into 3D scenes. Later, Chang et al. (2014) proposed a representation that converts an input text describing a scene to output a 3D scene by transforming the text to a set of constraints consisting of the objects and the spatial relations between them as well as by learning priors on how the objects occur in 3D scenes. Yuan

(2011) applied a shortest path dependency kernel for support vector machine (SVM) to identify spatial relations from text for geographic information retrieval. Kergosien et al. (2015) designed a framework to extract relevant spatial information from web textual data (newspaper articles) to annotate satellite images with additional meaningful information for use cases such as image annotation and land use planning. Collell & Moens (2018) used both visual and linguistic features to generate distributed spatial representations by feeding them into a neural network model that learns to predict 2D spatial arrangements of objects provided their instances and the relationship between them. More recently, Ulinski et al. (2019) designed the SpatialNet framework to encode spatial language based on frame semantic principles and additionally proposed ways to incorporate external knowledge sources for disambiguating the spatial expressions. All these highlight some important works relevant to spatial information representation in text.

2.2 Information Extraction from Radiology Reports

A range of existing work has focused on extracting isolated radiological entities (e.g., findings/locations) utilizing NLP in radiology reports (Hassanpour & Langlotz, 2016; Hassanpour et al., 2017; Cornegruta et al., 2016; Bustos et al., 2019; Annarumma et al., 2019). There exists recently published radiology image datasets labeled with important clinical entities extracted from corresponding report text (Wang et al., 2017; Irvin et al., 2019). In Table 2.1, we compare the specific information types or the radiology entities extracted in the previous studies from chest radiology reports using NLP. We primarily pay attention to the clinically-important entities which are common across various types of radiology reports. We also do not take into account the cases where uncertainty and negation information were used to detect the presence or absence of a particular finding or a disease (Wang et al., 2017). For example, *Hedge* is not considered as extracted in Table 2.1 when the uncertainty levels are classified into negative, uncertain or positive for each finding term extracted (Irvin et al.,

Table 2.1: Comparison of our corpus with the studies extracting clinically-relevant radiological entities from chest X-ray and chest CT reports.

Paper	Finding	Anatomy	Descriptor	Diagnosis	Device	Hedge	Negation	Relation
Hassanpour et al. Hassanpour & Langlotz (2016)	✓	✓	✓			✓	✓	
Cornegruta et al. Cornegruta et al. (2016)	✓	✓	✓		✓		✓	
Bustos et al. Bustos et al. (2019)	✓	✓		✓				
Hassanpour et al. Hassanpour et al. (2017)	✓							
Irvin et al. Irvin et al. (2019)	✓			✓	✓		✓	
Annarumma et al. Annarumma et al. (2019)	✓	✓	✓					
Wang et al. Wang et al. (2017)	✓			✓				

2019). Further, in Table 2.1, we have not considered studies dealing with specific body locations (e.g., mammography reports containing breast imaging information, and head CT reports) as the entities of interest are usually very domain-specific such as 'Clock face', 'Depth', 'BI-RADS category' etc. in the case of mammography reports. We also do not take into account the works which focused on detecting a specific disease such as pneumothorax (Wang et al., 2019b) or pulmonary lesion (Pesce et al., 2019) from chest radiographs. Note that most of these studies have targeted toward entity extraction from text without focusing on recognizing relations among these entities.

Among the studies that extracted relations, Friedman et al. (1994) proposed a formal model (MedLEE) based on grammar rules to map clinical information in radiology reports, including central findings and their contextual information like body location, degree, and certainty modifiers into a structured format utilizing controlled vocabulary and synonym knowledge base. They also worked toward providing an interface for using MedLEE for different applications (Friedman et al., 1995). In another work, Friedman et al. (2004) adapted MedLEE to generate the most specific Unified Medical Language System (UMLS) code based on a finding and its associated modifier information. Later, Sevenster et al. (2012) built a reasoning engine to correlate clinical findings and body locations in radiology reports utilizing the Medical Language Extraction and Encoding System (MedLEE). However, the major limitation of this work is the system's poor recall. Yim et al.

Table 2.2: Studies focusing on spatial relations in radiology reports.

Paper	Finding	Anatomy	Diagnosis	Hedge
Roberts et al. Roberts et al. (2012)	—	✓ (Spatially related to a finding)	—	—
Rink et al. Rink et al. (2013)	✓	✓ (Linked with finding)	—	✓ (Not linked with finding/location)

(2016) worked on extracting relations containing tumor-specific information from radiology reports of hepatocellular carcinoma patients. Steinkamp et al. (2019) extracted facts representing clinical assertions and recognized contextual information such as location, image citation, and description of change over time related to a target entity (e.g., finding) identified for that fact. However, this system does not necessarily capture the related entities from a spatial perspective and does not identify all the fine-grained spatial information. Another work (Alex et al., 2019) identified relations between observation entities with their location (deep/cortical) and recency (old/recent) modifiers from brain imaging reports. However, the location information includes two broad categories and is relevant to two specific observations (stroke and microbleed). In Table 2.2, we present the two works relevant to spatial information extraction from radiology reports. The main limitations of Rink et al. (2013) are the usage of appendicitis-specific lexicons and the requirement of manual effort in crafting rules based on syntactic dependency patterns to identify the spatially-grounded inflammation description. Besides being domain-specific, another limitation of Roberts et al. (2012) is that the study extracts only the location entities associated with an actionable finding and this required relying on heavy feature engineering. Thus, we see that relatively few studies have focused on relation extraction from radiology text. Moreover, the datasets were limited to specific report types (e.g., hepatocellular carcinoma) and the relations extracted do not capture spatial information.

An important dataset for radiology NLP is the Open-i radiology report dataset (Demner-Fushman et al., 2016). Open-i is a biomedical image search engine*. One of its data collections

*https://Open-i.nlm.nih.gov/

Indication: Abdominal pain and distention.

Findings: Frontal and lateral views of the chest show an unchanged cardiomediastinal silhouette. There is bibasal interstitial opacity and left basal platelike opacity XXXX due to discoid atelectasis and/or XXXX scarring. **There are emphysematous changes, particularly within the right upper lobe.** No XXXX focal airspace consolidation or pleural effusion.

Impression: 1. COPD. Basilar probable pulmonary fibrosis and scarring. 2. No acute cardiac or pulmonary disease process identified.

Manual Annotation

- Opacity/lung/base/bilateral/ interstitial
- Pulmonary Atelectasis/base/left
- Cicatrix/lung/base/left
- Pulmonary Emphysema
- Pulmonary Disease, Chronic Obstructive
- Pulmonary Fibrosis/base

Figure 2.1: A sample of radiology report annotations in Open-i dataset.

is a public chest X-ray dataset containing 3955 de-identified radiology reports from the Indiana Network for Patient Care released by the National Library of Medicine. We have presented an example of the manual annotation of a sample report in the Open-i dataset in Figure 2.1 (the annotations are inspired by MeSH terms). Although most of the Open-i annotations embody the relationship between finding and location, there are, however, a few missing relations. For example, note that in Figure 2.1 the Open-i manual annotations contain the normalized finding *Pulmonary Emphysema* corresponding to the phrase *'emphysematous changes'* in the report, but do not annotate the associated location *'right upper lobe'*. Although many studies have used the Open-i dataset, most of them focused on the extraction of only the disease/finding (Wang et al., 2017, 2018; Peng et al., 2018; Daniels & Metaxas, 2019; Zech et al., 2018). Two studies worked on automatically annotating

both disease and disease descriptions (e.g., location, severity) (Shin et al., 2016; Huang et al., 2019) similar to the human annotations in Demner-Fushman et al. (2016). However, all these works ignored distinguishing diagnosis terms from findings (except for Peng et al. (2018)), and annotating correlations between them. We describe annotation-specific limitations of each of these works in Table 2.3.

Table 2.3: Studies who have used Open-i manual annotations.

Paper	How Open-i chest X-ray dataset is involved	Limitation (Radiology entities annotated/considered for model evaluation)
Demner-Fushman et al. (2016)	Manually annotated or coded the collected reports with findings, diagnoses, body parts using MeSH terms supplemented by RadLex codes. Automatic annotation was also produced by the Medical Text Indexer (MTI).	This is a manual annotation process relying on MeSH terms and standard qualifier terms. The coded terms were not well-distinguished between findings and diagnoses. Moreover, the annotation lacks other information such as relation between findings and diagnoses. The automatic labeling does not include the related body parts for the labeled finding. (**Positive Findings/Diagnoses and Body parts**)
Shin et al. (2016)	Trained CNNs using existing image annotations from Demner-Fushman et al. (2016) and considered images labeled with a single disease using unique MeSH term combinations (this accounted for around 40 percent of the full Open-i dataset and 17 unique disease annotation patterns). Generated image annotations including disease as well as its contexts such as location, severity, and the affected organs by taking into account image/text contexts while training CNNs.	Although the annotation includes disease context and that way it generates different image captions based on severity/location contexts, it is limited to one major disease provided an image. (**Findings/Diagnoses and their context such as location and severity**)

Wang et al. (2017)	Used text mining, DNorm Leaman et al. (2015) and MetaMap Aronson & Lang (2010), to label disease names using reports. Evaluated their image labeling method on Open-i reports using the key findings/disease names coded by human annotators as gold standard (Demner-Fushman et al., 2016). Note that additional datasets are also used.	Only used the available annotations for evaluating their proposed method. (**Findings/Diagnoses**)
Wang et al. (2018)	Evaluated a text-image embedding auto-annotation framework on the Open-i dataset using the key findings/disease names coded by human annotators as the gold standard (Demner-Fushman et al., 2016). Additional datasets are also used.	Used the annotated Open-i dataset for evaluating proposed disease classification method for 14 diseases. (**Findings/Diagnoses**)
Peng et al. (2018)	Defined rules utilizing universal dependency graphs to identify negation or uncertainty related to findings. Manually checked the annotations in Open-i and organized the findings into 14 domain-important and generic types of medical findings.	Used the Open-i dataset both for designing the patterns and testing. Although they mentioned that organizing the findings into fine-grained categories can facilitate in correlating findings with the diagnosis, the terms distinguished as diagnoses or body parts were not utilized in the study for showing any correlation. (**Findings**)
Daniels & Metaxas (2019)	Proposed a deep neural network that predicts one or more diagnoses given an image by jointly learning visual features and topics from report findings.	Used the Open-i dataset and their corresponding 'findings' annotations both for fine-tuning and evaluating the model. (**Findings/Diagnoses**)

Huang et al. (2019)	Proposed a neural sequence-to-sequence model by leveraging "indication" information of the report which includes annotating the relationship between the positions where the finding term appears. They used the Open-i manual annotations as a reference annotation for evaluating the model.	Although this generated annotations for multiple diseases per image and also aimed to improve the results of Shin et al. (2016) in annotating disease along with context such as location and severity, they did not annotate other useful contexts including spatial information of the finding as well as the associated diagnosis. (**Findings/Diagnoses and their context such as location and severity**)
Zech et al. (2018)	To assess the generalizability of a deep learning model for screening pneumonia across 3 hospital systems. Used human-annotated pathology labels of the Open-i dataset for testing.	Used Open-i only for evaluation. (**Findings/Diagnoses**)
Candemir et al. (2018)	Fine-tuned several deep CNN architectures to detect presence of cardiomegaly. Used Open-i dataset both for training and testing.	Manually annotated each Open-i image into one of the following severity categories: borderline, mild, moderate, severe, and non-classified using the corresponding reports having cardiomegaly. (**Findings, specifically cardiomegaly and their severity levels**)

2.3 Medical Concept Normalization

2.3.1 Data

The NCBI disease corpus, consisting of 793 PubMed abstracts, was annotated for disease name normalization (Doğan et al., 2014). There exists an annotated dataset of narrative clinical reports for the normalization task of disorders as part of ShARe/CLEF eHealth 2013 challenge[†]. Disorder normalization corpora constructed from MIMIC clinical notes are also available through SemEval-2014 Task 7[‡] and SemEval-2015 Task 14[§]. A few studies created medical concept normalization corpora for mapping user generated text on social media to standard vocabularies like SNOMED. CADEC consists of annotated concepts from 1253 social media posts taken from AskaPatient[¶] associated with adverse drug events (ADEs) of patients (Karimi et al., 2015). PsyTAR, also constructed from AskaPatient, contains 887 patient posts annotated with ADEs related to psychiatric medications (Zolnoori et al., 2019). Sarker et al. (2018) developed an annotated corpus for normalizing expressions denoting adverse drug reactions (ADRs) from Twitter to MedDRA (Brown et al., 1999) (Medical Dictionary for Regulatory Activities) Preferred Terms (PTs), which was released in the shared task – Social Media Mining for Health (SMM4H). Roberts et al. (2017) also released an annotated dataset of 200 drug labels for TAC2017 where the ADR expressions in the labels were mapped to MedDRA Lower Level Terms and PTs. Luo et al. (2019b) has also released an annotated corpus of 100 discharge summaries in a 2019 shared task covering entities corresponding to medical problems, treatments, and tests. Previous works have mapped the concept mentions to ontologies such as the Unified Medical Language System (UMLS) (Bodenreider, 2004), SNOMED, RxNorm, and MedDRA. Thus, we note that no work has focused on constructing normalization corpus from

[†] https://sites.google.com/site/shareclefehealth/
[‡] http://alt.qcri.org/semeval2014/task7/
[§] http://alt.qcri.org/semeval2015/task14/
[¶] https://www.askapatient.com/

radiology reports and mapping the important radiological entity spans to RadLex codes.

2.3.2 Methods

Some of the first deep learning approaches for concept normalization in the medical domain were based on convolutional neural networks (CNNs) and recurrent neural networks (RNNs) where a user phrase was converted to a semantic vector representation and eventually a softmax classifier was used to assign a standard medical concept to that phrase (Limsopatham & Collier, 2016; Tutubalina et al., 2018; Han et al., 2017). Tutubalina et al. (2018), however, incorporated additional semantic similarity features by leveraging prior domain knowledge (UMLS) to further enrich the phrase representations. Miftahutdinov & Tutubalina (2019) used contextualized word representations such as BERT and ELMo (Peters et al., 2018) for normalizing user generated phrases and achieved state-of-the-art performance on three benchmark normalization datasets – CADEC, PsyTAR, and SMM4H 2017. All these papers worked on user-generated text of social media posts and formulated normalization as a multi-class classification task. Luo et al. (2019a) has proposed a hybrid system by combining exact match, edit-distance, and deep learning methods for normalizing entities in the ShARe/CLEF 2013 challenge dataset. Their model architecture additionally integrated contextual information of an entity mention (left and right context words) and predicted the UMLS code using a softmax classification layer. Ji et al. (2019) have used BERT as a ranking model in a normalization task. They ranked the candidate concepts after generating them using the BM25 (Robertson et al., 1996) information retrieval method. Their BERT-based ranker outperformed the previous best results on ShARe/CLEF, NCBI, and TAC2017ADR normalization datasets. Moreover, BERT-based re-ranking has been shown to perform well on other information retrieval tasks such as passage retrieval (Nogueira & Cho, 2019).

To the best of our knowledge, due to the lack of annotated radiology corpus, no study so far has applied supervised learning techniques for radiology concept normalization. Tahmasebi et al.

(2019) has utilized an unsupervised semantic learning approach to normalize the anatomical phrases in the radiology reports to SNOMED CT anatomical concepts. However, their work was limited to normalizing only the anatomical terms and did not cover other commonly observed clinically-significant information such as clinical findings and modifier terms.

2.4 Weak Supervision in the Medical Domain

Numerous work has focused on open-domain NLP tasks using weak supervision. Many studies (Shang et al., 2018; Fries et al., 2017; Safranchik et al., 2020; Li et al., 2021; Lison et al., 2020; Zeng et al., 2020; Zhao et al., 2021) have proposed weak supervision methods for named entity recognition, and a few for other tasks such as natural language generation and understanding (Chang et al., 2021) and discourse structures (Badene et al., 2019). Recently, there has been increasing work on automatically creating training data and adopting weakly supervised machine learning (ML) methods for NLP tasks in the clinical domain. Wang et al. (2019a) developed a rule-based NLP method to create labels for training ML models to classify clinical text. Cusick et al. (2021) proposed a rule-based approach based on NegEx to generate training labels for identifying current suicidal ideation. Dong et al. (2021) adapted a weak supervision approach with rules and contextualized representations to identify rare diseases. Shen et al. (2021) adopted a similar weak supervision approach with BERT where they used a rule-based NLP method to automatically generate training labels for classifying lifestyle factors for Alzheimer's disease. Banerjee et al. (2019) proposed a weak supervision method where domain-specific dictionaries are used to heuristically generate training labels to classify evidence of urinary incontinence and bowel dysfunction. Callahan et al. (2019) employed data programming and trained LSTM networks for identifying pain-anatomy and implant-complication relations from clinical notes. Peterson et al. (2020) trained a BERT model using weakly labeled data generated through data programming to classify relations (e.g., severity, stage,

etc.) that can be mapped to FHIR representations. Fries et al. (2021a) utilized data programming with BioBERT to classify medical entities and demonstrated comparable results to fully supervised models on multiple benchmark datasets. Very recently, Humbert-Droz et al. (2022) developed a data programming-based weak supervision pipeline using Snorkel to generate weak labels for identifying the presence or absence of symptoms. Moreover, in the biomedical domain, multiple studies have used the Snorkel framework for extracting chemical reaction relationships from biomedical abstracts (Mallory et al., 2020), biomedical relation extraction (Krasakis et al., 2019), and filtering biomedical research articles as relevant or non-relevant for drug repurposing in cancer (Dua et al., 2021).

We see that most studies in the clinical domain use a rule-based approach to create weak labels for binary classification tasks. Although Peterson et al. (2020) identified different relations associated with a problem description, their approach assumes a single clinical problem in a description. Moreover, only a few studies (Dunnmon et al., 2020; Wang et al., 2019a; Eyuboglu et al., 2021) so far have applied weak supervision on radiology report text, two of those for binary classification problems (classifying a report as normal vs abnormal (Dunnmon et al., 2020) and identifying hip fracture from report (Wang et al., 2019a)) while another (Eyuboglu et al., 2021) to generate weak anatomical region labels that are subsequently used for training imaging models.

2.5 Applications using Radiology Information

2.5.1 Phenotyping

Numerous work has focused on identifying certain subgroups of stroke patients using NLP techniques with the aim to facilitate timely patient triaging to select appropriate group of patients highly likely to encounter severe consequences. We highlight the relevant studies by categorizing them in the following three subsections:

2.5.1.1 Identifying stroke/ischemic stroke

Sedghi et al. (2015) converted medical narratives to codified text based on expert provided sign and symptom phrases and they applied ML algorithms on the codified sentences to predict the presence of stroke in a patient. Majersik Jennifer J et al. (2018) applied NLP-based approaches by adding context to n-grams that classified ischemic, hemorrhagic, and non-stroke cases with high precision by using different combination of clinical report types. Kim et al. (2019) utilized document-feature matrix vectorization techniques to classify brain MRI reports for identifying acute ischemic stroke. Govindarajan et al. (2020) developed ML-based NLP approaches to identify whether the stroke is ischemic or hemorrhagic based on some pre-defined symptoms and patient factors.

2.5.1.2 Classifying stroke subtypes

Two studies focused on automatically classifying stroke patients based on standard stroke subtype classification systems–the Trial of Org 10172 in Acute Stroke Treatment (TOAST) and the Oxfordshire Community Stroke Project (OCSP). Garg et al. (2019) developed ML-based approaches to classify patients according to the TOAST ischemic stroke subtyping using neurology progress notes and neuroradiology reports for better patient management and outcome prediction. Sung et al. (2020) constructed features based on the medical entities identified by MetaMap and then applied traditional ML techniques to classify stroke patients based on four clinical syndromes taken from OCSP classification system that considers the anatomical location of stroke.

2.5.1.3 Identifying stroke features

A recent study (Ong et al., 2020) classified radiology reports based on three outcomes - presence of stroke, involvement of MCA location, and stroke acuity by using text featurization methods such

as bag of words, term frequency-inverse document frequency, and GloVe. These are considered as three separate classification tasks and they employed traditional ML models and recurrent neural networks to predict the outcomes.

Most of the important information, especially those describing or relating to abnormal findings, are mentioned as part of the spatial descriptions between brain imaging observations and their corresponding anatomical structures. Often times, determining granular phenotypes is dependent on these specific information documented in the reports. Fu et al. (2019) developed both rule-based and ML methods to identify incidental silent brain infarct and white matter disease patients from the EHRs. As reported in Fu et al.'s work, some of the false positive errors generated by the ML-based text classification system are usually contributed by certain disease locations (e.g., right occipital lobe) that often co-exist with expressions related to the disease/outcome of interest (e.g., silent brain infarct in their case). Thus, developing a set of constraints using domain knowledge on the spatial information in the reports has the potential to diminish such false positive cases. Moreover, developing constraints based on the spatial relationships between imaging observations and anatomical locations forms a natural way to predict a stroke-associated outcome of interest. This also enhances the interpretability of the automatic phenotype construction system as it closely replicates a clinician's workflow to select eligible group of patients for treatment plans and clinical recommendations.

During the same time, Wheater et al. (2019) developed a rule-based NLP system to automatically label neuroimaging reports with a pre-defined set of 24 phenotypes. Their system incorporates manually crafted domain lexicons as well as a chunking step for extracting the radiological entities and relations from the text. Simple rules are then developed based on the presence of certain entities and relations to construct the final labels for each report.

We see that prior work has mostly focused on classifying relatively broad phenotypes (e.g., classifying if a report has evidence of a particular condition like *acute ischemic stroke*) and there is

still less research on using detailed spatial information from the reports for more fine-grained stroke phenotyping. Although Wheater et al. (Wheater et al., 2019) constructed 24 phenotypes, there is still a lot of reliance on the tedious process of developing manual rules for entity and relation extraction.

2.5.2 Automated Tracking

Most of the prior work using radiology report text has developed NLP systems to extract important entities such as findings, diagnoses, anatomical locations, and their respective descriptor terms (Hassanpour & Langlotz, 2016), with some focusing on more comprehensive information extraction (Steinkamp et al., 2019; Sugimoto et al., 2021). Some studies have targeted extracting information from the reports to automatically generate labels for the corresponding medical images (Syeda-Mahmood et al., 2020; Bradshaw et al., 2020; Wood et al., 2020).

In the context of automated tracking, existing research has highlighted the requirement of a tracking system to track radiological findings. Rubin et al. (2014) has extracted tumor-related quantitative assessments to facilitate automated tracking. More recently, Bozkurt et al. (2019) has focused on automatically identifying measurements and their corresponding descriptor terms from the reports with the aim to improve care delivery by tracking the same lesions across multiple patient encounters. Another study (Steinkamp et al., 2019) that extracted various important contextual information from radiology reports has also highlighted the benefits of automatic tracking. Two studies (Mabotuwana et al., 2018, 2019) have concentrated on automated matching of follow-up imaging recommendations from the reports using contextual information (e.g., recommended anatomy) and various other features (e.g., text-based similarity features). Interestingly, an earlier attempt (Son et al., 2004) was made where a probabilistic model was employed to correlate lung mass or lung lesion-related findings across different computed tomography documents associated with lung cancer patients. However, the study highlighted limitations such as requirement of more

refined definitions for locations (a highly weighted feature in the probabilistic model) in order to handle scenarios where multiple findings are detected around the same location.

One could consider the tracking problem as a cross-document (CD) coreference resolution task. In the general domain, there has been recent advancements to this relatively less explored and challenging task (Barhom et al., 2019; Cattan et al., 2021a,b; Bugert et al., 2021; Cattan et al., 2021c). Among the most recent contributions, Cattan et al. (2021a) developed the first end-to-end CD coreference resolution model where they applied the model over predicted mentions and achieved first baseline performances on the standard ECB+ dataset. Another work by Cattan et al. (2021b) proposed more realistic principles for evaluating CD coreference models (e.g., tackling lexical ambiguities involved in real-world CD coreferences). Cattan et al. (2021c) also proposed a hierarchical CD coreference resolution task where they identify the coreference clusters and hierarchy between them. In the medical domain, Wright-Bettner et al. (2019) provided insights on the challenging aspects of this task both from model and annotator perspectives through complex illustrative examples from a colon cancer dataset. They suggested relying more on schematic rules and less on annotator intuition to annotate more realistic and consistent CD coreference relations. Their work also highlighted the difficulties associated with creating human-annotated CD gold annotations on a sizable dataset and, thereby, restricted their annotation scope (e.g., limit the CD relations to a set of three notes per patient).

Besides the two works on CD coreference resolution on a medical corpus, (Son et al., 2004; Wright-Bettner et al., 2019), a few more studies have targeted within-document task (Apostolova et al., 2012; Miller et al., 2017). Thus, CD coreference is still under-explored in the clinical domain and we tackle this challenging problem specifically focusing on all radiological findings and devices.

2.5.3 Automated Image Labeling

Numerous work has extracted labels for the images using NLP on their corresponding reports (Wang et al., 2017, 2018; Peng et al., 2018; Daniels & Metaxas, 2019; Zech et al., 2018). Most of these studies generated coarse classification image labels for a set of diseases/findings utilizing rule-based NLP on the associated chest X-ray report text. The labels mainly include whether certain diseases are present or absent. More recently, two works have directed their focus on generating fine-grained image labels applying NLP on reports. Yan et al. (2019) extracted semantic labels to tag lesions in CT images using deep learning-based NLP module on the reports. The labels are more granular compared to the studies described above and include a lesion's body part, type, and attributes. Syeda-Mahmood et al. (2020) proposed an automatic labeling algorithm to generate fine-grained labels describing chest X-ray findings. The algorithm uses domain vocabulary to identify the core findings and a natural language parser to identify the finding modifiers from reports. These fine-grained labels are used to train a image classification model. Later, the labels predicted for a given image are used to automatically generate a report.

Although the latter two works focus on utilizing the rich information embedded in the reports with the aim to learn fine-grained descriptions of findings from images, one of the limitations mainly relates to the narrow scope of findings considered, i.e., CT lesions and chest X-ray findings. Another limitation is the lack of capturing other crucial information required for facilitating fine-grained diagnosis. Such crucial information includes fine-grained spatial information of the findings, for example, location specificity of a lesion as exemplified in–"*lesion in the posterior fossa centered posterior and to the left of the medulla, in the region of the cerebello-pontine cistern.*"

3

Spatial Representation Schema for Radiology Language

We propose two representation frameworks, one based on spatial role labeling (SpRL) (Kordjamshidi et al., 2010) and the other based on SpatialNet (Ulinski et al., 2019), to encode spatial language in radiology report text. We describe these frameworks in detail in the following sections.

3.1 RadSpRL - Radiology Spatial Role Labeling

In the general domain, earlier studies (Kordjamshidi et al., 2010, 2017) have formulated and evaluated the spatial role labeling task for extracting spatial information from text by mapping language to a formal spatial representation. In the SpRL annotation scheme, *an object of interest*

(Trajector) is associated with *a grounding location* (Landmark) through *a preposition or spatial trigger* (Spatial Indicator). For example, in the sentence, *"The book is on the table"*, the spatial preposition *'on'* indicates the existence of a spatial relationship between the object *'book'* (Trajector) and its location *'table'* (Landmark). In the medical domain, a limited number of studies have utilized the SpRL scheme. Kordjamshidi et al. (2015) extracted relations between bacteria names and their locations from scientific text. Roberts et al. (2015) utilized SpRL in the extraction of spatial relations between symptoms/disorders and anatomical structures from consumer-related texts.

We construct similar spatial roles for radiology texts based on SpRL. For instance, in a radiology report sentence, *"Mild streaky opacities are present in the left lung base"*, the location of a clinical finding *'opacities'* (Trajector) has been described with respect to the anatomy *'left lung base'* (Landmark) using the spatial preposition *'on'* (Spatial Indicator). Moreover, radiologists oftentimes document potential diagnoses related to the clinical findings which are spatially grounded. Consider the following example:

> *Stable peripheral right lower lobe opacities seen* ***between*** *the anterior 7th and 8th right ribs which may represent pleural reaction or small pulmonary nodules.*

Here, presence of a finding – *'stable peripheral right lower lobe opacities'* at a specific location – *'anterior 7th and 8th right ribs'* may elicit the radiologist to document two possible diagnoses – *'pleural reaction'* and *'small pulmonary nodules'*. As the actual occurrence of a disorder is highly dependent on various patient factors such as other physical examinations, laboratory tests, and symptoms, the radiologists usually describe diagnoses with uncertainty phrases or hedges. For instance, in the example above, the hedge term *'may represent'* is used to relate a finding and its corresponding body location with the most probable diagnoses.

In this dissertation, we propose the spatial role labeling-based framework as a preliminary step

to understand textual spatial semantics in chest X-ray reports. We define a basic spatial representation framework that extends SpRL for radiology (Rad-SpRL) involving interactions among common radiology entities. As most of the actionable clinical findings in all types of radiology reports are spatially located and represent a probable diagnosis, Rad-SpRL can potentially be extended to other report types. Consider the following sentence from a head CT report:

> *A well circumscribed hypodense 1 cm lesion is seen **in** the right cerebellar hemisphere consistent with prior stroke.*

Here, the spatial preposition *'in'* describes that the finding *'lesion'* is located inside the anatomical structure *'right cerebellar hemisphere'* which is also consistent with the diagnosis *'stroke'*.

3.1.1 Schema Description

Our spatial representation framework (Rad-SpRL) consists of 4 spatial roles (Trajector, Landmark, Hedge, and Diagnosis) with respect to a Spatial Indicator. The spatial roles and the Spatial Indicator are defined as follows:

1. Spatial Indicator: term (usually a preposition, e.g., *in*, *within*, *at*, *near*) that triggers a spatial relation
2. Trajector: object (finding, anatomical location) whose spatial position is being described
3. Landmark: location of the Trajector (may also be chained as a Trajector to another Landmark)
4. Hedge: phrase indicating uncertainty (e.g., *could be*, *may represent*), generally in reference to the Diagnosis and very rarely in the Trajector
5. Diagnosis: disease/clinical condition the radiologist associated with the finding

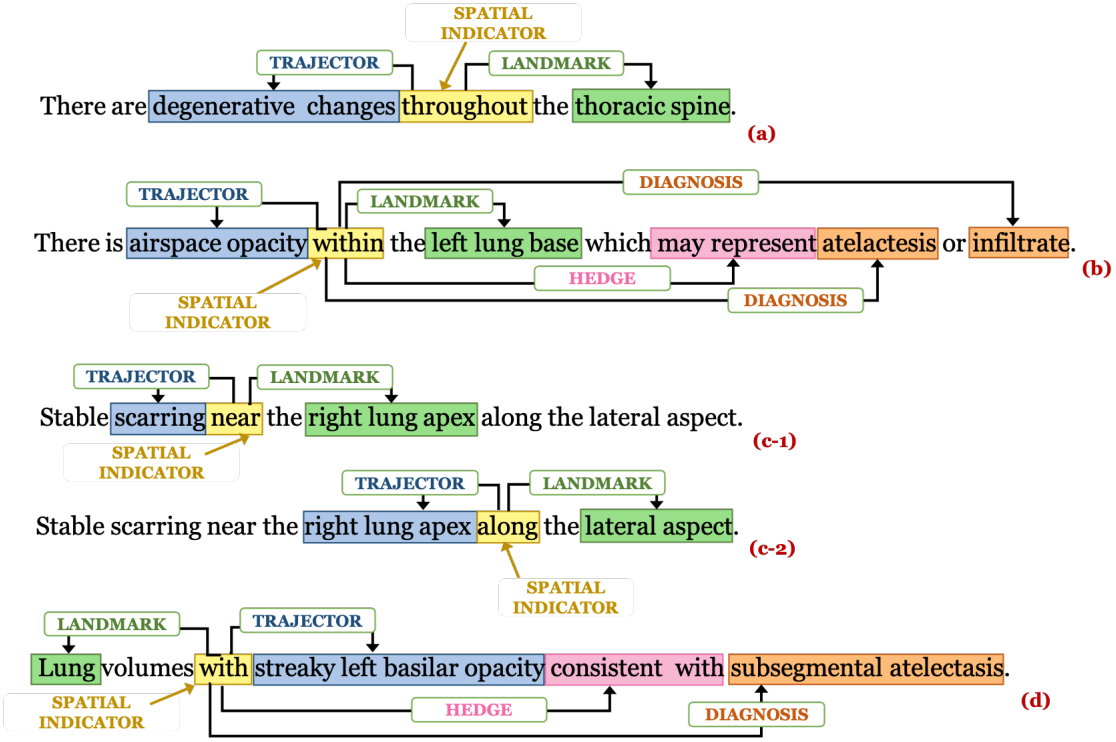

Figure 3.1: Examples of spatial role annotations: (a) Sentence having TRAJECTOR and LANDMARK, (b) Sentence having TRAJECTOR, LANDMARK, HEDGE, and DIAGNOSIS, (c-1) and (c-2) show the annotations of the same sentence containing 2 SPATIAL INDICATORS where the same entity *right lung apex* acts as a LANDMARK in (c-1) and a TRAJECTOR in (c-2), and (d) Sentence where a LANDMARK is described with a TRAJECTOR.

In most of the cases where a sentence contains spatial information, a finding (TRAJECTOR) is usually detected at a particular body location (LANDMARK) where the TRAJECTOR term appears to the left of the SPATIAL INDICATOR and the LANDMARK to its right. However, there are instances where a spatial preposition describes the body location (LANDMARK) with its associated abnormality (TRAJECTOR) and the TRAJECTOR term appears to the right of the SPATIAL INDICATOR and LANDMARK to the left (refer to example in Figure 3.1(d)). We have presented a few specific examples

to highlight how various spatial roles and Spatial Indicators are identified in sentences following the above definitions of Rad-SpRL in Figure 3.1. Please note that we have considered disease/condition terms as Diagnosis only when they are documented in conjunction with any spatially-located finding, or in other words are entirely probable diagnoses inferred from the finding. Also note that there is some ambiguity between a finding and a diagnosis, such that the same phrase may appear as a Diagnosis in one relation while being a Trajector in another. Our purpose here is not to formally distinguish between a finding and a diagnosis, but rather to identify the spatial relationships in radiology reports where the Trajector is generally a finding (or artifact in the image) and the Diagnosis is generally a well-understood disease term.

3.1.2 Dataset Annotation

A subset of 2000 reports from a total of 2470 non-normal reports as judged by two human annotators in Demner-Fushman et al. (2016) was used to create our spatial relation corpus. This newly annotated chest X-ray corpus contains spatial relations between findings and body locations as well as the correlated probable diagnoses and the hedging terms used in qualifying the diagnoses. We have presented a simple comparison between the Open-i manual annotations and our spatial annotations of a sample report in Figure 3.2. Note that we have not annotated other findings appearing in the report such as *Opacity* and *Pulmonary Fibrosis* as their corresponding body locations are not described through any spatial preposition.

3.1.2.1 Annotation Process

Two annotators annotated the spatial roles for each identified Spatial Indicator in each of the 2000 reports independently. They also were the annotators that manually coded the findings/diagnoses available as part of the Open-i dataset (Demner-Fushman et al., 2016). The spatial relation annotations were conducted in two rounds and reconciled after each. The first round consisted of annotating

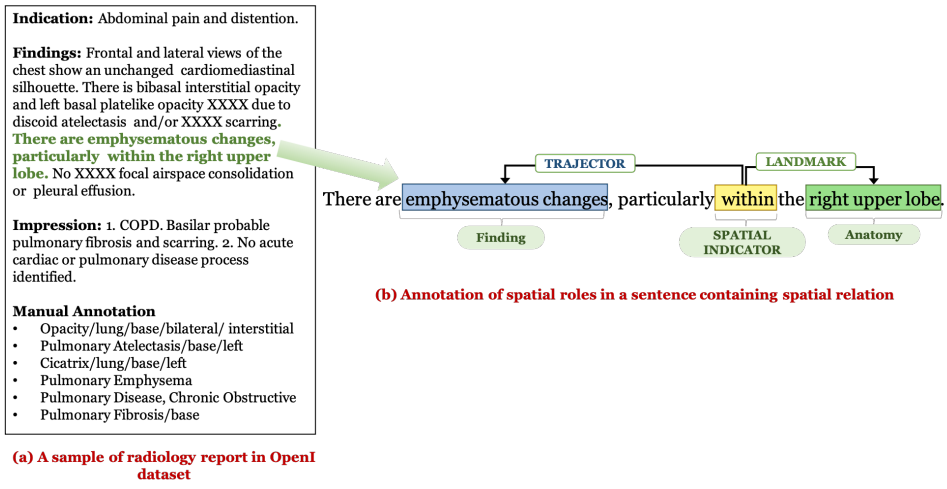

Figure 3.2: Examples of manual annotations: (a) Open-i annotations, (b) Our spatial relation annotations.

the first 500 reports and the second round consisted of annotating the remaining 1500. Figure 3.3 shows a sample annotated report from the corpus.

3.1.2.2 Annotation agreement

The inter-annotator agreement statistics for both Spatial Indicator and spatial roles are shown in Table 3.1. The Kappa (κ) agreement between the two annotators has been calculated for Spatial Indicator (as this is a binary classification task) whereas we report the overall F1 agreement for annotating the spatial role labels (as this is a role identification task). The Kappa agreement is high for Spatial Indicators in both annotation rounds. The F1 agreements for the 4 spatial roles are fairly low in the first round with much improvement in the second round. This is mainly because it is relatively easy and unambiguous to locate a spatial preposition in a sentence compared to identifying the spatial roles. All conflicts were reconciled with an NLP expert following each round of annotation. The moderate agreement rate for Trajector and Diagnosis

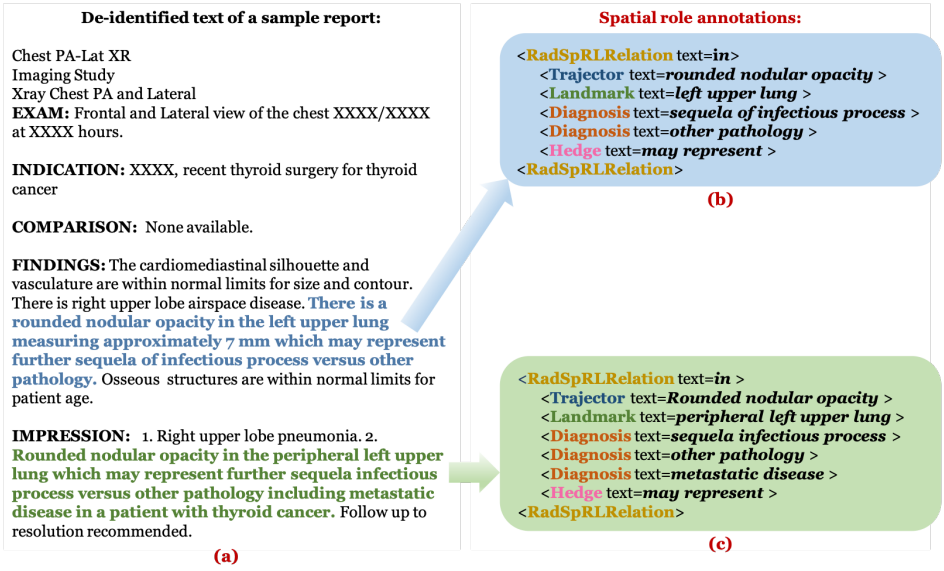

Figure 3.3: (a) Example of a de-identified report in our corpus, (b) Spatial role label annotations for the sentence represented by blue text in (a), and (c) Spatial role label annotations for the sentence represented by green text in (a). RadSpRLRelation indicates the text of the respective SPATIAL INDICATORS implying the existence of a spatial relation in both the sentences.

roles was likely due to ambiguity in distinguishing the two roles in a sentence, especially when the language pattern is different from the usual. Consider the examples below:

1. *Probably scarring **in** the left apex, although difficult to exclude a cavitary lesion.*
2. *There are irregular opacities **in** the left lung apex, that could represent a cavitary lesion **in** the left lung apex.*

In the first example, *'scarring'* was annotated as a TRAJECTOR after reconciliation as its spatial location (*'left apex'*) is described directly, although there is a higher chance of annotating it as a DIAGNOSIS since most of the probable diagnoses terms are usually preceded by a HEDGE term (*'Probably'* in this case). Similarly, *'cavitary lesion'* is indirectly connected to the same body location

Table 3.1: Annotator agreement.

Number of Reports	Kappa (κ)	Overall F1			
	Spatial Indicator	Trajector	Landmark	Diagnosis	Hedge
First 500	0.88	0.44	0.50	0.25	0.49
Remaining 1500	0.93	0.66	0.71	0.62	0.57
Complete 2000	0.92	0.59	0.64	0.49	0.55

Table 3.2: Descriptive Statistics of the annotations.

Parameter	Frequency
Average length of sentence containing spatial relation	13
Spatial Indicator	1962
Trajector	2293
Landmark	2167
Diagnosis	455
Hedge	388
Sentences containing at least 1 Spatial Indicator	1742
Maximum number of Spatial Indicator in any sentence	4
Spatial relations containing only Trajector and Landmark	1589
Spatial relations containing only Trajector, Landmark, and Diagnosis	9
Spatial relations containing only Trajector, Landmark, and Hedge	70
Spatial relations containing all 4 spatial roles	304
Spatial relations containing more than 1 Diagnosis	118
Maximum Diagnosis terms associated with any spatial relation	4

(*'left apex'*) and has been interpreted as an additional finding. So, *'cavitary lesion'* was also annotated as a Trajector and not as a Diagnosis. In the second example, *'cavitary lesion'* was annotated as a Diagnosis in context to the first 'in' in the sentence, whereas the same term *'cavitary lesion'* was annotated as a Trajector when its role was identified in context to the second 'in'. As previously noted, this difference where the same term can be both a Trajector and Diagnosis in different sentences is a consequence of focusing on explicitly representing the spatial language as described as well as the natural ambiguity between a finding and diagnosis in radiology. As a result, some downstream processing or interpretation is still required, which we leave to future work.

3.1.2.3 Annotation statistics

A total of 1962 spatial relations are annotated in our corpus of 2000 reports. Most of the Trajector terms were findings. However, 176 out of 2293 terms annotated as Trajectors were anatomical locations (example shown in Figure 3.1(c-2)). 118 Spatial Indicators had more than one probable Diagnosis, out of which 98 were associated with 2 Diagnosis terms, 17 were associated with 3 Diagnosis terms, and 3 had 4 associated Diagnosis terms. There are 1052 reports containing at least one sentence triggering a spatial relation. In those reports, there are 1742 sentences each containing at least one Spatial Indicator (1522 sentences containing exactly one Spatial Indicator and remaining 220 containing more than one Spatial Indicator). We have highlighted some brief descriptive statistics of our corpus based on the reconciled version of the annotations in Table 3.2.

3.2 Rad-SpatialNet - Radiology SpatialNet

A flexible way of incorporating fine-grained linguistic representations with knowledge is through the use of frames for spatial relations (Petruck & Ellsworth, 2018). The Berkeley FrameNet project (Baker, 2014) contains 29 spatial frames with a total of 409 spatial relation lexical units. Notably, SpatialNet (Ulinski et al., 2019) extends the use of FrameNet-style frames with enabling the connection of these frames to background knowledge about how entities may interact in spatial relationships using resources such as FrameNet (Baker, 2014). We propose an extension of SpatialNet (Ulinski et al., 2019) for radiology, which we call Rad-SpatialNet. Rad-SpatialNet is composed of 8 broad spatial frame types (as instantiated by the relation types), 9 spatial frame elements, and 14 entity types.

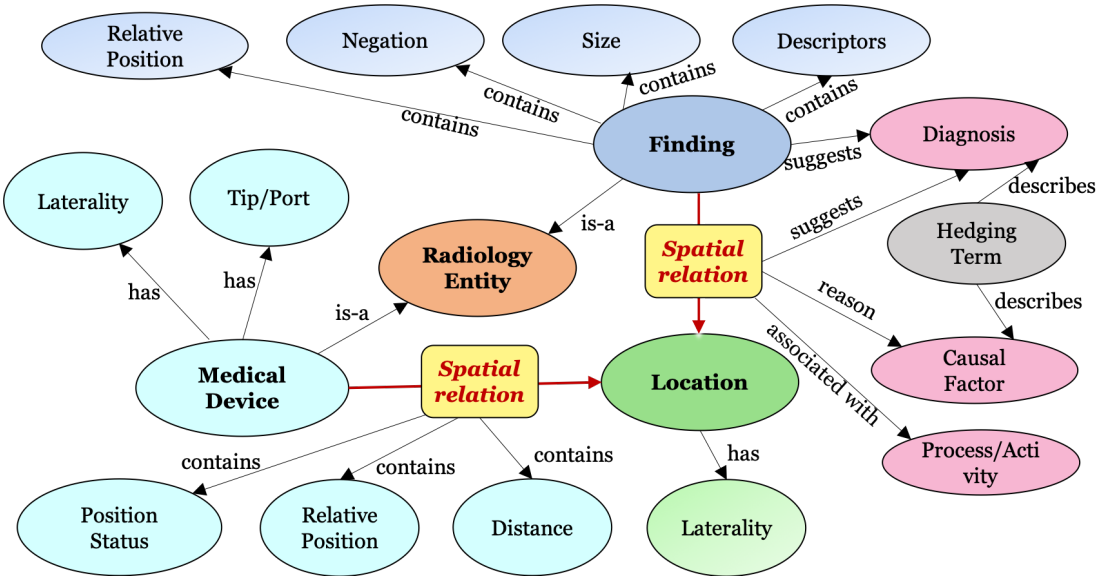

Figure 3.4: Relationship between entities in Rad-SpatialNet

3.2.1 Schema Description

Rad-SpatialNet provides a framework description to represent fine-grained spatial information in radiology reports by converting the linguistic expressions denoting any spatial relations to radiology-specific spatial meanings. For this, we extend the core design proposed in the general domain SpatialNet (Ulinski et al., 2019), which is based on FrameNet and VigNet, and tailor the framework specifically to encode spatial language in the domain of radiology. We have presented an overview of radiological spatial relations and the main participating entities in Figure 3.4. SpatialNet describes spatial frames by linking surface language to lexical semantics and further mapping these frames to represent the real spatial configurations. We update SpatialNet in our work with the aim to disambiguate the various spatial expressions used by radiologists in documenting their interpretations from radiographic images.

To build Rad-SpatialNet, we first utilize the language in radiology reports to construct a set of

spatial frames leveraging linguistic rules or valence patterns. The fundamental principle in forming the radiology spatial frames is the same used for constructing frames in FrameNet/SpatialNet. However, the main difference is that the target words (*lexical units*) of the frames are the spatial trigger words which are more common in radiology and are usually prepositions, verbs, and prepositional verbs. We then construct a list of ***spatial vignettes*** to transform these high-level spatial frames into more fine-grained frame versions by incorporating semantic, contextual or relation type constraints as well as radiology domain knowledge. The fine-grained frames reveal the true meaning of the spatial expressions from a radiology perspective. We describe these final frames containing the actual spatial configurations as the ***spatio-graphic primitives***. Unlike SpatialNet that uses the VigNet ontology to map the different lexical items into semantic categories, we utilize the publicly available radiology lexicon, RadLex, to map different radiological entities mentioned in the reports to standard terminologies recommended in radiology practice.

The main components involved in the proposed Rad-SpatialNet framework are described in the following sub-sections.

3.2.1.1 Ontology of Radiology Terms

We leverage RadLex (Langlotz, 2006) to map the various radiological entities along with other clinically important information in the report text to standard unified vocabularies to facilitate standardized reporting and decision support in radiology practice as well as research. RadLex consists of a set of standardized radiology terms with their corresponding codes in a hierarchical structure. Rad-SpatialNet utilizes the RadLex ontology to map all possible contextual information with reference to any spatial relation in the reports to the broader RadLex classes which capture the semantic types of the various information. For example, '*Endotracheal tube*' which is a Tube and a type of Implantable Device is mapped to the broad RadLex class Medical Device. Similarly, '*Ground-glass opacity*' belongs to the RadLex class Opacity which falls under the broad class

Imaging Observation. Thus, terms such as '*opacity*', '*opacification*', and '*Ground-glass opacity*' are mapped to Imaging Observation. We also link the various modifier or descriptor entities to standard RadLex descriptors. For instance, in '*rounded parenchymal opacity*', '*rounded*' is mapped to the Morphologic Descriptor class of RadLex, which is one of the categories of RadLex descriptors. Unlike VigNet, RadLex does not contain any graphical relations representing spatial configurations. So, we utilize Radlex mainly to map the terms in reports to radiology-specific semantic categories and not for creating spatio-graphic primitives. The mapped entities are utilized in the following steps to construct the spatial frames and subsequently the radiology-specific spatio-graphic primitives.

3.2.1.2 Spatial Frames

The spatial frames organize information in a radiology report sentence containing any spatial relation between common radiological entities (e.g., imaging observations and anatomical structures) according to the frame semantic principles, similar to FrameNet and SpatialNet. All the spatial frames created are inherited from the Spatial-Contact frame in FrameNet. We adopt similar valence patterns as defined in SpatialNet by specifying various lexical and syntactic constraints to automatically identify the frame elements from a sentence. However, there are differences in the set of frame elements in Rad-SpatialNet compared to SpatialNet. For example, besides Figure and Ground, some of the other common elements in the Rad-SpatialNet frames are Hedge, Diagnosis, Distance, and Relative Position.

To do this, we identify the most frequent words or phrases expressing spatial relations in radiology. Such a word/phrase also forms the lexical unit for a spatial frame. The type or the sense of the spatial trigger is also recognized to include the spatial relation type information in the frame. For example, in the sentence describing the exact position of a medical device - '*The umbilical venous catheter tip is now 1 cm* ***above*** *the right hemidiaphragm*', '*above*' is the spatial trigger having a

Table 3.3: Broader categories of spatial relations in radiology

Relation Type	Description
Containment	Denotes that a finding/observation/device is contained within an anatomical location ("*There is again seen high T2 signal **within** the mastoid air cells bilaterally*")
Directional	Denotes a directional sense in which a radiological entity is described wrt location ("*An NGT has its tip **below** the diaphragm*")
Contact	Denotes an entity is in contact with an anatomical structure ("*NGT **reaches** the stomach*")
Encirclement	Denotes a finding is surrounding an anatomical location or another finding ("*Left temporal hemorrhage with **surrounding** edema is redemonstrated*")
Spread	Denotes traversal of an entity toward an anatomical location ("*An NG tube **extends to** the level of the diaphragms.*")
Description	Denotes an anatomical location being described with any abnormality or observation ("*There is also some opacification **of** the mastoid air cells.*")
Distance	Denotes a qualitative distance between a radiographic finding and an anatomical location ("*There are areas of T2 hyperintensity **near** the lateral ventricles.*")
Adjacency	Denotes a radiographic finding is located adjacent to a location ("*There is a small amount of hypodensity **adjacent to** the body of the right lateral ventricle.*")

directional sense. This instantiates a spatial frame with **Directional** as the relation type and *above.prep* as the lexical unit. Some other common spatial relation types in Rad-SpatialNet are **Containment**, triggered by lexical units such as *in.prep*, *within.prep*, and *at.prep*; **Descriptive**, triggered by lexical units such as *shows.v*, *are.v* and *with.prep*; and **Spread** triggered by lexical units such as *extend (into).prep*, *throughout.prep*, and *involving.v*. The spatial relation types are shown in Table 3.3 and the elements identified for the spatial frames are described in Table 3.4.

The semantic type of the FIGURE and GROUND elements are identified for each spatial frame constructed using the RadLex ontology. For the example above related to the positioning of the

Table 3.4: Frame elements in Rad-SpatialNet

Element	Description
Elements with respect to a spatial trigger	
Figure	The object whose location is described through the spatial trigger (usually refers to finding/location/disorder/device/anatomy/tip/port)
Ground	The anatomical location of the trajector described (usually an anatomical structure)
Hedge	Uncertainty expressions used by radiologists (e.g., '*could be related to*', '*may concern for*' etc.)
Diagnosis	Clinical condition/disease associated with finding/observation suggested as differential diagnoses, usually appears after the hedge related terms
Reason	Clinical condition/disease that acts as the source of the finding/observation/disorder
Relative Position	Terms used for describing the orientation of a radiological entity wrt to an anatomical location (e.g., '*posteriorly*' in "*Blunting of the costophrenic sulci* ***posteriorly*** *is still present*", '*high*' in "*The UV line tip is* ***high*** *in the right atrium.*")
Distance	The actual distance of the finding or device from the anatomical location (e.g., '*1 cm*' in "*ETT tube is* ***1 cm*** *above the carina.*")
Position	Any position-related information, usually in context to a device (e.g., '*terminates*'
Status	in "*A right PIC catheter* ***terminates*** *in the mid SVC.*")
Associated Process	Any process/activity associated with a spatial relation (e.g., '*intubation*' in "*may be related to recent* ***intubation***")
Elements with respect to a radiological entity	
Status	Indicating status of entities (e.g., '*stable*', '*normal*', '*mild*')
Morphologic	Indicating shape (e.g., '*rounded*')
Density	Terms referring to densities of findings/observations (e.g., '*hypodense*', '*lucent*')
Modality	Indicating modality characteristics (e.g., '*attenuation*')
Distribution	Indicating distribution patterns (e.g., '*scattered*', '*diffuse*')
Temporal	Indicating any temporality (e.g., '*new*', '*chronic*')
Composition	Indicating composition of any radiological observation (e.g., '*calcified*')
Negation	The negated phrase related to a finding/observation (e.g., '*without evidence of*')
Size	The actual size of any finding/observation (e.g., '*14-mm*' describing the size of a lytic lesion)
Laterality	Indicating side (e.g., '*left*', '*bilateral*')
Quantity	Indicating the quantity of any radiological entity (e.g., '*multiple*', '*few*')

umbilical venous catheter tip, the semantic type of Figure is Medical Device and the type of Ground is Anatomical Location. All this information–combining the semantic types and the spatial relation type–are used to further refine the spatial frame **Figure-Directional-Ground-SF** as **Medical Device-Directional-Anatomical Location-SF**.

3.2.1.3 Spatial Vignettes

The main idea of spatial vignettes are also adopted from SpatialNet. Here, we develop vignettes primarily to resolve the ambiguities involved in using the same spatial trigger words or phrases to

describe different radiological contexts. In other words, the same spatial expressions might have different spatial configurations based on the context or the radiological entities associated.

The vignettes connect the spatial frames to spatio-graphic primitives utilizing the RadLex ontology, semantic/relation type constraints as well as domain knowledge to generate more accurate spatial representations of radiology language. Consider the following two sentences having the same spatial trigger '*extends into*':

1. *There is interval increase in the right pleural effusion which* ***extends into*** *the fissure.*

2. *There is an NG tube which* ***extends into*** *the stomach.*

The first sentence contains a radiologist's description of a fluid disorder '*pleural effusion*' with respect to an anatomical reference–fissure in the pleural cavity (a closed space around the lungs), whereas the second sentence describes the positioning of a feeding tube. It is difficult to interpret the actual spatial meaning of the same prepositional verb '*extends into*' in these two different contexts solely from the lexical information. The spatial vignettes map the spatial frames corresponding to these sentences to different spatio-graphic primitives representing the actual spatial orientation of '*extends into*' utilizing radiology domain knowledge.

Semantic constraints are applied to the Figure and Ground frame elements, whereas a spatial relation type constraint is also added to the relation sense. If the semantic category of the Figure is Medical Device and the relation type is Directional, with the lexical unit *extends (into).prep*, then a spatial vignette will generate the spatio-graphic-primitive **Medical Device-Terminates Into-Anatomical Location-SGP** from the spatial frame **Medical Device-Directional-Anatomical Location-SF** (corresponding to example 2 above). Another vignette will produce the spatio-graphic primitive **Disorder-Extends Into-Anatomical Location-SGP** if the semantic category of the Figure is Disorder instead of Medical Device and particularly refers to fluid-related disorders like '*pleural effusion*'(corresponding to example 1

above). The vignettes here determine the spatial meanings of radiology sentences based on both the semantic types of FIGURE element and specific properties of the DISORDER type (for example, *fluidity* in this case). Similar ambiguities are observed for spatial expressions containing lexcial units such as *projects (over).prep* and *overlying.v* as they often occur in both device and observation or disorder-related contexts. Thus, the vignettes differentiate the real orientation of the spatial expression when it is used in context to a medical device versus any other radiographic finding. Consider another set of examples below where both sentences are related to the tip position of two medical devices:

1. *The umbilical arterial catheter tip **projects over** the T8-9 interspace.*

2. *The tip of the right IJ central venous line **projects over** the upper right atrium.*

Here, the spatial vignettes produce a more specific spatial representation of the prepositional verb '*projects over*' based on the details of anatomical location. A spatial vignette will produce the spatiographic primitive **MEDICAL DEVICE TIP-IN FRONT OF-ANATOMICAL LOCATION-SGP** for the first sentence. IN FRONT OF is derived as the anatomical structure is corresponding to the '*T8-9 interspace*' in the spine which is often used as the level of reference to indicate the position of catheters and tubes and these tubes and catheters are in front of the spine. Another vignette will produce **MEDICAL DEVICE TIP-TERMINATES AT LEVEL OF-ANATOMICAL LOCATION-SGP** for the second sentence as the IJ venous line lies within the internal jugular vein and might go upto the '*right atrium*'.

Consider the following sentences containing '*overlying.v*' as the lexical unit:

1. *There is a less than 1 cm diameter rounded nodular opacity **overlying** the 7th posterior rib level.*

2. *The left IJ pulmonary artery catheter's tip is currently **overlying** the proximal SVC.*

If the FIGURE is IMAGING OBSERVATION and the GROUND is particularly associated with anatomical locations such as ribs, a spatial vignette will produce the primitive **IMAGING OBSERVATION-PROJECTS OVER-ANATOMICAL LOCATION-SGP** for the first sentence. Ribs are also used the same way as spine to describe the level of objects or pathology. However, for the second sentence, the spatiographic primitive will be **MEDICAL DEVICE TIP-TERMINATES AT LEVEL OF-ANATOMICAL LOCATION-SGP** as the FIGURE is MEDICAL DEVICE and the GROUND (anatomical location) is '*SVC*'.

The following examples contain the spatial trigger '*of*' connecting two anatomical structures or parts of anatomical structures:

1. *There is increased signal identified within the pons extending to the right side **of** the midline.*
2. *The UA catheter tip overlies the left pedicle **of** the T9 vertebral body.*

In the above cases with respect to the spatial preposition '*of*', the semantic types of both FIGURE and GROUND are referring to ANATOMICAL LOCATIONS. However, there is a difference in the interpretation of the same trigger word '*of*'. The vignette adds a constraint that if the spatial relation type is DESCRIPTIVE and both FIGURE and GROUND have semantic type as ANATOMICAL LOCATION, then the meaning of the preposition is determined based on the words of the FIGURE element. If the words are either '*side*' or '*aspect*', then '*of*' refers to a subarea/side of the GROUND anatomical location. Whereas for other words, '*of*' refers to a specific identifiable anatomical structure contained within the GROUND element, similar to '*pedicle*' present at each vertebra in Example 2 above. The spatial vignettes corresponding to the three spatial expressions described above - *extends (into).prep*, *projects (over).prep*, and *of.prep* are illustrated in Figure 3.5.

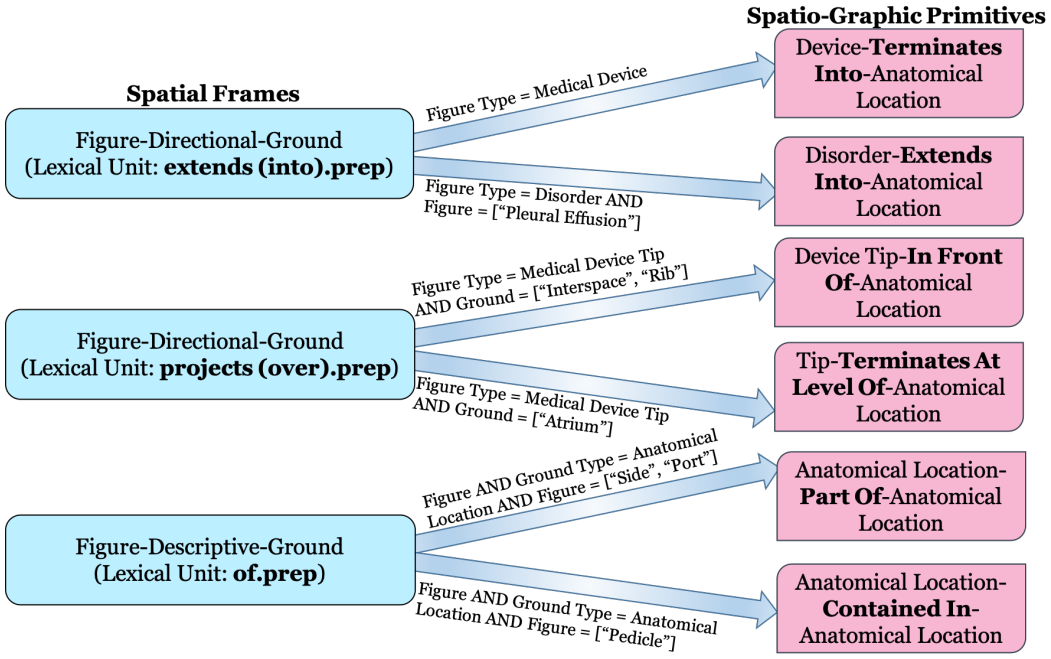

Figure 3.5: Examples of spatial vignettes for differentiating the spatial meanings of three commonly found spatial expressions in radiology reports.

3.2.2 Dataset Annotation

We annotated a total of 400 radiology reports–Chest X-ray reports (136), Brain MRI reports (127), and Babygram reports (137)–from the MIMIC III clinical database (Johnson et al., 2016). The language used in MIMIC reports is more complex and the report lengths as well as sentence lengths are long compared to other available datasets such as open-i chest X-ray reports (Demner-Fushman et al., 2016). We filtered the babygram-related reports following the '*babygram*' definition, that is, an X-ray of the whole body of an infant (usually newborn and premature infants). Since babygram reports have frequent mentions of medical device positions and involve multiple body organs, we incorporate this modality mainly with the intention to build a corpus with balance

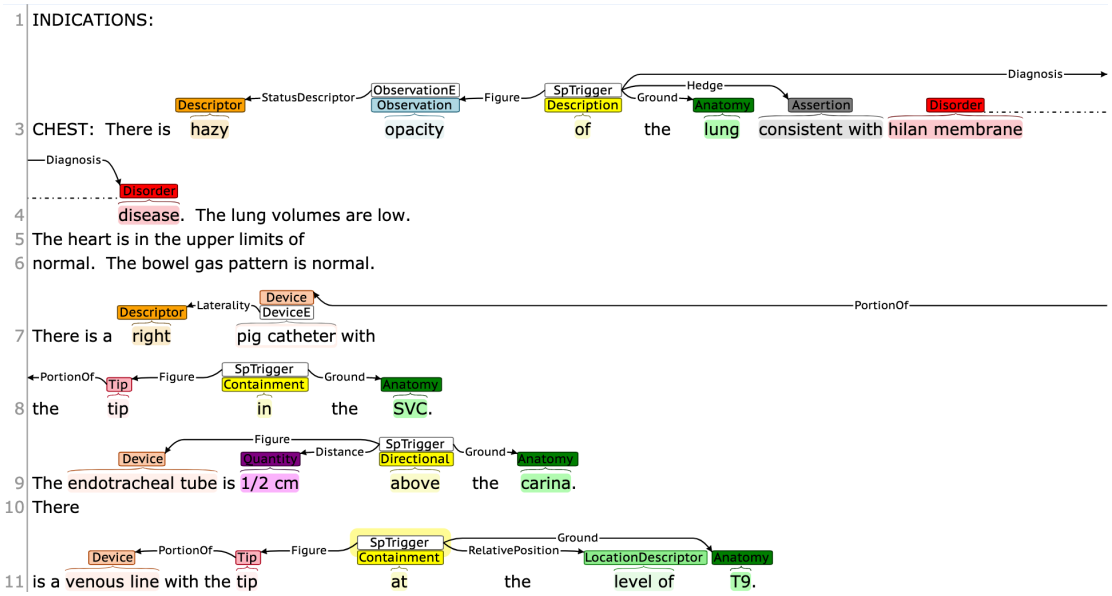

Figure 3.6: Examples of annotations.

between two major spatially-grounded radiological entities–imaging observations/clinical findings and medical devices.

3.2.2.1 Annotation Process

We pre-processed the reports to de-identify some identifiable attributes including dates and names and also removed clinically less important contents. All the sentences in the reports containing potential spatial relations were annotated by two human annotators. The annotation was conducted using Brat. The annotations were reconciled three times–following the completion of 50, 200, and 400 reports. Since identifying the spatial frame elements involves interpreting the spatial language from the contextual information in a sentence, the annotations of the first 50 reports (i.e. our calibration phase) differed highly between the annotators. Some of the major disagreements related to spatial relation sense were discussed and the annotation guidelines were updated after first two rounds of annotation. Examples of sample annotations are provided in Figure 3.6.

To provide more insights into some of the complexities encountered in the annotation process, we highlight a few cases here. **First**, oftentimes two anatomical locations are present in a sentence and in such scenarios, an intermediate anatomical location (first location occurrence) is chained as a FIGURE element in context to the broader anatomical location (second location occurrence and the GROUND element). However, this chaining is not valid if the two locations are not connected. Consider the sentences below:

1. *There are few small air fluid levels in* ***[mastoid air cells]***$_{FirstLocation}$ *within the left* ***[mastoid process]***$_{SecondLocation}$.
2. *Area of increased signal adjacent to the left lateral* ***[ventricle]***$_{FirstLocation}$ *at the level of* ***[corona radiata]***$_{SecondLocation}$.

For the first sentence, '*mastoid air cells*' is contained within '*mastoid process*' and these anatomical locations are connected. Therefore, '*mastoid air cells*' is annotated as FIGURE element and '*mastoid process*' as the GROUND in context to the spatial frame formed by the lexical unit '*within.prep*'. Note that '*mastoid air cells*' is the GROUND element associated with the FIGURE '*air fluid levels*' through the spatial trigger '*in*'. However, for the second sentence, '*ventricle*' and '*corona radiata*' are two separate anatomical references and are not connected. Hence, two separate spatial relations are formed with the radiographic observation '*signal*', one between '*signal*' and '*ventricle*' described through '*adjacent to*' and the other between '*signal*' and '*corona radiata*' described through '*at*' and the RELATIVE POSITION '*level of*'. **Second**, some instances require correct interpretation of whether prepositions such as '*of*' are SPATIAL TRIGGERS or are part of location descriptors. Note the examples below:

1. *PICC line with its tip located at the* ***junction of*** *superior vena cava and left brachiocephalic vein.*

2. *Abnormal signal in posterior portion **of** spinal cord.*

In the first sentence, '*junction of*' is annotated as the RELATIVE POSITION describing the connection between '*superior vena cava*' and '*brachiocephalic vein*', whereas in the second sentence, '*of*' is a SPATIAL TRIGGER connecting the FIGURE–'*portion*' and the GROUND–'*spinal cord*'. **Third**, although location descriptor words such as *anterior*, *lateral*, and *superior* are usually annotated as RELATIVE POSITION, in a few cases they are annotated as SPATIAL TRIGGER. For example, note the following two sentences:

1. *There is an area of high signal intensity extending into the **[anterior]**$_{\text{RELATIVE POSITION}}$ mediastinum.*
2. *There is a 6 mm lymph node **[anterior to]**$_{\text{SPATIAL TRIGGER}}$ the carina.*

In the first sentence, '*anterior*' is used to describe '*mediastinum*' and is annotated as a RELATIVE POSITION, whereas in the second sentence, '*anterior*' contributes in perceiving the actual spatial sense and hence '*anterior to*' is annotated as a SPATIAL TRIGGER.

3.2.2.2 ANNOTATION STATISTICS

The inter-annotator agreement results are shown in Table 3.5. We calculate the overall F1 agreement for annotating the spatial relation types, the main entities, and the spatial frame elements. The agreement measures are particularly low (around 0.4) for FINDING/OBSERVATIONS as often there are higher chances of boundary mismatch in the process of separating the descriptor-related words from the main finding or observation term. There are very few instances of PROCESS entities and cases where the disorder terms act as REASONS in the corpus which have resulted in the agreement measures being zero. Ultimately, Rad-SpatialNet requires an extremely knowledge-intensive annotation process, so low agreement at this stage is not unreasonable. Future work will include additiional quality checks to ensure the semantic correctness of the annotations.

Table 3.5: Annotator agreement.

Item	First 50	Next 150	Last 200
Relation Types			
Containment	0.73	0.81	0.81
Directional	0.29	0.73	0.46
Contact	0.44	0.1	0.03
Encirclement	0.67	0.57	0
Spread	0.24	0.55	0.45
Description	0.42	0.54	0.58
Distance	0	0.5	0.4
Adjacency	0	0.25	0.44
Main entities			
Spatial Trigger	0.58	0.81	0.78
Anatomy	0.48	0.68	0.76
Device	0.35	0.82	0.83
Tip	0.38	0.98	0.97
Finding/Observation	0.28	0.43	0.38
Descriptors	0.29	0.64	0.71
Frame Elements			
Figure	0.33	0.58	0.62
Ground	0.42	0.67	0.70
Diagnosis	0	0.51	0.54
Hedge	0.19	0.48	0.45
Reason	0	0.38	0
Relative Position	0.07	0.48	0.58
Distance	0.4	0.86	0.71
Position Status	0	0.62	0.42
Associated Process	0.57	0	0

3.2.2.3 Corpus statistics

358 (89.5%) of the reports contain spatial relations. The reconciled annotations contain a total of 1101 sentences with mentions of spatial triggers. There are 1372 spatial trigger terms in

total (average of 3.8 triggers per report). The frequencies of the entity types, the types of spatial relations, and the spatial frames are presented in Table 3.6. The predominant spatial trigger types to instantiate spatial frames are 'Containment', 'Description', and 'Directional'. 81 of the 330 DEVICE entities are described using various descriptor terms, 327 of the 436 OBSERVATION entities, 240 of the 367 CLINICAL FINDINGS, 190 of the 390 DISORDERS, and 564 of the 1492 ANATOMICAL LOCATIONS contain descriptors. The distribution of various types of descriptors (consistent with RadLex) across the main radiological entities are shown in Table 3.7. Note that the figures in Table 3.6 and Table 3.7 are considering only those entities and their descriptors which are involved in a spatial relation. There are 1328 frame instances which correspond to the main radiological entities as demonstrated in Table 3.6. Among the remaining 44 spatial triggers, 11 are related to Port/Lead of devices and 33 describes how a specific part of an anatomical structure is linked to the main part.

3.3 LIMITATIONS OF RAD-SPRL AND RAD-SPATIALNET

Rad-SpRL only covers prepositional spatial expressions as SPATIAL INDICATORS, and do not consider the non-prepositional ones (e.g., verbs such as *'demonstrates'*, *'shows'* etc.). Moreover, multi-word spatial expressions are not covered (e.g., *'projects in'*, *'projecting through'*, *'projected over'*), although such expressions occur rarely in the Open-i chest X-ray reports dataset to describe the location of findings. Additionally, the Rad-SpRL framework does not capture other important and common spatially-grounded radiology entities such as medical devices in the reports. The following example illustrates a sample sentence where the medical device *'Right IJ venous catheter'* acts as the TRAJECTOR in reference to its associated location *'proximal SVC'* that acts as the LANDMARK:

*Right IJ venous catheter terminates **at** the proximal SVC.*

Rad-SpRL also does not capture other important contextual information related to spatial relations such as the relative position of a finding w.r.t. its location (e.g., "*posteriorly*" in the sentence–'*Blunting*

Table 3.6: General corpus statistics.

Item	Freq
General	
Average sentence length	17.6
Spatial triggers	1372
Sentences with 1 spatial trigger	874
Sentences with more than 1 spatial trigger	227
Entity Types	
Medical Device	330
Tip of Device	142
Device Port/Lead	11
Imaging Observation	436
Clinical Finding	367
Disorder	390
Anatomical Location	1492
Descriptor	1548
Assertion	326
Quantity	67
Location Descriptor	398
Position Info	167
Process	19
Spatial Frames (SFs)	
Containment	642
Description	387
Directional	168
Spread	69
Contact	54
Adjacency	32
Encirclement	14
Distance	6
Most frequent Lexical Units - Containment SF	
'in.prep'	410
'within.prep'	134
'at.prep'	86
Most frequent Lexical Units - Description SF	
'of.prep'	277
'are.prep'	37
'with.prep'	17
Most frequent Lexical Units - Directional SF	
'above.prep'	43
'projecting (over).prep'	24
'below.prep'	18
Spatial Frames based on semantic types	
Medical Device-related	194
Medical Device Tip-related	142
Imaging Observation-related	436
Clinical Finding-related	344
Disorder-related	212
Spatial frame elements	
Figure	1491
Ground	1537
Hedge	249
Diagnosis	190
Reason	33
Relative Position	398
Distance	45
Position Status	167
Associated Process	21

Table 3.7: Descriptive statistics of the radiological entities in the annotated corpus. (Obs - Observation, Fndg - Finding, Dis - Disorder, Devc - Device, Aty - Anatomy)

Element	Obs	Fndg	Dis	Devc	Aty
Status	231	121	77	1	22
Quantity	50	31	14	5	30
Distribution	43	14	6	0	2
Morphologic	37	20	6	0	6
Size Desc.	33	19	31	1	9
Negation	32	50	20	0	1
Temporal	23	26	50	14	0
Laterality	20	3	22	68	499
Size	19	1	3	0	0
Composition	5	8	2	0	2
Density	5	0	0	0	0
Modality	2	0	0	0	0

of the costophrenic sulci posteriorly is still present.') and shape of a finding (e.g., "*rounded*" in the sentence–'*Again seen is a somewhat rounded parenchymal opacity in the right mid-lung.*'). Also, both positive and negative spatial relations are considered in Rad-SpRL as the primary focus was on identifying the spatial relationship itself, not the presence or absence of the condition to which the relation refers.

The Rad-SpatialNet framework is an advancement over the Rad-SpRL framework and thus captures more detailed spatial information from the reports than the Rad-SpRL. However, some other complex information in context to a spatial trigger could be considered for later work. Consider the sentence - '*The tip of the catheter has a* ***mild rightward curve****, suggesting that it may be directed into a portal vein.*' Here, information about '*intermediate position change*' might also be annotated as a frame element. Further, some phrases such as '*needs repositioning*' can be differentiated from Position Status as Position Recommendation.

In both Rad-SpRL and Rad-SpatialNet, only the intra-sentence spatial relations are covered. Oftentimes, we encounter inter or cross-sentence relations and scenarios where the differential diagnoses are documented in the sentence following the spatial relation or even far apart in the 'Impression' section (around 12.75% of the reports in the Rad-SpatialNet corpus). Covering inter or cross-sentence relations is left for future work. Moreover, the same entities that are referred to multiple times in the same report (e.g., once in the 'Findings' section and again in the 'Impression' section) are not linked in both the Rad-SpRL and Rad-SpatialNet datasets.

4

Deep Learning-based Natural Language Processing Methods for Spatial Information Extraction

This chapter describes our proposed deep learning-based methods for spatial information extraction from the radiology reports. The methods are based on sequence labeling and question answering. We additionally present the results obtained by applying these methods on both the Rad-SpRL and the Rad-SpatialNet datasets.

4.1 Sequence Labeling

We apply a set of deep learning models, primarily the pre-trained transformer language models like BERT (Bidirectional Encoder Representations from Transformers), to extract spatial information from report text. For the Rad-SpRL dataset, we apply bidirectional long short-term memory (Bi-LSTM) conditional random field (CRF) neural network as the baseline model and additionally utilize two pre-trained transformer language models (BERT and XLNet) for extracting the Spatial Indicators in a sentence and consequently to extract the associated spatial roles for each Spatial Indicator. Similarly, we apply BERT-based model on the Rad-SpatialNet dataset to first extract the spatial triggers in a sentence and then the spatial frame elements associated with each trigger. For both spatial role and spatial frame element extraction, we evaluate using both the gold and the predicted Spatial Indicators or spatial triggers in a sentence.

4.1.1 Description for Rad-SpRL

4.1.1.1 Baseline model

We formulate the spatial role extraction as a sequence labeling task. We utilize a Bi-LSTM CRF framework similar to the proposed architecture in Lample et al. (2016) both for Spatial Indicator extraction and spatial role labeling. The CRF in the decoding layer takes into account the sequential information in the sentence while predicting the sequence labels related to any spatial role (Trajector, Landmark, Diagnosis, and Hedge). We utilize a Bi-LSTM that incorporates a character embedding x_i^{ce} (where each character is denoted $c_{i,j}$) for each word w_i in a sentence. Here, i represents the word position and j stands for the position of the character in the word w_i. For every word, this character embedding is then concatenated with the respective pre-trained word embedding x_i^{we}. For extracting the spatial role labels, additionally a Spatial Indicator embedding x_i^{ind} is concatenated to the word and character embeddings to distinguish the indicators from non-

indicator words. The final concatenated representation $[x_i^{we}; x_i^{ce}; x_i^{ind}]$ is fed into the final Bi-LSTM network with one hidden layer. The overall architecture is presented in Figure 4.1.

4.1.1.2 BERT and XLNet-based models

First, we fine-tune BERT for extracting the Spatial Indicators in a sentence and second, we apply the fine-tuned model for labeling the four spatial roles provided the Spatial Indicator in a sentence. In this work, we represent a sentence obtained after WordPiece tokenization as [[CLS] sentence [SEP]] for constructing a single input sequence following the original BERT paper (Devlin et al., 2019), where [CLS] is a symbol added at the beginning of each input sentence and [SEP] is a separator token for separating sentences. The input sequences are then fed into the BERT model to generate contextual representations. For spatial role labeling, we mask the Spatial Indicator term with an identifier '$spin$' to better encode the positional information of the specific Spatial Indicator in a sentence for which the spatial roles are annotated. The contextual BERT representation corresponding to each word in the sequence [[CLS] sentence [SEP]] is then concatenated with a Spatial Indicator embedding similar to the baseline Bi-LSTM CRF model. The concatenated representation is fed into a simple linear classification layer for predicting the final labels for each token. The model architecture is illustrated in Figure 4.2.

To fine-tune BERT for spatial role labeling for the Rad-SpRL corpus, we initialize the model with the publicly available pre-trained checkpoints of the BERT large model ($\text{BERT}_{\text{LARGE}}$). We also initialize the model parameters obtained by pre-training BERT on medical corpus (MIMIC-III clinical notes). We have adopted these pre-trained parameters from a previous work (Si et al., 2019) where clinical domain embedding models were pre-trained on MIMIC-III clinical notes, referred to as $\text{BERT}_{\text{LARGE}}$ (MIMIC), after initiating from the $\text{BERT}_{\text{LARGE}}$ released checkpoint. Owing to the best performance of $\text{BERT}_{\text{LARGE}}$ (MIMIC) on clinical concept extraction for four benchmark datasets (Si et al., 2019), we initiate our model with the pre-trained parameters of $\text{BERT}_{\text{LARGE}}$

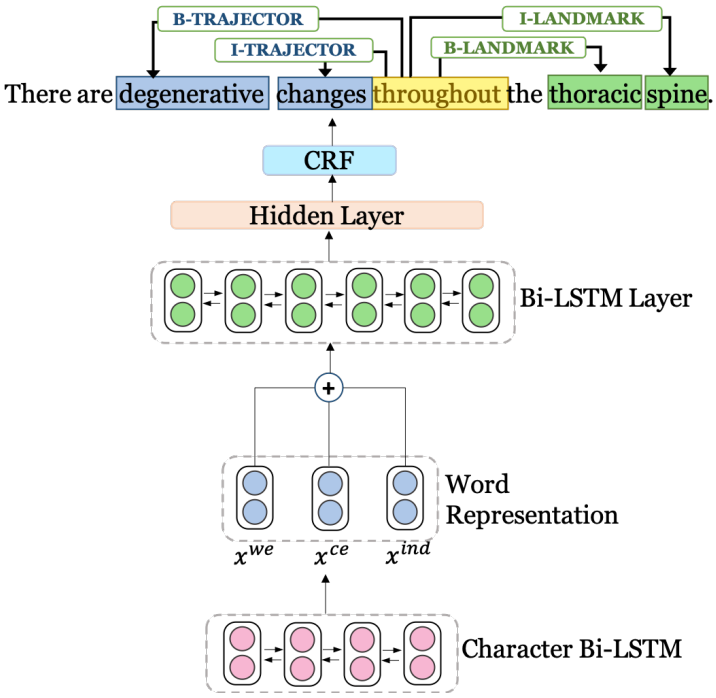

Figure 4.1: Baseline model architecture. For each word, a character representation is fed into the input layer of the Bi-LSTM network. For each word, x^{we} represents pre-trained word embeddings, x^{ce} represents character embeddings, and x^{ind} represents indicator embeddings. The final predictions for the spatial role labels in a sentence are made combining the Bi-LSTM's final score and CRF score.

(MIMIC) to fine-tune on our spatial role labeling task.

For XLNet, the model input is similar to BERT and we feed [sentence [SEP] [CLS]] into the model. We have utilized a similar simple architecture as BERT for fine-tuning XLNet on Rad-SpRL. However, we have initialized the model with the released pre-trained model parameters ($\text{XLNet}_{\text{LARGE}}$) for fine-tuning as experimenting with the MIMIC pre-trained parameters has yet to result in further performance improvement.

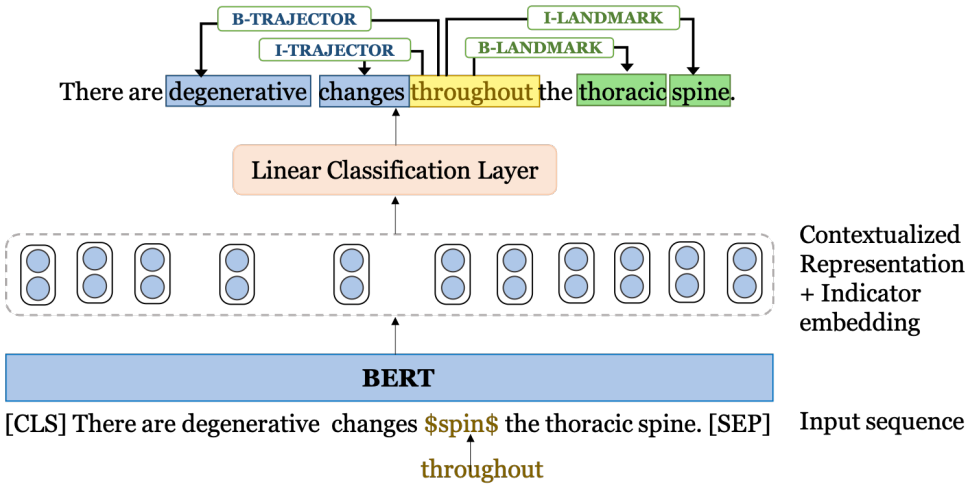

Figure 4.2: BERT-based model.

4.1.1.3 Pre-processing

Spatial Indicator extraction We preprocess the Rad-SpRL dataset to generate input sequence for the models. We follow Beginning (B), Inside (I), and Outside (O) tagging scheme to label the words in a sentence. The input to the models consists of the sequence of words and the corresponding BIO tags. The following example shows how a sentence containing two Spatial Indicators is tagged.

[*Stable*]$_{\text{O}}$ [*scarring*]$_{\text{O}}$ [***near***]$_{\text{B-Indicator}}$ [*the*]$_{\text{O}}$ [*right*]$_{\text{O}}$ [*lung*]$_{\text{O}}$ [*apex*]$_{\text{O}}$ [***along***]$_{\text{B-Indicator}}$

[*the*]$_{\text{O}}$ [*lateral*]$_{\text{O}}$ [*aspect*]$_{\text{O}}$

Spatial role labeling For each Spatial Indicator in a sentence, we create an instance or sample of the sentence. For each instance, we tag all the spatial roles (Trajector/Landmark/Diagnosis/Hedge) as well as the Spatial Indicator. Creating separate sentence instance for each Spatial Indicator helps in dealing with cases where the same word can be both a Trajector and a Landmark in context of two different Spatial Indicators in the sentence (example shown in (c-1) and

(c-2) in Figure 3.1 in chapter 3). Also, annotating only the roles associated with a single Spatial Indicator provides the model unambiguous information about the position of the specific indicator term to which these roles are associated. We again follow the BIO tagging scheme. The input to the final model consists of words and the corresponding B, I, O labels for a set of sentences. However, in the case of applying BERT and XLNet, the input sentence is tokenized by WordPiece and SentencePiece tokenizers before feeding into the BERT and XLNet encoders, respectively. The following example shows the tagged words for the sentence – *"Minimal degenerative changes **of** the thoracic spine"*.

[*Minimal*]$_{\text{B-Trajector}}$ [*degenerative*]$_{\text{I-Trajector}}$ [*changes*]$_{\text{I-Trajector}}$

[***of***]$_{\text{Indicator}}$ [*the*]$_{\text{O}}$ [*thoracic*]$_{\text{B-Landmark}}$ [*spine*]$_{\text{I-Landmark}}$

4.1.1.4 Experimental settings and evaluation

We use pre-trained medical domain MIMIC-III word embeddings of 300 dimensions * in our Bi-LSTM experiments. The character and the indicator embeddings are initialized randomly and altered during training. The dimensions of character and indicator embeddings are 100 and 5 respectively. The model is implemented using TensorFlow (Abadi et al., 2016), and the hyperparameters are chosen based on the validation set. LSTM hidden size is set at 500, dropout rate at 0.5, learning rate at 0.01, and learning rate decay at 0.99. We use the Adam optimizer and train the model for a maximum of 20 epochs.

For fine-tuning BERT, both for BERT$_{\text{LARGE}}$ and BERT$_{\text{LARGE}}$ (MIMIC), we largely followed the standard BERT parameters, including setting the maximum sequence length at 128, learning rate at 2e-5, and using the cased version of the models. Additionally, we set the number of training epochs at 4 based on the performance of the models on the validation set. For BERT$_{\text{LARGE}}$ (MIMIC), we initialize the model parameters pre-trained on MIMIC after 320000 steps. For XLNet, the

*https://northwestern.app.box.com/s/eprxyxmee37p3d6khqbpn125tyttq4u6

maximum sequence length and learning rates are the same as used for BERT, casing is also preserved, and the number of training steps is set at 2500 based on the validation set performance. In both BERT and XLNet, the dimension of indicator embedding is set at 5.

First, we perform 10-fold cross validation (CV) – with data splits at the report level – to evaluate the performance of the three models for Spatial Indicator extraction. The training, validation, and test sets are split in the ratio of 80%, 10%, and 10% respectively. There are a total of 1742 sentences with at least one Spatial Indicator and 31779 sentences without any Indicator in the dataset. To ensure that the performance of the models is not impacted due to the imbalance in the number of sentences with and without Spatial Indicators, we additionally run both the Bi-LSTM CRF and the $\text{BERT}_{\text{LARGE}}$ (MIMIC) models by randomly undersampling the negative sentences (i.e., sentences without an Indicator) while training. We experiment using different number of negative instances such that #negative sentences after undersampling = n * #positive sentences in each train and validation sets, where n = 1,2,3,4,5,6). We found that the performance of both the models (average F1 score of a 10-fold CV) improves as n is increased from 1 through 3 and starts to decline 4 onwards. Therefore, we select the value of n as 3 for conducting all our experiments. However, to evaluate the performance of the models on the full original dataset, we include all sentences in the reports of the test sets so that we get a more realistic sense of how well the models perform.

To better assess the generalizability of the models, we randomize the fold creation 5 times and conduct 10-fold cross validation for each fold variation. We then report the average Precision, Recall, and F1 measures across 50 (5*10) different instantiations for each model. We also include the 95 % confidence intervals of the average F1 measures.

Second, we evaluate the performance of the three models in extracting the spatial roles in context to a Spatial Indicator. We use the same fold settings and the same training, development, and test splits as in the Spatial Indicator extraction for spatial role labeling. For training and

validation, we utilize only the sentences containing a gold Spatial Indicator in the sentence. However, for testing, we experiment providing both the gold and the predicted Spatial Indicators (i.e., the output of the first model). The same trained model weights are used in predicting the roles using gold and predicted Indicators. We report the average Precision, Recall, and F1 measures of each of the 4 spatial roles across 50 instantiations for each model. We also calculate the overall measures of the three metrics considering all the roles collectively. We report the 95 % confidence intervals of the average overall F1 measures. Exact match is performed for evaluating the performance on the test set.

4.1.2 Description for Rad-SpatialNet

Here, we formulate both the tasks of spatial trigger (lexical units of spatial frames) identification and frame elements extraction as sequence labeling task. The BERT-based model architecture for sequence labeling is similar to the one described for Rad-SpRL in section 4.1.1.2. As an initial step of extracting the spatial trigger terms and the associated spatial information (spatial frame elements) in report sentences, we utilize both $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ pre-trained language models as our baseline systems. We initialize the model parameters obtained by pre-training $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ on MIMIC-III clinical notes (Si et al., 2019) for 300K steps and fine-tune the models on our constructed Rad-SpatialNet corpus. For fine-tuning, we set the maximum sequence length at 128, learning rate at 2e-5, number of training epochs at 4 and use cased version of the models.

10-fold cross validation is performed to evaluate the model's performance with 80-10-10% training, validation, and test splits of the reports. First, the spatial triggers are extracted in a sentence. Second, we extract the common frame elements with respect to the trigger. For element extraction, we evaluate the system performance using both the gold spatial triggers and the predicted triggers on the test set.

Table 4.1: Spatial Indicator extraction results: Average Precision, Recall, and F1 measures of 10-fold CV across 5 different fold variations. CI - 95% confidence intervals of the average F1 measures across 50 iterations.

Models	Precision(%)	Recall(%)	F1 (CI)
Bi-LSTM CRF	84.73	92.38	88.33 ($\pm$0.56)
$\text{BERT}_{\text{LARGE}}$	94.07	83.54	87.85 ($\pm$2.49)
$\text{BERT}_{\text{LARGE}}$ (MIMIC)	90.69	91.60	91.08 ($\pm$3.68)
$\text{XLNet}_{\text{LARGE}}$	88.62	94.40	**91.29** ($\pm$0.70)

4.1.3 Results on RadSpRL

The average results of the 10-fold CV across 5 different runs with fold variation are shown in Table 4.1 for Spatial Indicator extraction on the Rad-SpRL corpus. Note that we test the models on all sentences (both with and without Indicator). We see that either the recall or precision is higher than 90% for Bi-LSTM CRF, $\text{BERT}_{\text{LARGE}}$, and XLNet models. $\text{BERT}_{\text{LARGE}}$ (MIMIC) had better balance in precision and recall (both higher than 90%). The highest F1 score is obtained by $\text{XLNet}_{\text{LARGE}}$, which is 91.29.

For spatial role extraction, we report the average performance metric values of the 10-fold CV across 5 different fold variations, both considering the gold and the predicted Spatial Indicators in sentences of the test sets. Note that the test sets for each of the 50 different runs of the models are same for both Indicator and role extraction. We create a separate instance of a sentence for each of the predicted Spatial Indicators (in case multiple indicators are extracted by a model). When extracting the spatial roles using the predicted Spatial Indicators, we take into account all the spatial roles predicted for the false positive Spatial Indicators in calculating the precision loss, and consider the spatial roles predicted for the false negative Spatial Indicators in assessing the recall loss. This provides a more realistic end-to-end evaluation of the models.

The results using gold and predicted indicators are presented in Table 4.2 and Table 4.3, respectively.

Table 4.2: Spatial role extraction results using gold Spatial Indicators: Average Precision (P %), Recall (R %), and F1 measures of 10-fold CV across 5 different fold variations. CI - 95% confidence intervals of the average F1 measures across 50 iterations. BLSTM-C - Bi-LSTM CRF, BERT-L - $\text{BERT}_{\text{LARGE}}$, BERT-LM - $\text{BERT}_{\text{LARGE}}$ (MIMIC), XLNet-L - $\text{XLNet}_{\text{LARGE}}$.

Models	Trajector			Landmark			Diagnosis			Hedge			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1 (CI)
BLSTM-C	88.8	87.3	88.0	94.1	89.9	91.9	76.6	75.0	75.2	78.4	76.3	77.0	89.0	86.4	87.6 ($\pm$ 0.55)
BERT-L	89.7	91.8	90.7	95.4	96.1	95.8	72.7	85.5	78.4	72.8	84.1	77.8	88.8	92.4	90.5 ($\pm$ 0.42)
BERT-LM	91.2	93.1	92.1	95.6	96.6	96.1	72.3	83.9	77.4	75.0	86.1	80.1	89.5	93.3	91.4 ($\pm$ 0.54)
XLNet-L	92.8	94.1	**93.5**	96.1	96.8	**96.4**	78.6	88.0	**82.8**	79.6	88.6	**83.7**	91.6	94.2	**92.9** ($\pm$ 0.38)

We note that contextualized word representations help in improving spatial role extraction except for $\text{BERT}_{\text{LARGE}}$, which performed slightly inferior to the baseline model (Bi-LSTM CRF) when the predicted Spatial Indicators are used (see Table 4.3). XLNet performed the best (highest average overall F1 score of 92.9) in extracting the spatial roles when gold Indicators are used, however, its performance is comparable to $\text{BERT}_{\text{LARGE}}$ (MIMIC) when predicted Indicators are used (85.4 for XLNet and 85.6 for BERT with the same confidence interval). For Trajector, the highest average F1 for the end-to-end evaluation is 85.7, whereas for Landmark the highest average F1 is 89.3, both obtained by $\text{BERT}_{\text{LARGE}}$ (MIMIC) (Table 4.3). For all the models, the average F1 measures for Diagnosis and Hedge are comparatively lower than Trajector and Landmark, with the highest values being 79.0 and 78.6, respectively. Although the highest overall F1 is achieved by $\text{BERT}_{\text{LARGE}}$ (MIMIC) for the end-to-end evaluation, XLNet performed better in extracting the Diagnosis and Hedge roles.

4.1.3.1 Discussion

The results in Table 4.2 and Table 4.3 demonstrate that the models achieve promising results in extracting the spatial roles from the Rad-SpRL corpus. We observe that incorporating contextualized word representations by fine-tuning BERT (pre-trained on MIMIC) and XLNet models on the

Table 4.3: Spatial role extraction results using predicted Spatial Indicators: Average Precision (P %), Recall (R %), and F1 measures of 10-fold CV across 5 different fold variations. CI - 95% confidence intervals of the average F1 measures across 50 iterations. BLSTM-C - Bi-LSTM CRF, BERT-L - $\text{BERT}_{\text{LARGE}}$, BERT-LM - $\text{BERT}_{\text{LARGE}}$ (MIMIC), XLNet-L - $\text{XLNet}_{\text{LARGE}}$.

Models	Trajector			Landmark			Diagnosis			Hedge			Overall		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1 (CI)
BLSTM-C	77.7	83.9	80.6	83.0	86.6	84.7	72.3	72.2	71.7	71.5	72.5	71.6	78.8	83.1	80.8 (± 0.76)
BERT-L	85.7	73.9	78.1	90.0	77.2	82.8	73.1	75.3	73.8	71.4	71.8	71.1	85.1	75.2	79.3 (± 2.16)
BERT-LM	85.8	85.9	**85.7**	89.6	89.2	**89.3**	72.5	83.0	77.3	72.0	81.6	76.3	84.8	86.6	**85.6** (± 0.65)
XLNet-L	82.3	88.9	85.3	86.0	90.9	88.2	73.8	85.6	**79.0**	73.2	85.3	**78.6**	82.1	89.1	85.4 (± 0.65)

Rad-SpRL dataset performs better than a Bi-LSTM CRF network in extracting the Spatial Indicators as well as the spatial roles. Thus, $\text{BERT}_{\text{LARGE}}$ (MIMIC) and $\text{XLNet}_{\text{LARGE}}$ are currently the best performing models. However, more work is needed to determine which between these two models is more robust in extracting spatial information from chest X-ray reports. We also note that the average F1 measures are high for Trajector and Landmark roles and are comparatively low for Diagnosis and Hedge. The reason behind this can be attributed to the lesser number of Diagnosis and Hedge terms in the dataset (5 to 6 times less than both Trajector and Landmark terms) and greater distance between the Spatial Indicator and the Diagnosis/Hedge terms compared to the Trajector/Landmark terms.

Taking into account the relatively low F1 measure for Diagnosis and Hedge, we performed a brief analysis of the errors. On average, the best performing $\text{BERT}_{\text{LARGE}}$ (MIMIC) model in the end-to-end evaluation (shown in Table 4.3) misses around 10 % of the gold annotated Diagnosis terms, misclassifies 1 % of the terms as Trajectors, and misidentifies the beginning of around 2.6 % of the Diagnosis terms as inside. Some of the Diagnosis terms that are misclassified as Trajectors include *'bronchovascular crowding'*, *'edema'*, *'pulmonary fibrosis'*, *'atelectasis'*, and *'scarring'*. This is mainly because of the different ways certain common radiographic findings are also described as differential diagnoses. For example, in the sentence – *"Low lung volumes **with***

bibasilar opacities may represent bronchovascular crowding.", the DIAGNOSIS *'bronchovascular crowding'* is falsely classified as a TRAJECTOR. This might be because there are instances in the dataset where *'bronchovascular crowding'* appears as TRAJECTOR (e.g., in the sentence – *"There are low lung volumes **with** bronchovascular crowding as a result."*), as often a DIAGNOSIS term itself appears in a spatial relationship. The main reason for the errors associated with incorrect starting boundary of a predicted DIAGNOSIS term is that sometimes an extra adjacent term to the left of the actual DIAGNOSIS term is predicted by the model. For example, in *"Increasing prominence **of** the superior mediastinum may be secondary to enlarging thyroid mass."*, the model outputs *'enlarging thyroid mass'* as the predicted DIAGNOSIS instead of the annotated *'thyroid mass'*. For HEDGE, one of the major contributing factors of incorrect predictions of gold terms is that the $\text{BERT}_{\text{LARGE}}$ (MIMIC) model misses around 14 % of the gold annotated HEDGE terms. Most of these missed terms (e.g.,*'questionable'*, *'suggestion of'*, *'appears'*, *'alternatively'*) occur very infrequently in the dataset. Another challenge could be the variety of ways the hedging terms are used and positioned in a sentence to suggest any finding or differential diagnosis. Future work should attempt to improve the models to better handle complex description of sentences. Future work should also be directed toward building an end-to-end system based on neural joint learning models (Li et al., 2017; Miwa & Bansal, 2016) that would extract both SPATIAL INDICATOR and the spatial roles together, reducing discongruencies between predicted roles. From a method perspective, some alternative deep learning methods such as highway networks (Srivastava et al., 2015) and tree-based LSTMs (Miwa & Bansal, 2016) could be explored to further improve the performance of spatial role extraction from the Rad-SpRL corpus.

4.1.4 RESULTS ON RAD-SPATIALNET

For element extraction, we evaluate the system performance using both the gold spatial triggers and the predicted triggers on the test set. The results are shown in Table 4.4, Table 4.5, and Table 4.6.

Table 4.4: 10 fold CV results for spatial trigger extraction. P - Precision, R - Recall.

Model	P (%)	R (%)	F1
$BERT_{BASE}$ (MIMIC)	92.20	43.04	57.52
$BERT_{LARGE}$ (MIMIC)	93.72	67.13	77.89

Table 4.5: 10 fold CV results for extracting spatial frame elements using $BERT_{BASE}$ (MIMIC). P - Precision, R - Recall.

Main Frame Elements	Gold Spatial Triggers			Predicted Spatial Triggers		
	P (%)	R(%)	F1	P(%)	R(%)	F1
Figure	75.14	84.10	79.35	63.26	40.48	48.00
Ground	84.96	88.90	86.87	68.83	42.95	51.64
Hedge	64.31	77.46	69.78	56.91	31.84	38.69
Diagnosis	53.99	79.54	63.93	38.89	26.48	29.18
Relative Position	81.27	75.56	78.12	67.01	39.91	48.81
Distance	86.64	87.50	84.87	77.17	73.67	74.29
Position Status	62.29	64.07	62.74	56.86	52.05	52.39
Associated Process, Reason	0.0	0.0	0.0	0.0	0.0	0.0
Overall	76.53	82.69	79.48	63.98	40.73	48.55

The results demonstrate that $BERT_{LARGE}$ performs better than $BERT_{BASE}$ for both spatial trigger prediction and frame elements prediction. However, in case of spatial trigger, the recall is low for both the models (Table 4.4). For element extraction, $BERT_{BASE}$'s overall F1 combining all the frame elements is 79.48 (using gold spatial triggers) and 48.55 (using predicted triggers), whereas, for $BERT_{LARGE}$, the difference in the overall F1 between using gold (81.61) versus predicted triggers (66.25) is much lower (15.4 vs 30.9). The F1 values are zeroes for Reason and Associated Process because of few occurrences in the dataset.

4.1.4.1 Discussion

The results of the baseline system illustrates that it is difficult to identify the spatial trigger expression. This might be because of their wide variation in the reports and many of these appear

Table 4.6: 10 fold CV results for extracting spatial frame elements using $\text{BERT}_{\text{LARGE}}$ (MIMIC). P - Precision, R - Recall.

Main Frame Elements	Gold Spatial Triggers			Predicted Spatial Triggers		
	P(%)	R(%)	F1	P(%)	R(%)	F1
Figure	77.42	85.75	81.35	65.51	65.44	65.12
Ground	88.92	91.57	90.22	73.31	70.21	71.51
Hedge	67.20	77.94	71.59	60.43	57.26	57.82
Diagnosis	51.96	79.81	62.31	47.06	57.64	50.76
Relative Position	81.31	78.42	79.57	66.02	67.76	66.33
Distance	86.83	87.50	87.00	86.50	90.24	88.05
Position Status	65.73	65.83	65.38	58.59	63.63	60.37
Associated Process, Reason	0.0	0.0	0.0	0.0	0.0	0.0
Overall	78.83	84.64	81.61	66.63	66.28	66.25

as multi-word expressions. We only use the developed corpus to extract the core frame elements in a spatial frame. The results indicate that there is enough scope for improving the predictions by developing more advanced methods in the future.

4.2 Information Extraction as Question Answering

Recent research has focused on formalizing various information extraction (IE) tasks as question answering (QA). More specifically, a few studies (Li et al., 2020b; Sun et al., 2020; Banerjee et al., 2020) have demonstrated the effectiveness of formulating named entity recognition (NER) as machine reading comprehension (MRC) instead of the traditional sequence labeling technique in both general and biomedical domains. Apart from NER, prior work has also framed relation (Levy et al., 2017) and EE (Liu et al., 2020) tasks as MRC. Additionally, several studies (Li et al., 2019, 2020a; Wang et al., 2020) have explored a new paradigm by formulating relation and event extraction (EE) tasks as multi-turn QA, an approach that involves performing multiple turns of machine reading comprehension (MRC) successively. Advantages of framing extraction tasks as MRC include leveraging prior knowledge through queries, jointly modeling entities and relations

Table 4.7: Target entities extracted in turn 1.

Target entity type	Description	RadLex Class
Spatial trigger	Spatial prepositions (e.g., *in*), verbs (e.g., *demonstrate*), verb followed by prepositions (e.g., *projected at*), etc.	Not applicable
Finding	Terms related to radiological observations, clinical findings (including those suggesting diagnoses)	Clinical finding, Imaging observation
Anatomy	Anatomical location	Anatomical entity
Device	Medical device	Medical device
Tip	Tip of a medical device	Portion of medical device
Location descriptor	Describing how a finding is located with respect to an anatomy	Location descriptor
Other descriptor	Modifiers describing a radiological observation or finding	RadLex descriptor (except Location and Certainty)
Assertion	Uncertainty and negated phrases used by radiologists	Certainty descriptor
Position	Position status of a device (e.g., *good position*)	Not applicable
Quantity	Any quantitative term in the report text (e.g., *3 mm*)	Not applicable
Process	Describing motion, change, etc.	Process

in the form of natural language questions, and making use of advanced MRC models. In addition, multi-turn QA captures the hierarchical dependency of entities and is therefore suitable for complicated scenarios where extraction of certain entities depends on previously extracted entities. Previous approaches have formulated spatial trigger and element (or role) extraction as a sequence labeling task either in a pipelined or joint learning fashion (Bastianelli et al., 2013). However, inspired by the advantages that multi-turn QA provides, we propose to adopt a two-turn QA approach by harnessing MRC models for identifying fine-grained spatial and descriptor information of radiographic findings and medical devices described in radiology reports..

Table 4.8: Frame Elements extracted in turn 2, their descriptions, and associated entity types. ST - Spatial Trigger. Desc - Descriptor.

	Frame Elements	Description	Entity Types of Related Entities
Spatial	Figure	Object whose location is described	ST; Finding/Anatomy/Device/Tip
	Ground	Anatomical location of Figure	ST; Anatomy
	Hedge	Uncertainty expressions used by radiologists	ST; Assertion
	Diagnosis	Clinical condition or disease associated with a radiological finding	ST; Finding
	Position Status	Any position-related information, usually in context to a device	ST; Position
	Relative Position	Terms used for describing the orientation of a radiological entity wrt to an anatomical location	ST; Location descriptor
	Distance	Actual distance of finding or device from the anatomical location	ST; Quantity
	Reason	Clinical condition or disease that acts as the source of a radiological finding	ST; Finding
	Associated Process	Any process or activity associated with a spatial relation	ST; Process
	Morphologic	Indicates shape	Finding/Anatomy; Desc
	Size Desc	Indicates size description	Finding/Anatomy/Device; Desc
	Distribution Pattern	Indicates distribution patterns	Finding/Anatomy; Desc
	Composition	Indicates composition of a radiological finding	Finding/Anatomy; Desc
	Laterality	Indicates side	Finding/Anatomy/Device; Desc
	Size/Measurement	Actual size of a finding	Finding; Desc
Desc	Status	Indicates status of entities	Finding/Anatomy/Device; Desc
	Quantity	Indicates quantity of a radiological entity	Finding/Anatomy/Device; Desc
	Temporal	Indicates temporality	Finding/Device; Desc
	Negation	The associated negated phrase	Finding/Anatomy; Desc

4.2.1 Description for Rad-SpatialNet

4.2.1.1 Problem Formulation

We formulate the spatial IE problem as a machine comprehension problem where information is extracted from a given text (treated as a context paragraph) using templates posed as queries to elicit specific information (triggers and frame elements). The answer spans returned by the MRC system are treated as the extracted entities. In case the system returns a special token NONE, this indicates that the specific entity that is queried for is not present in the report text. Analogous to how a conventional entity-relation extraction system is employed, i.e., first identifying the target entity and then identifying the related entities, our MRC formulation is also designed in two turns/steps as one-time QA is not sufficient to capture this dependency in the information extracted. The target entities extracted in the first turn are mentioned in Table 4.7. These entities cover a wide range of common radiology terms curated as part of the radiology lexicon, RadLex (Langlotz, 2006) (see Table 4.7). The second turn identifies the spatial frame elements (FEs) associated with a spatial trigger (e.g., Figure, Ground, Diagnosis) as well as spatial (e.g., Laterality) and descriptive (e.g., Status) FEs associated with a radiological entity. These frame elements are described in Table 4.8.

4.2.1.2 Query Construction

Entity and frame element (FE) type modification The entity and FE types are used in forming the queries. The entity types except 'Spatial Trigger', 'Location Descriptor', and 'Quantity' are modified, as shown in Table 4.9, while incorporating in a query. We modified the entity types to incorporate more information about the entities in the queries as well as to make the queries sound more natural. The FE types (corresponding to element names) are used in the queries of the second turn without modification except for the 'Diagnosis' element in which case it is modified to 'Potential diagnosis'.

Table 4.9: Modified entity types to be used in queries.

Target entity type	Modified entity type
Finding	Clinical finding
Anatomy	Anatomical structure
Device	Medical device
Tip	Medical device tip
Other descriptor	Descriptor
Assertion	Assertion-related
Position	Position-related
Process	Associated process

Target entity extraction The modified entity type (of a spatial trigger and a main radiological entity) is converted to a query using a template. This query variant is referred to as $\text{Query}_{\text{find}}$ (shown in Table 4.10).

Spatial and descriptive frame element extraction Each FE is converted to a query, $\text{Query}_{\text{find}}$ (see Table 4.10), such that the query asks for identifying the text span (belonging to a specific entity type) from the report that has the particular FE relation to a target entity type. The query template contains a slot corresponding to the target entity type (ENT_1) that is filled by the previously extracted entity (ENT_1_SPAN) from the first turn. Thus, this query jointly extracts the FE relation (REL) as well as the related entity (of type ENT_2) in the form of an answer span that is predicted by the MRC model. Using this template, queries are formed such that all FE relations are covered for all possible pairs of target and related entity types. For example, "*find all **medical device** entities in the context that have a **figure** relationship with **spatial trigger** entity **above**.*" is the query constructed for the triplet {spatial trigger, Figure, medical device} where spatial trigger is ENT_1, Figure is REL, and medical device is ENT_2 that is extracted by answering this query. If the answer is NONE, this means there is no such related entity in the text that is associated to the target entity through REL.

We also experiment with another query variation for the second turn. In this, we encode domain knowledge in the query by incorporating a general description of the FE. That is, we prepend a description of the spatial FE (SFE) or descriptive FE (DFE) at the beginning of a query. We refer

Table 4.10: Query template and example. Q_f: $\text{Query}_{\text{find}}$. Q_{f+d}: $\text{Query}_{\text{find + desc}}$.

Extraction step	Query template	Example
Target entities	***Entity type*** **:** ENT Q_f **:** find all **ENT** entities in the context.	***Entity type*** **:** Spatial trigger Q_f **:** find all **spatial trigger** entities in the context.
Spatial and descriptive frame elements	***Frame element type*** **:** REL ***Target Entity type*** **:** ENT_1 **ENT_1 span from turn 1 :** ENT_1_SPAN ***Related Entity type*** **:** ENT_2 Q_f **:** find all **ENT_2** entities in the context that have a/an **REL** relationship with **ENT_1** entity **ENT_1_SPAN**. Q_{f+d} **:** a general **description** about **REL** + Q_f	***Frame element type*** **:** Figure ***Target Entity type*** **:** Spatial trigger **Spatial trigger span from turn 1 :** *in* ***Related Entity type*** **:** Clinical finding Q_f **:** find all **clinical finding** entities in the context that have a **figure** relationship with **spatial trigger** entity ***in***. Q_{f+d} **: Figure** refers to finding or device or tip entities that are described with respect to an anatomical structure. + Q_f

to this query variation as $\text{Query}_{\text{find + desc}}$ (see Table 4.10 for template and example). The descriptions developed for each of the spatial and descriptive FEs are listed in Tables 4.11 and 4.12, respectively.

4.2.1.3 MRC Framework

The MRC architecture is based on the pre-trained language model BERT (Devlin et al., 2019). Previous work achieved promising results using BERT-based MRC models for QA (Devlin et al., 2019; Liu et al., 2019; Qu et al., 2019). We select this model framework owing to the promising performance of using BERT for QA as well as to tackle multi-answer QA. We follow the standard format to feed input into the BERT model for answering queries. We split the whole content of a radiology report into overlapping passages by sliding window and use each passage as context c into the BERT model after combining with the query q. This sliding window technique proved to be effective as evidenced by prior work (Wang et al., 2019c; Li et al., 2019) . After WordPiece tokenization of both query and context, we merge the query q and the context c as [[CLS] q [SEP]

Table 4.11: Descriptions for spatial frame elements used in Queryfind + desc for the second turn.

Spatial Frame Element	Description
Figure	Figure refers to finding or device or tip entities that are described with respect to an anatomical structure
Ground	Ground refers to the main anatomical structures
Hedge	Hedge refers to uncertainty phrases describing a finding or diagnosis. Examples include may represent and suggestive of
Diagnosis	Diagnosis refers to a clinical condition or disease associated with a finding. This is suggested as potential diagnosis and usually appears after the hedge related terms
Position Status	Position status refers to any position-related information, usually in context to a device. Examples include terminates and expected position
Relative Position	Relative position refers to terms describing the orientation of a finding wrt to an anatomical structure. Examples include posteriorly, level of, and high
Distance	Distance refers to the actual distance of finding or device from the anatomical structure. Examples include 3.6 cm and 4 mm
Reason	Reason refers to a clinical condition or disease that caused a radiographic finding to be detected. This usually appears after hedge terms like could be due to
Associated Process	Associated process refers to any process/activity associated with a spatial relation. Example includes intubation
Morphologic Descriptor	Morphologic descriptor indicates shape of findings. Examples include rounded
Size Descriptor	Size descriptor indicates size of a finding. Examples include large and small
Distribution Pattern	Distribution pattern indicates patterns such as scattered and diffuse
Composition Descriptor	Composition descriptor indicates composition of a finding. Example includes calcified
Laterality	Laterality indicates side. Examples include left and right
Size/Measurement	Size refers to the actual size of a finding. Examples include lesion size such as 14-mm
Density Descriptor	Density descriptor refers to densities of findings. Examples include hypodense and lucent
Modality Characteristic	Modality characteristic refers to characteristics such as attenuation

Table 4.12: Descriptions for descriptive frame elements used in Queryfind + desc for the second turn.

Descriptive Frame Element	Description
Status Descriptor	Status descriptor indicates status of findings. Examples include stable and mild
Quantity Descriptor	Quantity descriptor refers to quantity of any radiological entity. Examples include multiple and few
Temporal Descriptor	Temporal descriptor indicates temporality. Examples include new and chronic
Negation	Negation refers to a negated phrase related to a finding. Examples include without evidence of and no

c [SEP]] to construct the input sequence where [CLS] and [SEP] are special BERT tokens. As explored in previous work (Li et al., 2019, 2020b), the span extraction mechanism enables queries that have multiple answers given the context passage. Traditional approaches strategize span extraction as two *n*-class classification problems where one classifier predicts the start index and the other predicts the end index from all the context tokens (*n* refers to the length of the context passage). However, this strategy is only applicable for single-answer QA settings. To overcome this shortcoming, the two *n*-class classification task is converted to *n* 5-class classifications where the softmax function is applied to each token in the context to predict a BMESO (B-begin, M-middle, E-end, S-single, O-outside) label. This is suitable to our problem where there can be multiple entities of the same spatial or descriptor role that are associated to a single spatial trigger or a radiological finding (see Figure 4.3). The BERT models for both target entity (turn 1) and FE extraction (turn 2) are trained jointly. The MRC framework and the training mechanism are adopted from a previous work (Li et al., 2019).

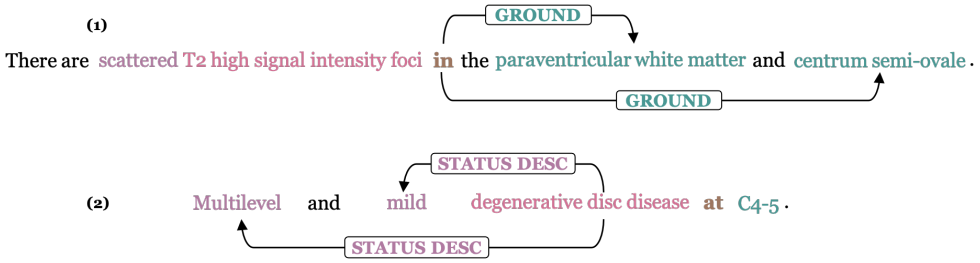

Figure 4.3: (1) Two Ground elements are linked to a spatial trigger and (2) two status descriptors are linked to a radiological finding. For (1), the query for extracting anatomical locations with respect to the spatial trigger *in* should return two spans – *paraventricular white matter* and *centrum semi-ovale*. For (2), an MRC model is expected to return the spans corresponding to *Multilevel* and *mild* as output when queried for the status descriptive elements of *degenerative disc disease*.

4.2.1.4 Sequence Labeling Baseline

We compare our approach to the BERT-based sequence labeling approach described in the section 4.1.2 above. In that paper, a $\text{BERT}_{\text{LARGE}}$ model pre-trained on MIMIC-III is fine-tuned to first extract all the spatial triggers in a sentence and then extract the spatial FEs associated with each trigger. A report sentence is represented as [[CLS] *sentence* [SEP]] to feed into BERT to identify the triggers and FEs. Additionally, while extracting the spatial FEs, we mask the spatial trigger identified in the first step to better encode the position of the specific spatial trigger in a sentence for which the FEs are to be identified. The encoder output is then fed into a linear classification layer to predict labels per token. The BIO (B-begin, I-inside, O-outside) scheme is used to tag the triggers and the FEs.

4.2.1.5 Experimental Settings and Evaluation

We experiment with both cased and uncased $\text{BERT}_{\text{LARGE}}$ variants in the MRC framework (referred to as Uncased and Cased hereafter). Additionally, we also experiment using a $\text{BERT}_{\text{LARGE}}$ cased version that is pre-trained on MIMIC-III clinical notes for 300K steps Si et al. (2019) (referred as MIMIC+Cased). The hyperparameters used in our experiments are selected based on the validation set and are shown in Table 4.13. For training the MRC model in the second turn, we only consider the relationships between target and related entities where there is at least one instance of such a relationship in the training data.

We perform 10-fold cross-validation (CV) for evaluating our MRC approach for spatial IE. For each of the 10 iterations, we split the dataset such that reports in 8 folds are used for training and 1 fold each are used for development and testing. We report the average F1 measures for extracting the FEs. Since the query format is the same for the first turn (i.e., target entity identification) and only varied in the second turn (corresponding to using $\text{Query}_{\text{find}}$ and $\text{Query}_{\text{find + desc}}$), we report the average of the two 10-fold CV runs for target entity extraction. We use exact match to evaluate the performance of the MRC approach for both target entity and FE extraction on the test splits. Exact matches of both the target and the related entity spans are required to consider a FE relation extraction as a true positive. We compare our approach to the baseline method for identifying spatial triggers and SFEs connected to triggers. For a fair comparison, we use the same fold settings for 10-fold CV for both the MRC and the baseline methods. The baseline method is also evaluated using exact match for spatial trigger and SFE extraction.

4.2.2 Results on Rad-SpatialNet

The average F1 measures of 10-fold CV evaluation for extracting the spatial and descriptive FEs are shown in Table 4.14. This includes the results for both the query variations. The average

Table 4.13: Hyperparameters used in the experiments.

Parameter	Value
Sliding window size for context passage	200
Overlap between adjacent windows	45
Maximum number of training epochs	10
Learning rate	2e−5
Trade-off between two turns	0.25
Maximum norm for gradients	1
Warmup ratio	0.1

F1 scores of the BERT-based sequence labeling baseline method are also shown in Table 4.14 for comparison. Since density descriptor and modality characteristics occur very infrequently in the dataset (5 and 2 times, respectively), we do not report the results for these two FEs. For extracting SFEs associated with triggers, we see that $\text{Query}_{\text{find + desc}}$ helps in achieving a better performance than $\text{Query}_{\text{find}}$ for all elements (except for Hedge) in the case of MIMIC+Cased model. Whereas, for Uncased and Cased variants, $\text{Query}_{\text{find + desc}}$ performed better for some of such SFEs. We also note that the performance of less frequent FEs: Reason and Associated Process improved to 49.81 and 54.63 compared to baseline system's F1 (0 for both). For the majority of the FEs associated with a radiological entity, the average F1 scores lie in the range of 60-75. However, for Laterality and Size Descriptor, the values are relatively high with the highest F1 scores being 89.35 and 78.98, respectively.

The results for target entity extraction are shown in Table 4.15. We observe that the best F1 score for identifying the spatial triggers obtained by our proposed method (90.07) is around 12 points higher compared to the baseline system's performance of 77.89 (as reported in Table 4.4 above). The entity labels of the target entities are included during the FE extraction to make the queries more informative and are not part of the FE performance evaluation.

Table 4.14: Average F1 measures of $\text{BERT}_{\text{LARGE}}$ models over 10-fold CV for spatial and descriptive frame element extraction. Desc: Descriptive. Q_f: $\text{Query}_{\text{find}}$. Q_{f+d}: $\text{Query}_{\text{find + desc}}$. M+Cased: MIMIC+Cased. Count: Number of annotations in the dataset. Dash (–): not available for baseline method.

	Frame Elements	Proposed approach						Baseline M+Cased	Count
		Uncased		Cased		M+Cased			
		Q_f	Q_{f+d}	Q_f	Q_{f+d}	Q_f	Q_{f+d}		
Spatial	Figure	**78.13**	77.29	76.72	77.57	76.44	77.40	65.12	1491
	Ground	83.76	83.40	83.31	82.27	83.17	**83.77**	71.51	1537
	Hedge	75.47	76.44	**77.18**	76.42	75.90	74.97	57.82	249
	Diagnosis	69.32	73.32	**73.94**	72.67	65.47	67.92	50.76	190
	Position Status	68.72	68.75	66.98	67.12	68.43	**70.37**	60.37	167
	Relative Position	77.19	76.42	**77.53**	76.71	75.78	76.15	66.33	398
	Distance	84.65	86.54	85.36	85.20	87.94	**90.09**	88.05	45
	Reason	39.51	32.34	39.51	**49.81**	17.71	44.89	0	33
	Associated Process	48.52	**54.63**	43.15	42.29	38.95	41.36	0	21
	Morphologic	52.48	58.14	49.92	**60.52**	48.04	45.53	–	69
	Size Desc	76.16	73.80	78.16	78.94	78.56	**78.98**	–	93
	Distribution Pattern	57.45	63.62	59.74	64.01	59.22	**66.03**	–	65
	Composition	41.46	33.63	41.67	**46.88**	26.49	20.48	–	17
	Laterality	88.43	88.51	**89.35**	87.49	87.78	87.32	–	612
	Size/Measurement	45.43	**48.44**	41.51	43.59	34.46	32.06	–	23
Desc	Status	**64.67**	62.60	63.38	61.67	59.17	59.09	–	452
	Quantity	72.56	72.32	72.82	71.61	72.47	**73.11**	–	130
	Temporal	70.87	70.63	70.5	71.47	67.31	**71.78**	–	113
	Negation	58.08	61.06	**67.75**	65.04	60.95	61.83	–	103

4.2.2.1 Discussion

The results in Tables 4.14 and 4.15 demonstrate the performance improvement in extracting spatial information from radiology reports when the problem is framed as MRC compared to traditional sequence labeling. The improvements are high: for example, improvement of average F1 scores from 65.12 to 78.13 and 71.51 to 83.77 for common FEs like Figure and Ground, respectively. This highlights the advantages of framing IE as MRC as described in the Section 2.2 of Chapter 2.

Table 4.15: Average F1 measures of $\text{BERT}_{\text{LARGE}}$ models over two 10-fold CVs for target entity extraction. M+Cased: MIMIC+Cased.

Target entities	Uncased	Cased	M+Cased
Spatial trigger	89.99	89.50	**90.07**
Finding	76.89	**78.26**	76.11
Anatomy	**87.56**	87.40	87.46
Device	91.87	92.68	**93.12**
Tip	99.18	98.41	**99.32**
Location descriptor	**81.50**	81.21	80.89
Other descriptor	84.19	**84.24**	84.09
Assertion	78.48	**80.85**	79.40
Position	69.68	71.41	**72.81**
Quantity	**85.54**	85.37	83.23
Process	**60.93**	59.26	60.19

We also note that casting IE problems as MRC is still under-explored on clinical domain datasets except for Banerjee et al. (2020). This is the only other case we are aware of MRC being used for IE from clinical reports, and there it is used only for entity extraction, not relation extraction and not with a two-turn QA approach. Moreover, we emphasize that our work covers more detailed radiological information from spatial and descriptor perspectives and extracts information from reports of multiple imaging modalities and anatomies (as opposed to previous work (Syeda-Mahmood et al., 2020; Sugimoto et al., 2021; Steinkamp et al., 2019) focusing on either single modality or anatomy). This is the first study to use MRC both for spatial IE and for extracting important radiology information.

Our investigation using two query variations for FE extraction suggests that incorporating more information about the element in the query helps in obtaining better results, especially in cases where the meaning of FE is not obvious solely based on the FE type name. For example, we see performance improvement for Position Status for all model variants when the following description about Position Status is included in the query:

Position status refers to any position-related information, usually in context to a device. Examples include terminates and expected position.

This provides more prior knowledge about what is meant by Position Status in a radiology report context. We also find that our proposed approach tends to perform better than sequence labeling for less frequent FEs (e.g., Reason). Note that the MIMIC pre-trained BERT model underperforms both the original Uncased and Cased models for the majority of the infrequent FEs such as Associated Process, Composition Descriptor, and Size/Measurement.

Alongside starting the queries with 'Find all', we explored two other query variations – beginning the queries with 'Get all' and 'What' (inspired by previous work by Banerjee et al. (2020)). Although these variations performed better than the baseline, we did not find any clear performance trend when compared to the 'Find all' variant. An exhaustive comparison of query variations could be investigated further, but that was not the focus of this work.

The moderate performance values as well as the performance variation for some FEs could be due to the infrequency of annotations in the dataset. This indicates that there is still scope for improving the results and we aim to evaluate our approach on an enlarged dataset that will have more such FEs. Although we apply our proposed method on a dataset that covers three types of radiology reports, we further intend to evaluate the generalizability of this method on multi-institutional datasets and on other imaging modalities (e.g., ultrasound and CT reports of different body parts) in a subsequent work.

5

Normalization of Radiological Entities using RadLex

Radiology reports contain a wide range of entities describing the interpretations of the corresponding images. Prior research (Steinkamp et al., 2019; Bozkurt et al., 2019; Fu et al., 2020) has focused on developing methods to identify clinically-significant information from these reports. They emphasize using this extracted information in a variety of downstream clinical applications including automated tracking of abnormal radiographic findings (e.g., lesions), summarization, and cohort selection for epidemiological research. However, to enable the use of the extracted entities in the process of developing the automated systems across multi-institutional reports, the entities need to be mapped to concepts in a standardized vocabulary of radiology terms. There has been limited efforts in this direction, and therefore, in this chapter, we aim to normalize the different entities or

Report 1: CHEST PRE-OP PA
There is **volume loss** in the **left upper lobe**.
There is **pleural thickening** involving the **left apex**.

Report 2: CHEST SINGLE VIEW
There is **marked volume loss** in the **left upper lobe region**.
Marked left apical thickening is noted and also **stable**.

Figure 5.1: Partial section of two radiology reports. Findings are shown in green, anatomical locations are in blue, and the descriptor terms in purple. The RadLex concepts corresponding to the two anatomical locations are upper lobe of left lung (RID1327) and left lung (RID1326) + apex (RID5946)

.

information types to RadLex (Langlotz, 2006) concepts to facilitate improved consistency in the structured representations of important radiological entities.

Radiologists use different phrases to express the same concept in a report. Normalization is the process of mapping these phrase spans in text to standard concepts in a vocabulary. For instance, *Right base* and *Right lower lung zone* are different forms of describing the same anatomical entity. Similarly, *Intramural or free air* is used by radiologists to indicate the clinical finding - *Pneumatosis intestinalis*, and *Central line*s are often used to denote the presence of *Central venous catheter*s in chest X-ray reports. Many natural language processing (NLP)-based clinical application systems rely on developing inference rules. Such inferences are often performed on the entities that are extracted from clinical text by entity recognition systems. Consider the example shown in Figure 5.1 to illustrate the usability of normalizing radiological entities in the report text for automated abnormality tracking systems.

The sentences in Figure 5.1 appear in two different reports of a patient. *Volume loss* and *Pleural*

thickening are the main finding-related entities while *left upper lobe* and *left apex* are the anatomy-related entities in the first report. The second report also contains the same findings except that the anatomical phrases are changed to *left upper lobe region* and *left apical* and there are additional clinically-relevant status descriptors such as *marked* and *stable*. Utilizing the extracted contextual information (e.g., anatomical phrases here) in the automated tracking of *volume loss* and *thickening* would require establishing a standardized way to represent the extracted anatomical entities (i.e., mapping both *left upper lobe* and *left upper lobe region* to RadLex concept ***upper lobe of left lung (RID1327)***). To our knowledge, only one study (Tahmasebi et al., 2019) so far has worked on normalizing the anatomical terms in the reports to SNOMED CT ontology. Here, we attempt to broaden the scope of entity types and consider all those that can act as contextual information in various potential clinical use cases. Moreover, since RadLex (Langlotz, 2006) is a publicly available radiology lexicon (containing 46,657 concepts) specifically developed for standardizing the language used in reporting imaging results, we utilize RadLex for mapping the various entity mentions in the reports. This will cover entities such as common modifiers and uncertainty phrases often encountered in radiology report text and may not be present in other ontologies such as SNOMED CT.

5.1 Dataset Annotation

We selected a subset of 50 radiology reports from MIMIC-III clinical database (Johnson et al., 2016). This consists of 17 chest X-ray reports, 16 Brain Magnetic Resonance Imaging (MRI) reports, and 17 babygram-related reports. This set of reports covers some common imaging techniques and a wide range of anatomical locations (as often babygrams contain descriptions of multiple body organs). We used the BRAT annotation tool (Stenetorp et al., 2012a) for annotating the radiological entities with their corresponding RadLex IDs as shown in Figure 5.2.

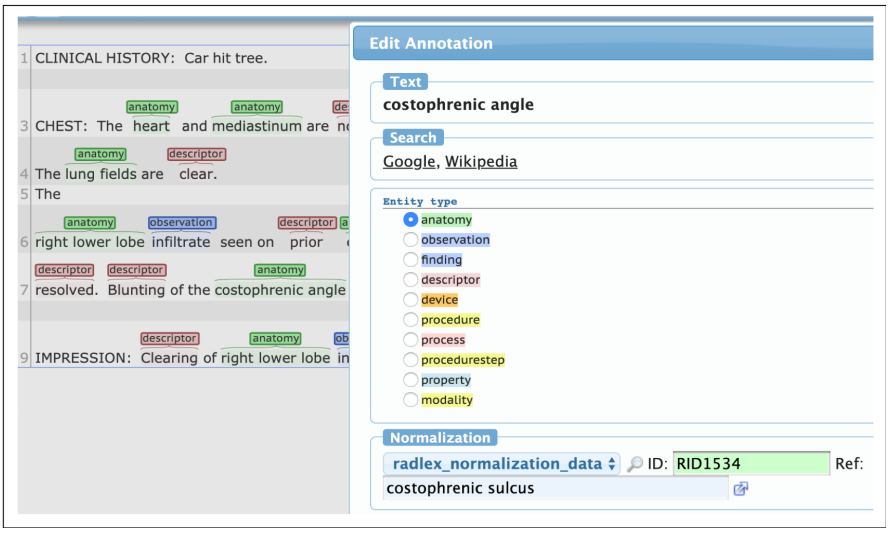

Figure 5.2: Example annotation to normalize "*costophrenic angle*" to RadLex term "*costophrenic sulcus*" corresponding to RadLex ID RID1534 in a sample report using BRAT 1.3.

5.1.1 Annotation Process

We describe the annotation process in the following sections.

5.1.1.1 Identifying Entity Spans

The first annotation task is to identify the entity mention in the report sentences whose type falls under one of the following broad RadLex classes:

1. **Clinical Finding** - Refers to pathophysiologic finding, and symptoms (e.g., *heart failure*)
2. **Imaging Observation** - Image-specific features as interpreted by radiologists (e.g., *infiltrate*)
3. **Anatomical Entity** - Refers to a body location (e.g., *apex of lung*)
4. **Medical Device** - Refers to a medical object (e.g., *endotracheal tube*)

5. **RadLex Descriptor** - Any modifier (usually adjectives) used to describe other entities like clinical finding (e.g., status descriptor - *stable*, composition descriptor - *osseous*, certainty descriptor - *no*, etc.)

6. **Procedure** - This includes different procedures such as imaging procedures, follow-up procedures, and treatment. (e.g., *catheter removal*)

7. **Procedure Step** - Includes any step in image processing (e.g., *multiplanar reformat*)

8. **Process** - Usually refers to treatment planning, change etc. (e.g., *motion*)

9. **Imaging Modality** - Form of imaging that depends on how the image is produced (e.g., *magnetic resonance imaging*)

10. **Property** - Modifier terms (usually noun phrases) associated with entities (e.g., *patient rotation position*)

5.1.1.2 Instructions for Assigning RadLex Codes

The next step involves assigning a single RadLex ID to each of the identified entity mentions. Note that we have mapped each entity to only one RadLex ID. For instance, the anatomical entity "*Midline structure*" is mapped to the RadLex concept "*Septum pellucidum*" with RadLex ID RID6525. While assigning the RadLex ID, the following instructions were given to the annotators:

1. Search for the exact entity span in RadLex

2. If not found using 1, search whether it appears in RadLex with different a variation such as with words rearranged in a different order (e.g., assigning RadLex concept *apex of lung* for the entity mention *lung apex*)

3. If not found using 1 and 2, search whether it appears as a synonym or in the decription of another RadLex concept (e.g., *Costophrenic sulcus* is present as a synonym of the RadLex concept *Costophrenic angle*)

4. If not found using the above, refer the web to look for the most semantically similar concept in RadLex (e.g., *Chest tube* is mapped to the RadLex concept *Thoracostomy tube* following this guideline)

5. If an entity cannot be mapped to any RadLex concept, it is assigned a label "RID-less"

Moreover, while annotating entities following the above instructions, the following points are taken into consideration:

- **Taking context into account** - Annotation of some entities may vary based on the context of the sentence or the anatomical entity associated with the imaging modality. For example, *Microangiopathic changes* refers to a disease/condition affecting small blood vessels. So depending on the anatomical entity associated with the imaging modality, the mapping would vary. When the imaging results are related to heart, *Microangiopathic changes* would be mapped to *Microvascular ischemia* in RadLex which is listed under *Cardiovascular disorder*, whereas for brain-related imaging results *Microangiopathic changes* will not be assigned any RadLex code.

- **Splitting entity spans to subspans** - Many times an entity mention cannot be mapped to a specific RadLex concept and annotators tend to use multiple RadLex concepts in different combinations to annotate that entity. For example, *Right middle lobe* is not directly normalizable to a RadLex concept and may be mapped to concepts like *lobe*, *middle lobe of lung*, or *right*. In order to resolve ambiguity in the annotation process, each entity mention is split such that the split with the largest subspan can be mapped to a RadLex code and all the other smaller

Table 5.1: Descriptive statistics of the annotated corpus.

Item	Frequency
Clinical Finding	282
Imaging Observation	77
Anatomical Entity	384
Medical Device	102
RadLex Descriptor	651
Procedure-Related	46
Process	28
Imaging Modality	51
Property	85
Total entity mentions	1706
Unlinkable mentions	151

subspans are also mapped to their corresponding RadLex codes. Thus, "**Right middle lobe**" will be split as "**middle lobe**" (largest RadLex mappable subspan) + "**right**" and not as "**right**"+ "**middle**" + "**lobe**". Further, "middle lobe" will be normalized to "middle lobe of lung" (RID1310) and "right" to "right" (RID5825).

However, there may arise cases when multiple possible variations of largest mappable subspans can be generated. For example, "Right lung apex" results in two valid (RadLex mappable) splits, one with "Right lung" + "apex" and the other with "Right" + "lung apex". This can be resolved during reconciliation phase by incorporating domain knowledge (e.g., knowledge on human anatomy) and further verification by a physician. This may favor the first option – "Right lung" + "apex" as this is more close to describing the apex of right lung.

Every report was double-annotated and reconciled with the clinical knowledge verified by a physician when required. The F1 agreement between the two annotators in annotating the spans of radiological entity mentions is 0.60. We considered an exact match in the entity spans for calculating the F1 score. The normalization agreement (accuracy) between the annotators on the reconciled version of the entity mentions is 76.7%. Basic statistics of our annotated corpus are shown in Table 5.1.

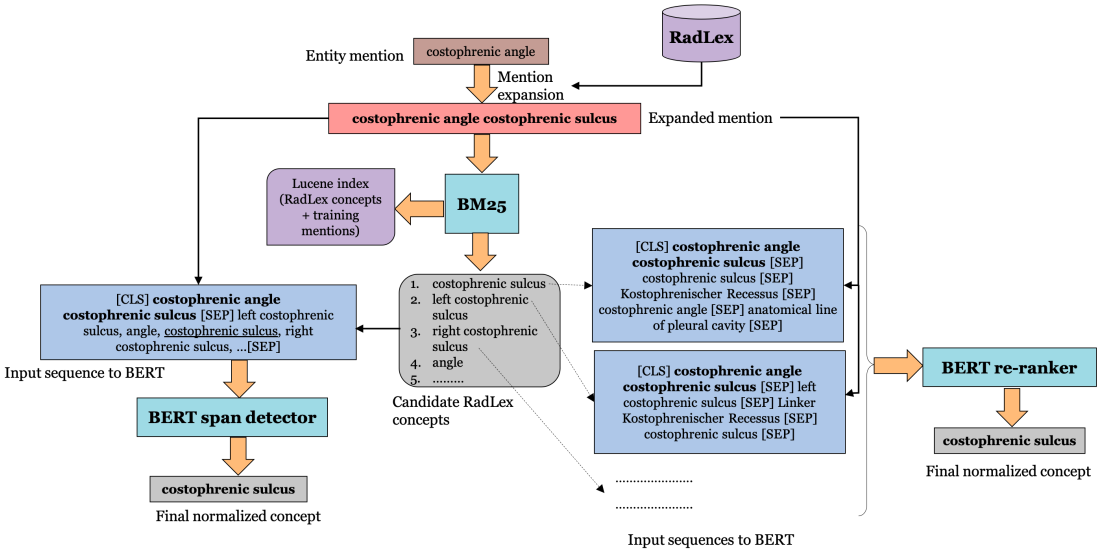

Figure 5.3: Overview of the normalization process using the proposed methods (demonstrated for the entity mention-"*costophrenic angle*").

5.2 Entity span detection

We formulate this as a sequence labeling task where each word that is part of any radiological entity of interest is tagged using Beginning and Inside tags whereas a word that is not a part of an entity is tagged as Outside. Each sentence in the reports is WordPiece-tokenized. This tokenized sentence is represented as [[CLS] sentence [SEP]] following the original paper (Devlin et al., 2019) and then fed into the BERT model.

5.3 Normalization Methods

The following subsections contain the descriptions of the three methods used for RadLex normalization. The overall framework of the normalization methods is illustrated in Figure 5.3.

5.3.1 BM25

We index all RadLex concepts (a total of 46,657 Preferred Names in RadLex) as well as the entity mentions present in the training sets of our annotated corpus using Anserini (Yang et al., 2018). We then use BM25 to retrieve and initially rank a set of n candidates for each entity mention. (In our experiments, we use $n = 10$.) We set the values of BM25 parameters, b and $k1$, as 0.75 and 1.2, respectively. In order to maximize the recall of BM25 in the candidate generation phase, each entity mention is transformed using the following two expansion techniques:

1. Using ***Synonyms*** in RadLex - If the entity mention (m) appears as a Synonym of a RadLex concept (r_c), the original mention is expanded using the Synonym. For example, "encephalopathy" is not present in RadLex but appears as a Synonym of the RadLex concept "disorder of brain" (RID5055). Thus, the mention "encephalopathy" is expanded to "encephalopathy disorder of brain".

2. ***Abbreviation*** expansion - Often, some common medical devices and clinical findings are abbreviated in the reports. We expand these mentions leveraging the medical abbreviations and acronyms of radiopaedia*. For instance, "NGT" is expanded to "nasogastric tube" and "NPH" is expanded to "normal pressure hydrocephalus".

5.3.2 BERT as re-ranker

We use the set of candidate concepts obtained from BM25 for each entity mention to train the $\text{BERT}_{\text{BASE}}$ model for re-ranking these candidates. The highest ranked candidate predicted by BERT is chosen as the final normalized RadLex concept for a given radiology entity mention. The model is trained as a binary classification task where for each candidate concept (c_i) and the mention (m) pair, the label is assigned as 1 when the candidate is the actual annotated normalized concept

*https://radiopaedia.org/articles/medical-abbreviations-and-acronyms-a?lang=us

for that mention. For each such pair of candidate concept and entity mention, a score is estimated to predict the likelihood of the candidate concept being the normalized concept. Note that we use the expanded version of the entity mentions as described above for BM25. The following input sequence is fed into BERT for each candidate and mention pair:

[CLS] expanded mention (m)[SEP]c_i[SEP]$syn_i(c_i)$[SEP]...$syn_n(c_i)$...[SEP]$is(c_i)$[SEP]

Here, syn_i refers to any synonym of the candidate concept c_i and is refers to the RadLex class to which the candidate concept (c_i) belongs. The order of the synonyms is random. The main intention behind using the synonyms is that they provide more variation of the candidate concepts and the is provides more information about the candidate concept's class obtained from the 'Is-A' attribute in RadLex. The final hidden vector corresponding to the [CLS] token in the input sequence is further fed into a single layer network to obtain the estimated probability of how likely the candidate concept is the normalized one. The probabilities corresponding to all the candidate concepts are then used for ranking. Note that the probability score calculated for a particular candidate is independent of the other candidates generated for an entity mention.

5.3.3 BERT as span detector

Alternatively, we formulate the normalization problem similar to the BERT framework for a question answering task. Given an expanded entity mention and its corresponding list of RadLex candidate concepts that is represented as a text sequence, the task is to identify the span of the normalized concept from the candidate list. The second part in the input sequence (the one followed by the first [SEP]) is constructed by joining all the candidate concept names separated by comma. The candidate concepts are placed in an arbitrary order to form this sequence. The final input sequence corresponding to an entity mention and its candidates is represented as follows:

[CLS] expanded mention (m)[SEP]c_1, c_2, ..., c_n[SEP]

The scoring mechanism of a candidate span from the sequence of candidate RadLex concepts is

same as the implementation in the original BERT paper (Devlin et al., 2019). The highest scoring span is identified as the normalized concept for a given mention.

5.4 Experimental Settings and Evaluation

For both BERT-based normalization methods (re-ranker and span detector), we use the $\text{BERT}_{\text{LARGE}}$ model by initializing the model parameters obtained after pre-training BERT on MIMIC-III clinical notes for 320, 000 steps (Si et al., 2019). We fine-tune $\text{BERT}_{\text{LARGE}}$ on our annotated dataset for normalizing the radiological entities. The number of epochs for fine-tuning is decided based on the accuracy of the models on the validation sets. The number of epochs is chosen as 4 for both the normalization models. We use a batch size of 8 for the BERT-based re-ranker model while the batch size of the span detector model is set at 10. We use the cased version of the models. We fine-tune the re-ranker model with a learning rate of 1e-6 and the span detector model with a rate of 3e-5. For the span detector model, we use a maximum sequence length of 384 and a maximum mention length of 64.

We also utilize $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ models, both pre-trained from clinical notes as mentioned above for automatically detecting the entity spans from the report text. We use the cased version of the models, fine-tune the sequence labeling task for 4 epochs with a learning rate of 2e-5 and maximum sequence length of 128. The batch size used for $\text{BERT}_{\text{BASE}}$ is 24 and $\text{BERT}_{\text{LARGE}}$ is 8.

We evaluate our proposed normalization methods - BERT-based re-ranker and the BERT-based span detector by performing 10-fold cross validation (CV) on our annotated radiology normalization corpus. We create the folds by splitting the corpus at the report level such that the training, validation, and test splits are divided in the ratio of 80-10-10% respectively. For comparison, we evaluate the predictions of BM25 by averaging the results obtained on the same test folds used for the BERT-

based models. Since the focus of this study is on normalizing the various radiological entities to RadLex concepts and not on joint prediction of entity mention spans and their normalized concepts, we conduct all our normalization experiments considering the gold entity mentions. In our experiments, any RID-less (RadLex ID-less) entity mention is represented using a special token-'XXXXX'.

We report the average accuracy of the models using the same fold settings. For BM25 and the BERT-based re-ranker, an exact match between the first ranked concept and the gold annotated concept for an entity mention is considered as a correct prediction. In order to handle cases where no candidates are retrieved by BM25 for a given entity mention, we adjust the performance metric (accuracy) by considering only those as correct predictions when their corresponding gold annotated normalized concept is tagged as 'unlinkable' or 'RID-less'.

For the BERT span detector model, we take into account an exact match between the predicted span and the gold annotated RadLex concept in the test sets to qualify a prediction as correct. Note that this model can predict any span from the text representing the sequence of comma-separated candidate concepts. Taking this into account, we evaluate the performance of this model in three ways. First, we evaluate using the original predicted text span. In this version, if more than one candidate concept is captured in the predicted span, the prediction is considered incorrect. Second, we employ post-processing of the predicted spans such that if a span contains more than one concept (indicated by comma), we perform an exact match only between the first concept (concept appearing to the left of the first comma in the predicted span) and the gold normalized concept. Third, we conduct a similar evaluation considering the last concept in the predicted span. We further report the average F1-measure of the 10-fold CV on our annotated dataset for detecting the boundaries of the entity mentions given a report sentence to the entity span detection model.

Table 5.2: 10-fold CV results for detecting the spans of entity mentions. Both $\text{BERT}_{\text{BASE}}$ and $\text{BERT}_{\text{LARGE}}$ models are pre-trained on MIMIC-III clinical notes.

Model	Precision(%)	Recall (%)	F1
$\text{BERT}_{\text{BASE}}$	65.27	73.64	69.14
$\text{BERT}_{\text{LARGE}}$	72.72	79.64	75.93

5.5 Results

The average performance measures of the BERT-based entity span detection system used as a sequence labeler (described in Section 5.2) over 10-fold CV are shown in Table 5.2. We notice that the average F1 is increased by around 6.8 points when the $\text{BERT}_{\text{LARGE}}$ model is used.

We first report the recall of BM25 for candidate generation. Recall here refers to the percentage of entity mentions for which the list of candidate concepts contains the gold normalized RadLex concept. The recall of BM25 as well as its accuracy in predicting the normalized concepts for 10 and 25 candidates is shown in Table 5.3. The average accuracies of 10-fold CV for the BERT-based methods in predicting the correct normalized RadLex concept when provided the 10 candidates generated by BM25 is shown in Table 5.4. We note that the normalization performance is the highest (accuracy of 77.72%) for the BERT-based span detector model when compared to both BM25 and BERT-based re-ranking models. The performance is further improved by 0.7% when either the first or the last concept in the predicted span (as predicted by the BERT span detector from the sequence of 10 candidate concepts) is considered as the normalized concept.

Table 5.3: BM25 results in predicting the normalized concepts using 10 and 25 candidate concepts.

Metric	10 candidates	25 candidates
Recall (%)	88.44	89.72
Accuracy (%)	76.10	76.10

Table 5.4: 10-fold CV results of the proposed BERT-based methods using 10 candidate concepts retrieved by BM25.

Method	Average accuracy (%)
BERT re-ranker	76.50
BERT span detector (using original predictions)	77.72
BERT span detector (first concept in the predicted span as the normalized concept)	78.43
BERT span detector (last concept in the predicted span as the normalized concept)	78.44

5.6 Discussion

We create a manually annotated corpus covering a broad range of radiology entity types that are usually of interest for information extraction research. To our knowledge, this is the first work in developing a corpus targeted toward radiology entity normalization. We propose methods for normalizing these entities to an existing lexicon-RadLex.

We also examine the performance of a sequence labeler, based on BERT, for identifying the spans of the entity mentions from the reports. The performance of the span detection system is decent (average F1-score of 75.93). The moderate performance may be attributed to the incorrect predictions for detecting the composite entities (e.g., "*right upper lobe*"). Note that the focus of this work is radiology entity normalization, hence we aim to further improve the performance of the entity span detection and develop joint learning methods for predicting the entity spans as well as mapping them to RadLex concepts simultaneously.

Most of the annotation-related challenges are related to the requirement of domain knowledge. For example - "*Lower pole of the right kidney*" usually refers to "*Inferior pole of right kidney*" and "*Temporal horns*" denotes "*temporal horn of lateral ventricle*". Besides being a time-consuming process, another generic challenge related to constructing a normalization corpus is the ambiguity involved in annotating composite entity mentions such as "*right lung apex*". Another difficulty in the annotation involves dealing with the inconsistencies in the RadLex lexicon. For example, the expression "*upper lobe of right lung*" is present in RadLex whereas "*middle lobe of right lung*" is not although both are anatomical expressions at the same hierarchical level. The closest term available in RadLex for the middle lobe is "*middle lobe of lung*". Also, there are cases where certain entity mentions when expressed using more general terms such as "*sulci*" do not appear in RadLex, although their specific types such as "*hypothalamic sulcus*" and "*cardiophrenic sulcus*" are present in RadLex.

Our proposed normalization methods achieve satisfactory performance with the highest average accuracy of 78.44%. However, we aim to further evaluate the performance of the proposed methods by augmenting the annotated corpus in the future. A brief analysis of the model outputs suggests that the BERT-based models make correct predictions for uncertainty or hedging-related entity mentions such as "*could indicate*" that are incorrectly predicted by BM25. Moreover, the BERT span detector model performs better in predicting the normalized concepts for plural entity mentions compared to BERT re-ranker. For instance, BERT span detector predicts "*lungs*" as the normalized concept for the mention-"*lungs*", whereas BERT re-ranker model predicts "*lung*" as the mapped concept. One of the reasons for the moderate performance improvement of the BERT-based models over BM25 may be that our annotated corpus mostly contains different variations of radiological terms unlike social media posts where there are more variations of natural language expressions.

Future work can explore an exhaustive set of ablation experiments mainly for the BERT-based

re-ranker model utilizing various combination of RadLex knowledge. Additional techniques can be used for expanding the radiological entity mentions, particularly the ones related to clinical findings (e.g., "*Scrotal herniation of bowel*"), by leveraging the co-occurring entity information from sources like Wikipedia and medical abstracts.

Among the 151 entity mentions for which a suitable RadLex concept is not found, some of the most common clinical finding and imaging observation-related entities include - "*respiratory distress*", "*hyaline membrane disease*", "*bowel gas pattern*", "*cabg*", "*portal venous gas*", "*mucosal thickening*", "*hydropneumothorax*", "*ventricular prominence*", "*v-fib arrest*", "*urinary incontinence*", "*reexpansion*", "*pleural margins*", "*neonatal pneumonia*", "*lyme disease*", "*intubation*", "*guaiac positive stools*", "*gonadal shielding*", "*fetal lung liquid*", "*dyspnea*", "*claustrophobia*", "*cardiomegaly*", "*anxiety*", "*altered mental status*", "*afib*", and "*aeration*". This can serve as a potential list of terms to expand RadLex in the future.

6

Generalizability of Rad-SpatialNet: Extending to Ophthalmology Domain

Ophthalmology notes contain important clinical information about a patient's eye findings. These findings are documented based on interpretations from imaging examinations (e.g., fundus examination), complications or outcomes associated with surgeries (e.g., cataract surgery), and experiences or symptoms shared by patients. Such findings are oftentimes described along with their exact eye locations as well as other contextual information such as their timing and status. Thus, ophthalmology notes comprise of spatial relations between eye findings and their corresponding locations, and these findings are further described using different spatial characteristics such as laterality and size. Although there has been recent advancements in using natural language processing (NLP) methods in the ophthalmology domain, they are mainly targeted for specific ocular conditions.

Some work leveraged electronic health record text data to identify conditions such as glaucoma (Wang et al., 2022), herpes zoster ophthalmicus (Zheng et al., 2019), and exfoliation syndrome (Stein et al., 2019), while another set of work extracted quantitative measures particularly related to visual acuity (Baughman et al., 2017; Mbagwu et al., 2016) and microbial keratitis (Woodward et al., 2021). Here, we aim to extract more comprehensive information related to all eye findings, covering both spatial and contextual, from the ophthalmology notes. Besides automated screening and diagnosis of various ocular conditions, identifying such detailed information can aid in applications such as automated monitoring of eye findings or diseases and cohort retrieval for retrospective epidemiological studies. For this, we propose to extend our existing radiology spatial representation schema–Rad-SpatialNet (described in Section 3.2.1 in Chapter 3) to the ophthalmology domain. We refer to this as the Eye-SpatialNet schema. We annotate a total of 600 ophthalmology notes following Eye-SpatialNet. Finally, we apply an advanced deep learning-based method to automatically identify the spatial and contextual information from the notes.

Ophthalmologists use spatial language to describe findings interpreted from imaging techniques. For example, in the sentence – "*OCT of the retinal nerve fiber layer shows normal thickness in both eyes.*", both eyes have been described using the finding *normal thickness* as interpreted from an Optical Coherence Tomography examination. Here, *thickness* is spatially associated to *eyes* through the preposition *in*, where *normal* describes the status of thickness and *both* describes the laterality. Similarly, symptoms presented by patients are also documented using spatial relations. In the sentence – "*She presented in [DATE] with weakness and numbness of her right eye as well as pain and vision loss in the left eye consistent with optic neuritis.*", the findings *weakness* and *numbness* are spatially related to *right eye* through the preposition *of*, whereas *pain* and *vision* are linked to *left eye* through *in*. Additionally, we note that the ophthalmologist also reports the potential diagnosis inferred from these findings, i.e., *optic neuritis*. *[DATE]* denotes the timing associated with the findings. Sometimes, eye procedures and drugs are also associated with anatomical locations and thus are

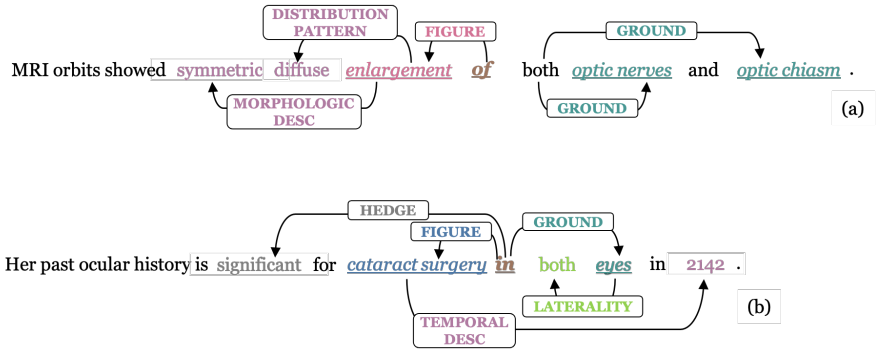

Figure 6.1: Example sentences from ophthalmology notes showing some of the spatial frame elements covered in the Eye-SpatialNet schema. The underlined and italicized texts denote the lexical units of the frames.

spatially-grounded. We capture all these important information in our Eye-SpatialNet schema.

The Eye-SpatialNet schema is based on frame semantics, where a lexical unit (LU) represents the word that invokes a frame and the participants of a frame form the frame elements (FEs). The spatial prepositions (e.g., *in*) and verbs (e.g., *reveals*) constitute the lexical units whereas the associated findings (e.g., *weakness*), the locations (e.g., *eye*), diagnosis (e.g. *optic neuritis*), and the various spatial and other descriptors (e.g., *left*, *normal*) constitute the frame elements. The spatial prepositions and verbs are also referred to as spatial triggers in this work. Following this schema, we create a manually-annotated corpus of 600 ophthalmology notes to represent important spatial information of clinical significance. Two sample examples from our ophthalmology corpus is illustrated in Figure 6.1. Note that for (a), 'Figure' and 'Ground' are the spatial frame elements of the frame evoked by the spatial trigger *of*, whereas 'Morphologic descriptor' and 'Distribution pattern' are the spatial frame elements of the frame evoked by the finding *enlargement*. 'Figure' usually refers to an entity whose location is described through a spatial trigger whereas 'Ground' denotes the actual anatomical location. In the second example (b), *cataract surgery* is spatially linked to *eyes* where *cataract surgery*

acts as the 'Figure' element of the frame evoked by the spatial trigger *in* and *2142* (year altered for de-identification) is a descriptive frame element of the frame instantiated by the procedure *cataract surgery*. To our knowledge, this is the first work to develop an annotated dataset with comprehensive representation schema for identifying detailed information from ophthalmology notes.

For automatic extraction of the spatial information, we adopt a two-turn question answering framework (Li et al., 2019) based on a transformer language model, BERT (Devlin et al., 2019). This is inspired by previous studies demonstrating the effectiveness of framing various information extraction tasks such as named entity recognition (Li et al., 2020b), relation extraction (Levy et al., 2017), and event extraction Liu et al. (2020) as question answering (QA) by harnessing the well-developed machine reading comprehension models. Further, some studies (Li et al., 2019, 2020a; Wang et al., 2020) investigated the formulation of relation and event extraction tasks as multi-turn QA both in the general and biomedical domain. In this chapter, we apply a two-turn QA method, similar to the one proposed for radiology domain (described in Section 4.2.1 in Chapter 4), to extract the spatial and descriptive frame elements from ophthalmology notes. In this, we extract the spatial triggers and the main entities (e.g., eye finding, anatomical location) in the first turn and subsequently extract all the spatial (e.g., laterality) and descriptive (temporal descriptor or the timing of a finding) frame elements in the second turn.

6.1 Eye-SpatialNet Schema Description

Our annotation schema is largely adopted from an existing frame-based spatial representation schema – Rad-SpatialNet (described in Section 3.2.1 in Chapter 3). The spatial language encoded in the ophthalmology notes are different from those in radiology reports. We represent the information in a way that can accurately capture ophthalmology-specific spatial meanings from the note text.

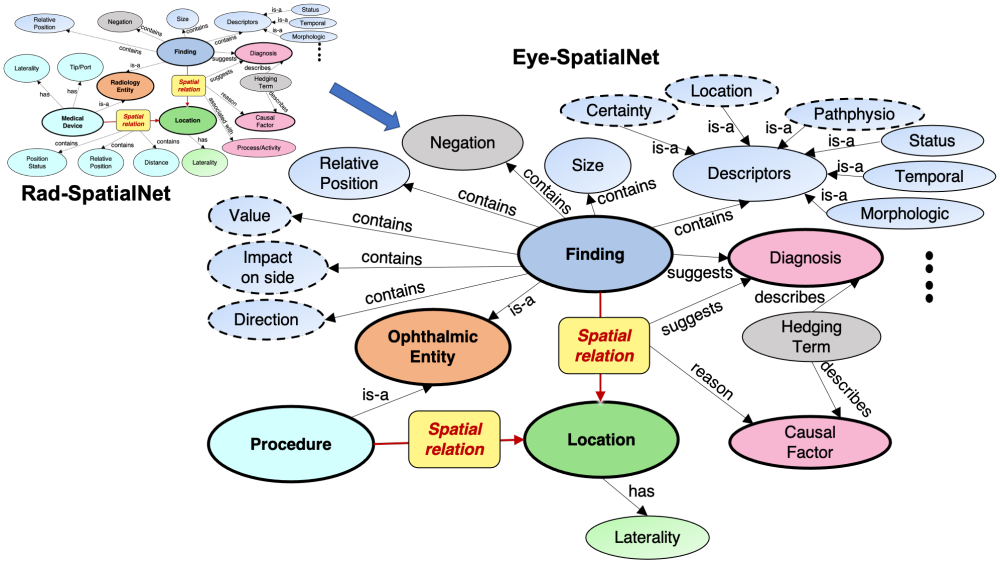

Figure 6.2: Eye-SpatialNet schema. The dashed circles indicate the newly added frame elements.

For this, in this schema, we incorporate specific spatial and descriptive frame elements or relations besides the common ones proposed in Rad-SpatialNet. The entity types included are spatial trigger, finding, anatomy, device, location descriptor, other descriptor, assertion, quantity, drug, and procedure. The spatial and descriptive frame elements are mostly similar to the ones described in Table 4.8 in Chapter 4. Additionally, we include the frame elements–medication, impact on side, pathophysiologic descriptor, direction, associated diagnosis, specific location descriptor, certainty descriptor, and value. Frame elements are either connected to the spatial trigger terms or the main clinical entities such as findings and anatomies. We describe the newly added ophthalmology-specific frame elements in the following subsections. The schema is illustrated in Figure 6.2.

Ophthalmologists often document detailed contextual information while describing the findings. To cover this, we add some ophthalmology-specific frame elements that we describe in the following subsections.

6.1.1 New spatial frame elements

We add three new spatial frame elements related to findings, namely, exact location descriptor, impact on side, and direction. For the exact location descriptor, let us consider the example below.

> *She was found to have 20/25 vision OD and CF vision OS with mild **disc** edema **in** the left eye.*

Here we see that there is a spatial relation between *mild disc edema* and *left eye* connected through the spatial trigger *in*. As per the Eye-SpatialNet schema, *edema* has the spatial role of a 'Figure' and its corresponding location *eye* acts as the 'Ground'. Moreover, we notice that *edema* has been described through a location descriptor *disc* besides the status descriptor *mild*.

Sometimes, a finding that has been detected in both sides (left and right) is described with different severity based on laterality or side.

> *External examination reveals a right relative proptosis with bilateral lid retraction **right greater than left**.*

In this example, *retraction* is the finding that is more pronounced in the right eyelid than the left eyelid. Moreover, *retraction* is described using laterality *bilateral* and location descriptor *lid*.

A finding's direction is also documented in the notes.

> *She reports that her right eye deviated **outward**, and she had difficulty walking with poor coordination.*

Here, *outward* is used to describe the direction of right eye deviation.

All these three frame elements–location descriptor, impact on side, and direction are associated with describing the detailed spatial aspects of a finding and, therefore, we include these elements in our representation schema.

6.1.2 New descriptive frame elements

We add four descriptive frame elements related to findings, namely, certainty descriptor, associated diagnosis, pathophysiologic descriptor, and value. Consider the example below.

> *His past ocular history is **significant** for optic neuritis and right optic atrophy.*

In this sentence, the term *significant* is used to describe the certainty of both *optic neuritis* and *optic atrophy* findings.

Oftentimes, some findings are described along with their associated diagnoses. In the following example, *occlusions* is linked to *Susac Syndrome*.

> *At this time the exact cause is unknown, however, with multiple retinal branch artery occlusions bilaterally one must entertain the diagnosis of **Susac Syndrome**.*

Note that this 'Associated Diagnosis' frame element is different from the 'Diagnosis' frame element proposed in Rad-SpatialNet. The 'Diagnosis' element is linked to a spatial trigger, whereas 'Associated Diagnosis' element is linked to an eye finding (e.g., *Susac Syndrome* in the sentence above). In "*She did show me video of her episodes of upturning of the eyes which appears consistent with **oculogyric crisis**.*", *oculogyric crisis* acts as the 'Diagnosis' element of the spatial frame instantiated by the spatial trigger *of* connecting *upturning* and *eyes*.

We also include pathophysiologic descriptor of a finding in the schema. For example, in "*She is seen in follow up for her left sided headache and retroorbital pain in the setting of presumed **autoimmune** retinopathy.*", *autoimmune* is the pathophysiologic descriptor associated with the finding *retinopathy*.

Ophthalmology notes also contain information about visual acuity scores and other eye-related measurements. We present two examples below.

1. *On his examination he found her to have **20/20** vision OD and **20/30** vision OS with a left RAPD.*

2. *INTRAOCULAR MEASUREMENT: Method: Applanation Right Eye:* ***15***

3. *LEFT EYE: Media: hazy view Cup/Disc Ratio:* ***0.4***

Note that in the first example, the first *vision* occurrence has a visual acuity score of *20/20* in the oculus dextrus (*OD*) or the right eye, while the second vision has a score of *20/30* in the oculus sinister (*OS*) or the left eye. Thus, the first *vision* finding is linked to *20/20* and the second *vision* is linked to *20/20* through the 'Value' relation or frame element. And *OD* is the laterality of the first *vision* while *OS* is the laterality of the second one. In the second example, *INTRAOCULAR MEASUREMENT* is linked to *15* through the 'Value' element, whereas the third example shows that the finding *Cup/Disc Ratio* is associated with its corresponding value *0.4*. Therefore, we capture all the important eye measurements in our schema.

Apart from above additions, this schema covers temporal information of findings that are expressed using a variety of phrases unlike the temporal descriptors of radiological findings annotated in Rad-SpatialNet (described in Section 3.2.1 in Chapter 3). These expressions include *one and a half to two years*, *> 8 years*, *next 3-4 months*, *within 1-2 months post-operatively*, *over the next few days*, and *early in the mornings*. This schema covers lateralities that are specific to ophthalmology such as *OS*, *OD*, and *OU*, besides the common ones such as *left*, *right*, and *bilateral*.

6.2 Dataset Annotation

We use a set of 600 notes for annotating the important ophthalmic entities and spatial relations. These notes are collected from the Robert Cizik Eye Clinic at McGovern Medical School at Houston. The notes contain information about a patient's history, detailed description of patients' experiences with their vision, interpretations of eye imaging examinations, information about past surgeries and their outcomes and complications, and associated neurological symptoms. We use the BRAT tool (Stenetorp et al., 2012a) for annotation.

Table 6.1: Basic statistics. Avg - Average.

Item	Value
Avg. note length (in tokens)	470.61
Avg. sentence length (in tokens)	20.34
No. of unique spatial triggers	49

6.2.1 Annotation Statistics

Each note was annotated by two annotators having medical background (one optometrist, one MD) and the annotations were reconciled iteratively through discussions. The overall F1 agreements are reported for annotating the main entities, the spatial and descriptive frame elements. We show the statistics of our annotated dataset as well as the inter-annotator agreement measures in Tables 6.1, 6.2, 6.3, and 6.4. The average sentence length (20.34) of the ophthalmology notes is slightly higher than that of the radiology reports Rad-SpatialNet dataset (refer Table 3.6). The spatial triggers in our annotated ophthalmology dataset contains spatial prepositions such as *in*, *behind*, and *within* as well as verbs such as *appear*, *reveals*, and *are*. The top three frequent trigger terms are *in*, *of*, and *are*. Among the entity types, the agreement is low (F1: 0.35) for Procedure as this involves identifying different eye surgery procedures or therapies that are often expressed in their abbreviated forms (e.g., LPI for Laser Peripheral Iridotomy and PRP for Pan-retinal photocoagulation). Among the spatial and descriptive frame elements, Diagnosis, Size, and Associated Diagnosis have low F1 agreements of 0.28, 0.30, and 0.23, respectively, as it is oftentimes difficult to correctly interpret the potential diagnoses terms and sizes of different eye entities. Diagnoses terms are difficult to differentiate from the Finding terms and are rather annotated as Findings. Another general challenge in the annotation process involved separating the eye-related findings from the neuroradiological findings as oftentimes the interpretations of brain images are embedded in the ophthalmology notes.

Table 6.2: Spatial frame elements.

Frame element	Frequency	F1 agreement
Figure	2261	0.77
Ground	2094	0.89
Hedge	397	0.69
Diagnosis	18	0.28
Relative Position	132	0.59
Reason	7	0.77
Medication	18	0.64
Morphologic	45	0.44
Size Desc	43	0.56
Distribution Pattern	83	0.29
Composition	36	0.59
Laterality	3464	0.78
Size	48	0.30
Impact on Side	97	0.75
Direction	85	0.56
Specific location	1636	0.72

6.3 Information Extraction as Question Answering

6.3.1 System overview

We frame the task of spatial information extraction (IE) from ophthalmology notes as two-turn question answering (QA). This formulation (both single and multi turn QA) has proven to perform well for various general and biomedical domain IE tasks. Table 4.14 in Chapter 4 has also demonstrated the improved performance of a two-turn QA framework over a more standard sequence labeling-based method to extract detailed information from radiology text. Inspired by these findings, we adopt a similar two-turn QA approach to identify the spatial triggers, the main ophthalmic entities, and their corresponding spatial and descriptive frame elements. This framework is suitable for IE scenarios where identification of relations or frame elements are dependent

Table 6.3: Main entities.

Entity type	Frequency	F1 agreement
Spatial trigger	1715	0.91
Finding	7308	0.80
Anatomy	2424	0.88
Device	14	0.90
Drug	22	0.60
Procedure	182	0.35
Other descriptor	9782	0.79
Quantity	366	0.88
Assertion	1616	0.70
Location descriptor	132	0.60

Table 6.4: Descriptive frame elements.

Frame element	Frequency	F1 agreement
Status	3051	0.59
Quantity	101	0.56
Temporal	1066	0.45
Negation	921	0.55
Pathphysio	75	0.60
Certainty	298	0.49
Associated Diagnosis	72	0.23
Value	318	0.83

on extracting the target entities or lexical units of the frames (i.e., spatial triggers and ophthalmic entities). In this, the aim is to query a machine reading comprehension (MRC) model for returning answers given a query and the context passage (ophthalmology note text). The MRC system is based on the pre-trained language model BERT (Devlin et al., 2019) where we follow the standard BERT input format by combining the query and the note text. The system allows for multiple answer extraction against a query, which is suitable for our schema as there can be multiple frame elements of the same type that are linked to a particular entity (spatial trigger or other ophthalmic entity).

The MRC framework involving two BERT models for the two turns are adopted from a previous work (Li et al., 2019).

6.3.2 Query generation

We construct queries for the newly added entities and frame elements in Eye-SpatialNet. We adopt the same query templates for both target entity and element extraction as used in Table 4.10 in Chapter 4. Queries for the first turn incorporate the entity types whereas queries for the second turn include information about the frame elements and the associated main entity that is extracted in the first turn. In this work, we use the $\text{Query}_{\text{find + desc}}$ variant to extract the frame elements in the second turn. The idea is to make the query more informative through incorporation of domain knowledge by adding a description of the particular frame element of interest at the beginning of a query. The following is an example query to extract 'ImpactOnSide' spatial element.

> *ImpactOnSide refers to which eye side is more impacted. Examples include right greater than left, smaller than left, and worse in the left eye. find all* ***descriptor*** *entities in the context that have a* ***impact on side*** *relationship with* ***clinical finding*** *entity optic neuropathy.*

Here, we see that the query includes description about ImpactOnSide as well as the finding entity (i.e., *optic neuropathy*) that is identified in the previous turn. If no answer is retrieved from the MRC system, this means there is no such entity of type 'Descriptor' in the note text that captures information about which eye side is more or less affected by *optic neuropathy*. The descriptions used to form the queries for all new frame elements are shown in Table 6.5.

6.4 Experimental Settings and Evaluation

We randomly split our annotated ophthalmology dataset of 600 notes such that 450 notes are used for training, 50 for development, and 100 for testing. We use clinical BERT $_{\text{LARGE}}$ model

Table 6.5: Descriptions used in the queries to extract additional frame elements.

Frame element	Description
Medication	Medication refers to a drug or solution that has been administered or applied to any eye location.
ImpactOnSide	ImpactOnSide refers to which eye side is more impacted. Examples include right greater than left, smaller than left, and worse in the left eye.
PathphysioDesc	Pathophysiologic descriptor refers to the functional changes that accompany a disease. Examples include autoimmune and physiologic.
Direction	Direction indicates direction of a finding. Examples include outward and to the right.
AssocDiag	Associated diagnosis refers to the clinical condition or disease associated with a finding. This usually appears after phrases such as associated with and secondary to.
LocationDesc	Location descriptor refers to the exact location of a finding. Examples include retrooorbital and optic disc.
CertaintyDesc	Certainty descriptor refers to uncertainty phrases describing a finding. Examples include significant and consistent with.
Value	Value refers to a visual acuity score or any measurement or ratio. Examples include 20/20, 20/40, 16, and 0.8.

that is pre-trained on MIMIC-III clinical notes for 300K steps (Si et al., 2019) as it performed better on our radiology reports dataset. We fine-tune $\text{BERT}_{\text{LARGE}}$-MIMIC (cased version) on our Eye-SpatialNet dataset for 10 epochs and use the same hyperparameter settings as reported in Table 4.13 in Chapter 4 for the Rad-SpatialNet dataset. We evaluate the performance metrics - precision, recall, and F1 score and report the results on the test set of 100 notes. We consider exact matches of the entity and frame element spans against the annotated spans for evaluation.

Further, to leverage an already available language model that is fine-tuned on the radiology reports dataset for the task of spatial information extraction, we evaluate any prospective benefits of

Table 6.6: Target entity extraction results using $\text{BERT}_{\text{LARGE}}$-MIMIC two-turn QA method. desc - Descriptor.

Entity	Precision(%)	Recall (%)	F1
Spatial trigger	86.86	91.89	89.31
Finding	75.71	83.41	79.37
Anatomy	85.37	85.15	85.26
Location desc	30.77	40.00	34.78
Other desc	76.57	83.04	79.67
Assertion	81.78	89.80	85.60
Quantity	82.89	82.89	82.89
Procedure	56.67	53.12	54.84

transfer learning through sequential fine-tuning, that is, by first fine-tuning the model on radiology reports followed by fine-tuning on ophthalmology notes. The radiology fine-tuning was performed on 288 reports and we further fine-tune on 450 ophthalmology notes. For this, we use the BERT LARGE-MIMIC sequence labeling model fine-tuned on radiology reports (described in Section 4.1.4 in Chapter 4). Note that we use the gold spatial triggers for this experiment to extract the elements that are connected to the triggers (and not the main ophthalmic entities). Using predicted triggers would provide a more realistic evaluation, but that is not the focus of this experiment. We evaluate the performance on the main spatial frame elements that are common between the two domains on the 100 test ophthalmology notes. For fine-tuning the sequence labeling model on the ophthalmology data, we set the maximum sequence length at 128, learning rate at $2e-5$, and number of training epochs at 4.

6.5 Results

The performance measures for extracting the main ophthalmic entities in the first turn from 100 ophthalmology test notes are reported in Table 6.6. The results are promising for the common

Table 6.7: Frame element extraction results using $\text{BERT}_{\text{LARGE}}$-MIMIC two-turn QA method. sptr - Spatial trigger. Desc - Descriptive.

	Frame Elements	Precision(%)	Recall (%)	F1
Spatial(sptr)	Figure	75.29	74.43	74.86
	Ground	85.89	91.21	88.47
	Hedge	89.47	86.44	87.93
	Relative Position	30.43	70.00	42.42
	Medication	50.00	100	66.67
Spatial(entity)	Laterality	80.59	83.15	81.85
	Distribution Pattern	47.37	64.29	54.55
	SizeDesc	60.00	42.86	50.00
	LocationDesc	69.26	76.21	72.57
	ImpactOnSide	72.73	84.21	78.05
	Direction	57.14	66.67	61.54
	Size	28.57	10.00	14.81
Desc(entity)	Status	70.11	70.93	70.52
	Quantity	63.64	43.75	51.85
	Temporal	53.33	43.78	48.09
	Negation	77.60	82.32	79.89
	Certainty	60.26	64.38	62.25
	Pathphysio	47.06	53.33	50.00
	Value	81.63	68.97	74.77

entities including 'Spatial trigger', 'Finding', and 'Anatomy' with F1 scores of 89.31, 79.37, and 85.26, respectively, while they are low for 'Location descriptor' and 'Procedure'. Note that the entities 'Drug' and 'Device' occur very infrequently in the dataset (with only 2 and 1 occurrences in the test set) and the performance measures are zero.

We show the results for extracting the spatial and descriptive frame elements in the second turn in Table 6.7. We see that the model performs well for common frame elements such as 'Ground', 'Hedge', 'Laterality', 'ImpactOnSide', and 'Negation'. The performance measures are particularly low for 'Relative Position', 'Size', and 'Temporal Desc'. This may be because of the wide variation

Table 6.8: F1 measures for different fine-tuning variations using $\text{BERT}_{\text{LARGE}}$-MIMIC sequence labeling method on 100 test ophthalmology notes. Oph - Fine-tuning only on Ophthalmology, Rad-Oph - Fine-tuning on Radiology followed by Ophthalmology, Rad - Fine-tuning only on Radiology.

Frame Element	Oph	Rad-Oph	Rad
Figure	78.76	80.88	51.29
Ground	95.38	95.19	91.95
Hedge	82.64	89.08	0
Relative Position	60.87	57.14	43.48

in the phrases used to express the sizes and temporalities of findings. Moreover, there are only 3, 1, and 7 instances of the frame elements 'Diagnosis', 'Reason', and 'Composition Desc', respectively in the test set with no occurrence for 'Morphologic Desc'. Although the element 'AssocDiag' occurs 21 times, the performance values are zero for this element. The reason could be that this is often a difficult task (even for humans) to differentiate these terms from findings.

The results of transfer learning experiment from radiology to ophthalmology domain is shown in Table 6.8. We see the F1 scores improve from 78.76 to 80.88 for 'Figure' and 82.64 to 89.08 for 'Hedge' when we use a model fine-tuned on radiology reports to further fine-tune on our ophthalmology corpus. We also note that the F1 measure for the 'Ground' element is 91.95 without the requirement of any fine-tuning on ophthalmology data. The results are zero for 'Diagosis' and 'Reason' as they are too infrequent in the dataset as stated above.

6.6 Discussion

We present a new dataset of 600 ophthalmology notes annotated with detailed spatial and contextual information. Although a few studies worked on identifying a certain set of entities from clinical notes, they are mostly focused toward visual acuity and features of microbial keratitis (Baughman et al., 2017; Mbagwu et al., 2016; Woodward et al., 2021). Our work is an initial effort

in building a schema that captures more detailed information from the notes that can potentially be used in various useful ophthalmology-related applications and research studies.

Most of the entities and frame elements used in encoding spatial language in the ophthalmology notes are adopted from our previously proposed Rad-SpatialNet schema built for radiology. This indicates the generalizability of the schema in that it captures most of the common and important spatial information usually encountered in clinical text. We incorporate additional frame elements for two reasons. First, to cover more detailed information about the findings that were not present in Rad-SpatialNet such as capturing implicit spatial relations through the 'Location Desc' frame element (e.g., scenarios where a spatial relation exists but a spatial trigger term is not present in the sentence). Second, to include ophthalmology-specific spatially-grounded entities (e.g., 'Procedure') and elements that are of interest to ophthalmology researchers (e.g., visual acuity and other important eye measurements through the 'Value' frame element). The results in Tables 6.6 and 6.7 show that the two-turn QA approach achieves satisfactory performance in identifying different entities and frame elements and are comparable to the results on the radiology report dataset (shown in Table 4.14 in Chapter 4). We achieve this without any modification of the query templates and the frame element descriptions (that are used to form the queries) for those elements that also exist in Rad-SpatialNet. This also indicates that the method is adaptable and generalizable enough to work satisfactorily well for frequent entity types and frame elements across medical domains (although the language style and the vocabulary differ a lot between radiology reports and ophthalmology notes).

To examine the effect of transfer learning from a different medical domain, our experiment with the sequence labeling model in Table 6.8 indicates that transfer learning holds potential in improving the performance for some frame elements, however, a more thorough evaluation covering all other elements is required to understand its real benefits. This includes experimenting with a small number of ophthalmology notes in the fine-tuning process, as often only a limited amount of

labeled data is available in a new domain. Interestingly, although the two-turn QA approach works well both for ophthalmology and radiology domains, our initial experiments with sequential fine-tuning did not yield good results using the QA approach. We leave this to future work where we plan to investigate this further and evaluate the less explored domain adaptation techniques such as the adaptive off-the-shelf approach proposed in Laparra et al. (2020).

To handle less frequent entities and frame elements better as well as to further improve the QA model's performance, the dataset can be augmented in the future by automatically generating a large weakly labeled ophthalmology dataset using domain-specific rules, a technique that has been validated to be useful by many recent studies in the medical domain (Smit et al., 2020; Fries et al., 2021b). Apart from reducing the annotation effort, this can particularly be useful for elements such as 'Size' and 'Value' that usually follow a set of patterns based on domain. For example, '4–> 3mm' is used to express pupil size in an ophthalmology note whereas '2.1 x 3.4 x 2.0 cm' denotes a tumor size in a radiology report. Finally, cross validation can be incorporated in a later work for a more exhaustive evaluation on this proposed dataset.

7

Weak Supervision for Spatial Information Extraction

Most information extraction studies in the clinical domain utilize exclusively supervised learning approaches. Such approaches rely on human-annotated reports that are not only tedious, time-consuming, and expensive, but also require extensive domain knowledge. Thus, it is difficult to achieve the scale of manual annotation for complex and fine-grained information. Moreover, manual annotations are often not generalizable across institutions because of limited coverage of language variations and/or reporting style. Meanwhile, deep learning-based supervised methods often demand large amounts of annotated training data to achieve substantial performance improvement over alternatives like rule-based methods. Recent research (Ratner et al., 2020; Fries et al., 2017; Shang et al., 2018; Safranchik et al., 2020) has proposed weak supervision to address the above issues

by programmatically creating very large training corpora with imperfect labels which have the potential to outperform fully supervised approaches. Such approaches have been applied in clinical natural language processing (NLP) tasks such as medical entity classification (Fries et al., 2021a), concept normalization (Pattisapu et al., 2020), relation classification (Peterson et al., 2020; Callahan et al., 2019), and sentence-level classification (Banerjee et al., 2019) for different use cases including patient-centered outcome assessment (Banerjee et al., 2019) and medical device surveillance (Callahan et al., 2019).

One recently explored weak supervision method is data programming (Ratner et al., 2020), which uses multiple supervision sources including patterns, domain rules, and domain-specific ontologies to automatically generate training data. Rules or labeling functions (defined based on domain knowledge from these sources) are applied on unlabeled data, the output of which is used to train a generative model to generate probabilistic training labels, thus obviating the laborious process of constructing human-annotated labels. Moreover, the labeling functions can be easily updated when applied to a different institution's data to incorporate any change in the downstream use case, or to keep in sync with the latest domain knowledge with feedback from subject matter experts. This thereby reduces the manual effort of re-labeling data based on revised annotation guidelines. Inspired by this, we use data programming to automatically construct a distantly labeled corpus of radiology reports following our previously proposed Rad-SpatialNet representation schema (described in Section 3.2.1 in Chapter 3).

This chapter describes our proposed data programming-based weak supervision approach to create a large labeled dataset of radiology reports for spatial relation extraction. We use the Snorkel framework (Ratner et al., 2020) to automatically create the weak relation labels. Our labeling functions are based on the radiology-specific lexicon - RadLex (Langlotz, 2006), regular expressions, language characteristics of report text, and other task-specific heuristics. The generated weak labels are used to train a transformer-based language model, BERT (Devlin et al., 2019). The overall weak supervision

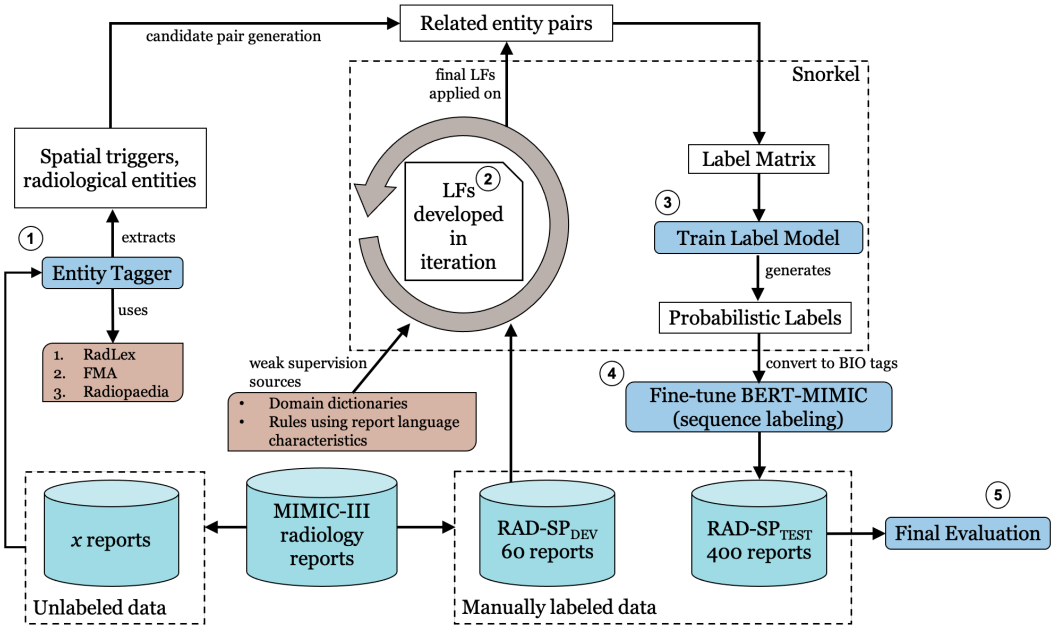

Figure 7.1: Overview of our weak supervision approach for radiology spatial information extraction. LF: Labeling Function. BIO: Beginning, Inside, Outside. RAD-SP$_{\text{DEV}}$: Development set. RAD-SP$_{\text{TEST}}$: Held-out test set. x represents the number of unlabeled reports used for training the Label Model (varies from 500 to 50k).

pipeline is shown in Figure 7.1. To assess our approach, we evaluate BERT that is fine-tuned only using weakly labeled reports. We also evaluate sequential fine-tuning performance (fine-tuning on weak labels followed by gold labels) and compare it with a fully-supervised variant. The evaluations are performed on 400 radiology reports (comprising of chest X-ray, brain MRI, and babygram reports) (more details are described in Section 3.2.2 in Chapter 3).

7.1 Data

We use 400 (358 containing spatial relations) manually labeled MIMIC-III (Johnson et al., 2016) radiology reports (Chest X-ray: 136, Brain MRI: 127, and Babygram: 137) as a held-out

test set for evaluating our weak supervision pipeline. We refer to this as RAD-SP$_{\text{TEST}}$. More details are found in Section 3.2.2 in Chapter 3. We randomly select a total of 50k unlabeled MIMIC reports (with almost equal proportion of the three report types) to train the generative model and subsequently the weakly supervised BERT$_{\text{LARGE}}$-MIMIC model. We manually annotate additional randomly selected 60 MIMIC reports (20 in each of the three categories) for building dictionaries, labeling functions, and hyper-parameter tuning (referred to as RAD-SP$_{\text{DEV}}$).

7.2 Method

We perform the following sequential steps to programmatically create the weak training labels. For this we employ data programming using the Snorkel framework (Ratner et al., 2020).

7.2.1 Candidate Generation

We identify all the candidate {spatial trigger, radiological entity} pairs where the radiological entity acts as a potential spatial FE with respect to the spatial trigger in a sentence. This involves the following steps:

1. Dictionary construction - We curate two dictionaries using pre-existing knowledge sources– 1) Rad-Entity$_{\text{dict}}$: This contains different types of radiological entities such as radiological findings and anatomical entities using RadLex (Langlotz, 2006). For this, all RadLex terms under the parent RadLex classes *Imaging observation* (RID5), *Clinical finding* (RID34785), *Anatomical entity* (RID3), *Medical device* (RID29033), and *Process* (RID39128) are used for constructing a comprehensive vocabulary representing important radiological entities. Additionally, we also add the terms in Foundational Model of Anatomy (FMA) ontology (Rosse & Mejino, 2008) to include more anatomical entities and add radiology-specific

acronyms and their corresponding expansions from Radiopaedia [*]. This results in a total of 153,944 terms. 2) Spatial-Trigger$_{\text{dict}}$: This contains potential phrases denoting spatial relations between finding/device and location. We develop this by combining the spatial triggers annotated in RAD-SP$_{\text{DEV}}$ to a set of hand-built trigger terms.

2. Expanding Rad-Entity$_{\text{dict}}$ - We manually add more finding and anatomy-related terms to Rad-Entity$_{\text{dict}}$ that are encountered in RAD-SP$_{\text{DEV}}$ but are not present in RadLex or FMA (e.g., *effacement*, *volume loss*, *caudate nucleus head*). Based on patterns identified using RAD-SP$_{\text{DEV}}$, we further prepend or append phrases to a set of terms in Rad-Entity$_{\text{dict}}$. Specifically, we prepend in two ways– 1) prepending phrases such as '*area(s) of*', '*region(s) of*' and '*focus/foci of*' to finding-specific terms (e.g., *hypodensity*) and 2) prepending descriptors to certain finding and anatomical entities (e.g., prepend '*petechial*' and '*intraparenchymal*' to a finding term *hemorrhage* and prepend combinations of two brain lobes such as '*frontoparietal*' to terms like *lobe* and *cortex*). Finally, we add the plural forms of all terms to the dictionary. For including terms related to 'RelativePosition', 'PositionStatus', and 'Hedge' FEs to Rad-Entity$_{\text{dict}}$, we construct a list of terms using both RAD-SP$_{\text{DEV}}$ and manually curated terms. Additionally, for 'RelativePosition' and 'Hedge', we add all RadLex terms under RadLex class *Location descriptor* (RID5817) and *Certainty descriptor* (RID29), respectively. This increases the total number of terms to 1,492,109.

3. Entity tagging - We apply an entity tagger that extracts all possible text spans in a sentence representing any spatial FE by exactly matching against the terms in Rad-Entity$_{\text{dict}}$. For the identified spans having any overlap, the longest span is selected except for a few special cases. Such exceptions include anatomy entities (e.g., *inferior cerebellar peduncle*) that contains location descriptor-related terms (*inferior*) in which cases we select both *inferior* and the

[*] https://radiopaedia.org/articles/medical-abbreviations-and-acronyms-a?lang=us

main anatomical entity *cerebellar peduncle* as two candidate entity spans instead of selecting the longest span. Similarly, candidate spatial triggers are identified using dictionary matching against terms in Spatial-Trigger$_{\text{dict}}$. For entities representing 'Distance' FE (e.g., '*2 mm*'), regular expressions (inspired from Bozkurt et al. (2019)) are applied for matching. Besides 'Distance', we also develop regular expressions for identifying certain anatomical entities representing segments (e.g., '*C5-C7*', '*T12*').

Finally, all possible {trigger, entity} pairs are generated by combining each identified trigger with all identified radiological entities in a sentence. All these pairs form the candidate spatially-related entities.

7.2.2 Labeling functions

This step involves writing rules or labeling functions (LFs) considering both radiology report-specific language characteristics and domain lexicons to vote on a {trigger, entity} pair's potential FE label (from a set of nine labels corresponding to nine spatial FEs). Given a {trigger, entity} pair as input, each LF either assigns an FE label for the pair or abstains (i.e., assigns no label). Most LFs include combining dictionary-matching and task-specific heuristics. Dictionaries used in LFs are constructed in a similar manner as for the entity tagging step (described in subsection 7.2.1) for broad categories such as finding, device, and anatomy. Matching against terms in specific dictionaries constrains the semantic type of an entity whereas the task-specific heuristics captures prominent cues to identify the potential spatial role of an entity with respect to a spatial trigger by using linguistic features of a sentence documenting any important clinical information about radiological findings. Examples of heuristics used in LFs to vote a candidate {trigger, entity} pair as 'Ground' and 'Diagnosis' FE are illustrated in Table 7.1.

Besides relying on domain ontologies (RadLex, FMA) through dictionary match, the task-

Table 7.1: Heuristics used in two sample LFs to label a {trigger (SpTrg), entity (RadEnt)} pair with Ground and Diagnosis frame element relations. SpTrgs are **bolded** and FEs are underlined. FE - Frame Element. LF - Labeling Function.

FE (Features used in LF)	Example sentence	Heuristics
Ground (Relative position of RadEnt with respect to SpTrg; semantic type of RadEnt)	The lungs **demonstrate** hazy bilateral opacity of hyaline membrane disease.	• SpTrg is any of [*with*|*without*|*show(s)*|*demonstrate(s)*|*is*|*are*|*reveal(s)*] AND RadEnt lies directly adjacent to the left of the SpTrg • For other SpTrgs, RadEnt lies to the right of SpTrg AND there is 0-2 words in between SpTrg and RadEnt AND RadEnt is anatomy
Ground (Closest SpTrg; semantic type of RadEnt)	There are scattered T2 high signal intensity foci **in** the periventricular white matter and centrum semi-ovale consistent with microvascular angiopathy.	RadEnt lies to the right of SpTrg AND • 0 word in between SpTrg and RadEnt AND RadEnt is anatomy • For greater than 0 word in between SpTrg and RadEnt, no other trigger in between AND RadEnt is anatomy
Diagnosis (Closest SpTrg; semantic type of RadEnt; presence of hedge to the left of RadEnt; RadEnt being the last term)	A patchy area of consolidation is seen **within** the right lower lobe concerning for pneumonia.	RadEnt is finding AND text span to the right of RadEnt is '.' AND • IF preposition-containing hedge term between SpTrg and RadEnt • ELSE IF a hedge term present to the left of the RadEnt with window length 4 and no additional spatial trigger between SpTrg and RadEnt
Diagnosis (Presence of specific hedge-related terms surrounding RadEnt; semantic type of RadEnt)	There is stable opacity **in** the right lower lobe as well as a retrocardiac opacity, these are likely related to atelactases versus pneumonia.	left window = [*represent*, *suggest*, *indicat*, *consistent with*]; right window = [*ruled out*, *excluded*, *vs*, *versus*] • any item in left window list present to the left of RadEnt with window length 4 AND RadEnt is finding • any item in right window list present to the right of RadEnt with window length 4 AND RadEnt is finding

specific heuristics are necessary for a complicated task like this where identifying a spatial role (or FE) against a radiological entity is context-dependent (e.g., a finding term could be both 'Figure' and 'Diagnosis' depending on what role it plays in a sentence). This becomes more challenging when there are multiple spatial triggers in a sentence and the same entity is associated with different triggers with different spatial roles (e.g., an anatomical entity could be both a 'Figure' and 'Ground'). Our LFs handle this complexity by considering the position of an entity with respect to a specific spatial trigger in a sentence. We manually examine the sentences in RAD-SP$_{\text{DEV}}$ to build the LFs. The LFs are developed and refined in iteration by evaluating on the annotated RAD-SP$_{\text{DEV}}$ set. We develop 19 LFs in total. The heuristics used in all LFs are included in Appendix (Table A.1).

7.2.3 Weak label generation

We use Snorkel's generative model (known as a Label model) that combines the noisy label outputs from all LFs for a {trigger, entity} pair by estimating the unobserved accuracy of each LF to assign a single probabilistic label for that pair. This generates probablistic training labels (or "weak" labels) for all candidate pairs extracted from the unlabeled report sentences. Since our task is to identify the FEs at the level of each spatial trigger in a sentence, we create separate instance for each trigger by combining all the RADENTS from the {trigger, entity} pairs for which an FE label has been predicted. These modified trigger-level instances are used for further processing.

7.2.4 Weak label filtering

We apply two additional constraints to filter the weak labels generated by the Label model to produce a sizable improved weakly labeled training data. First, since 'Figure' and 'Ground' constitute the two fundamental FEs of a spatial frame, we check for the presence of both these FEs among the weak FE label predictions in the trigger-level instances from subsection 7.2.3. Second, we check for the presence of certain frequent phrases surrounding common spatial triggers such as *of*, *with*, and

in. We ensure that no such phrase is found to the left or right of the trigger. This rule eliminates false positive triggers (the frequent phrase sets are taken from Datta & Roberts (2020)). Only if the two constraints satisfy, we select that trigger-level instance in our final weak labeled training set.

7.2.5 Weakly supervised model - BERT

We use the final weak labeled training data to fine-tune $\text{BERT}_{\text{LARGE}}$-MIMIC (pre-trained on MIMIC notes for 300K steps (Si et al., 2019)). We formulate this as a sequence labeling task where we extract spatial FEs provided a spatial trigger in a sentence. For this, we convert each trigger-level instance produced from Label model to a sequence of BIO (B-Beginning, Inside-I, and Outside-O) labels against each word in a sentence. The process of filtering the weak labels and transforming to BIO tag sequence is shown in Figure 7.2. Each sentence, represented using the standard input sequence format - [[CLS] sentence [SEP]], is then fed into BERT. As there can be multiple triggers in a sentence, we mask the words corresponding to a spatial trigger with an identifier $sptrg$ to encode the position of a specific trigger. The contextual representations from the BERT encoder output is fed into a linear classification layer to predict labels per word.

7.3 Experimental settings

7.3.1 Without gold data

We use varying amounts of unlabeled MIMIC-III radiology reports to generate weak spatial labels and then use these labels to fine-tune the $\text{BERT}_{\text{LARGE}}$-MIMIC model. Specifically, we use 10, 25, 50, 75, and 100 percent of the 50k selected MIMIC reports. We evaluate each variant on the 358 gold annotated test reports ($\text{Rad-Sp}_{\text{Test}}$).

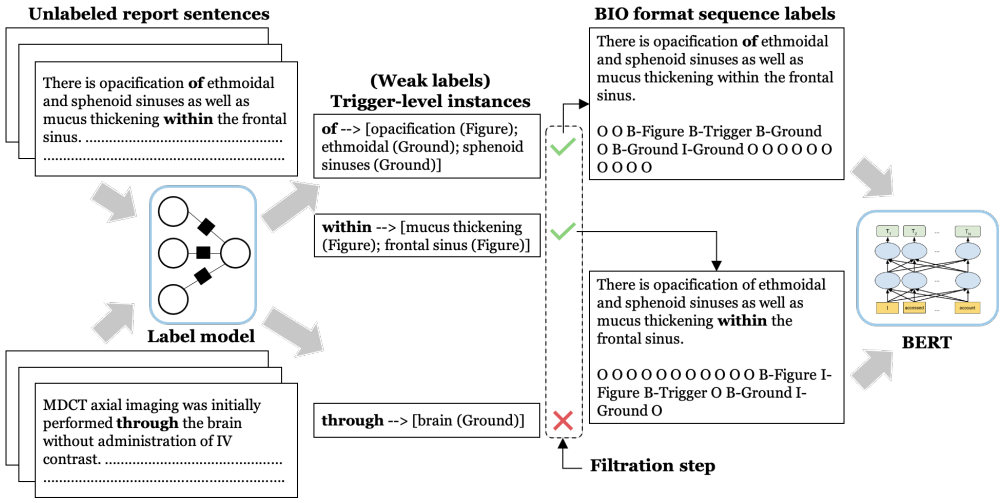

Figure 7.2: Filtering weak labels and converting the labels to feed into BERT model. All the candidate spatial triggers are shown in **bold**.

7.3.2 Sequential fine-tuning

We perform sequential fine-tuning where we first fine-tune BERT on weakly labeled reports followed by fine-tuning on gold annotated reports (a similar approach that proved to be effective in a recent work by Smit et al. (2020) to improve the performance of the automatic rule-based labeler like CheXpert (Irvin et al., 2019)). Specifically, we leverage the best $\text{BERT}_{\text{LARGE}}$-MIMIC model variant among the five variants trained on only weak labels in subsection 7.3.1 to initialize the model parameters to further fine-tune on gold reports. We report the average F1 measures of a 10-fold cross validation on $\text{Rad-Sp}_{\text{Test}}$. We use the results (average F1 using predicted triggers) of a fully supervised $\text{BERT}_{\text{LARGE}}$-MIMIC sequence labeling model reported in Table 4.6 in Chapter 4 for direct comparison. Note that we use the same 10-fold split settings as used in the previous work for this experiment.

7.3.3 Varying amounts of gold data

We also experiment using increasing amounts of gold annotated reports for sequential fine-tuning. Similar to subsection 7.3.2, the best trained model on weak labels is used to further fine-tune on the gold reports. We use 10 percent of 358 annotated reports, each for testing and development, and the remaining 80 percent (i.e., 288 reports) for training. Specifically, we use 10, 25, 50, 75, and 100 percent of the 288 gold reports for sequential fine-tuning.

7.4 Evaluation

To evaluate our weakly supervised BERT model's end-to-end performance in extracting the spatial FEs with respect to a trigger, we consider the spatial triggers that are predicted by the Label model on the test data (Rad-Sp_{Test}) after applying the two constraints (refer subsection 7.2.4 above) to form the model input to BERT. Here, we apply the first constraint with a slight modification (i.e., we filter the trigger-level instances for which a Ground FE is predicted by the Label model) in order to increase the recall of spatial triggers as the fine-tuning task uses the trigger positions to predict the associated FEs. We take into account the precision loss related to FEs that are predicted for false positive (FP) spatial triggers and recall loss related to FEs that are missed for false negative (FN) triggers. These FP and FN triggers are based on the predictions of the Label model on Rad-Sp_{Test}.

The hyperparameters for both the Label and $\text{BERT}_{\text{LARGE}}$-MIMIC models are tuned using the 60 annotated reports in Rad-Sp_{Dev} through grid search. For Label model, the number of training epochs, learning rate, L2 regularization, and precision initialization are set at 100, 0.0001, 0.01, and 0.7, respectively. For $\text{BERT}_{\text{LARGE}}$-MIMIC fine-tuning, the maximum sequence length, learning rate, training epochs are chosen as 128, 2e−5, and 4, respectively and we use the cased version of the model.

Table 7.2: F1 measures of the weakly supervised $\text{BERT}_{\text{LARGE}}$-MIMIC model on $\text{RAD-SP}_{\text{TEST}}$. All the values for the 'Associated Process' frame element are zero.

	# Weak labeled reports for training						
FRAME ELEMENTS	500	1*k*	5*k*	12.5*k*	25*k*	37.5*k*	50*k*
Figure	32.56	41.12	45.47	45.18	44.66	45.93	45.37
Ground	54.16	61.90	62.62	63.25	63.53	63.07	63.06
Hedge	42.75	41.34	53.65	50.51	51.84	55.00	56.38
Diagnosis	20.77	25.05	29.62	31.23	33.55	33.26	34.12
Position Status	43.75	45.08	42.66	38.32	39.88	40.92	41.04
Relative Position	50.36	50.15	48.34	48.88	49.13	49.22	49.17
Distance	70.48	64.46	65.49	63.79	66.67	66.67	67.31
Reason	0	0	19.51	27.69	28.12	30.77	31.43
Overall	43.40	49.01	51.64	51.64	51.92	52.47	52.43

7.5 RESULTS

The coverage of the candidate generation phase is 78.3%, or in other words, our candidate generator identifies 78.3% of the total gold {spatial trigger, radiological entity} pairs from $\text{RAD-SP}_{\text{TEST}}$. The performance measures of the weakly supervised $\text{BERT}_{\text{LARGE}}$-MIMIC model using increasing amounts of weakly labeled reports are presented in Table 7.2. We see that the best overall F1 on $\text{RAD-SP}_{\text{TEST}}$ is obtained when 37.5k weakly labeled reports are used. The precision, recall, and F1 measure for identifying the spatial triggers on $\text{RAD-SP}_{\text{TEST}}$ are 70.84%, 75.07%, and 72.89, respectively. We additionally present the results for 500 and 1*k* weakly labeled reports (in Table 7.2) to highlight the performance trend even when less than 10% of reports are used in fine-tuning.

Table 7.3 shows the sequential fine-tuning results when the model checkpoint corresponding to using 37.5k weakly labeled reports are used to further fine-tune on gold reports. We note that the sequential fine-tuning helps to improve the performance for all spatial FEs except for the 'Relative Position' FE when compared to the fully-supervised variant. To demonstrate the effect of increasing

Table 7.3: Average F1 measures of $\text{BERT}_{\text{LARGE}}$-MIMIC model over 10-fold CV through sequential fine-tuning (using the model checkpoint obtained after fine-tuning on weak labels of 37.5*k* reports). FS-F1 - Fully supervised F1 measures.

Frame Elements	Precision(%)	Recall (%)	F1	FS-F1
Figure	71.25	64.90	67.69	65.12
Ground	74.64	69.33	71.74	71.51
Hedge	67.60	61.87	64.04	57.82
Diagnosis	58.12	56.14	56.54	50.76
Position Status	66.18	71.22	68.05	60.37
Relative Position	65.99	65.18	64.96	66.33
Distance	88.00	90.83	88.36	88.05
Reason	53.33	51.29	45.19	0
Associated Process	60.00	45.00	50.00	0
Overall	71.26	66.71	68.76	66.25

size of gold annotated data on the model's performance, we present the sequential fine-tuning results with varying gold reports in Table 7.4 on a randomly selected 35 annotated test reports. As expected, the performance of the $\text{BERT}_{\text{LARGE}}$-MIMIC model improves as the gold data size is increased, however, we observe that the highest or a comparable overall F1 measure is achieved using 75 percent (i.e., 213 reports) of the total annotated reports available for training. Also note that the results for the 'Reason' and 'Associated Process' FEs are zero in many cases as these are found very infrequently in the dataset.

7.6 Discussion

We develop a weakly supervised pipeline based on data programming technique to extract spatial relations from radiology text. This is an early attempt to automatically create weak labels in the radiology domain covering detailed and important spatial information of clinical importance

Table 7.4: Sequential fine-tuning results (F1 measures) of $\text{BERT}_{\text{LARGE}}$-MIMIC (using the model checkpoint obtained after fine-tuning on weak labels of 37.5*k* reports) on randomly selected 35 test reports with increasing amount of gold reports used in the fine-tuning process. All the values for the 'Reason' frame element are zero. 100% corresponds to 288 gold reports.

	% of gold reports used				
Frame Elements	10	25	50	75	100
Figure	51.72	54.75	56.82	59.55	57.78
Ground	64.52	68.09	66.32	66.31	67.38
Hedge	55.00	68.18	76.92	76.92	73.17
Diagnosis	48.48	54.55	53.33	70.97	62.50
Position Status	54.55	61.54	61.54	57.14	66.67
Relative Position	53.66	60.47	58.54	55.00	55.81
Distance	40.00	33.33	66.67	57.14	75.00
Associated Process	0	0	100	100	100
Overall	56.05	60.43	62.03	63.62	63.01

that could be used for various clinical informatics applications, unlike the three previous studies that employed weak supervision for simpler binary classification and anatomy tagging (Dunnmon et al., 2020; Wang et al., 2019a; Eyuboglu et al., 2021). The results in Table 7.2 demonstrate that our proposed pipeline performs decently given the complexity of information extracted and without any reliance on the time-consuming and expensive manual labeling process. Although they do not surpass the fully-supervised model's performance, they hold the potential to identify a variety of important clinical information without using any hand-labeled training data. Further, our findings on sequential fine-tuning (Table 7.3) also reflect the advantages of leveraging a MIMIC pre-trained model first fine-tuned on domain and task specific data (weakly labeled radiology reports) and then on gold annotated data instead of just fine-tuning on the gold data. This is in line with the findings demonstrated for the CheXbert model where combining the annotations of a rule-based labeler and expert annotations resulted in better performance (Smit et al., 2020).

Our weak supervision approach provides sufficient flexibility as the labeling functions can be fairly easily modified to incorporate any change in reporting style and other institutional reports. For example, let's consider the following two reporting styles:

1. *There are calcified atherosclerotic changes in the brain parenchyma.*

2. *Brain parenchyma: There are calcified atherosclerotic changes.*

While the first style is more commonly encountered in radiology reports to describe findings (e.g., *atherosclerotic changes*) and their locations (e.g., *brain parenchyma*), some institutions or radiologists may prefer the second format (i.e., location: findings). Such changes in reporting format may necessitate some updations in the labeling functions which could be easily incorporated as and when needed. Moreover, as new frame elements are added to the representation schema for different downstream use cases, we can add additional labeling functions to cover those. Additionally, as the labeling functions are developed using the more general language characteristics of radiology text and the finding/anatomy dictionaries are primarily based on standard medical ontologies (e.g., FMA), they are mostly generalizable, i.e., they hold potential in identifying the spatial FEs belonging to different imaging modalities and human anatomies (beyond the three report types used in this work). For instance, our labeling functions will be able to identify spatial information from a pelvic ultrasound report sentence as well (e.g., identifying the finding *leiomyomas* and its location *uterus* from the sentence–"*Multiple leiomyomas in the uterus.*"). This is not explored in this work, however, we plan to do a thorough analysis to examine the performance of our weak supervision pipeline when applied to reports of multiple institutions, modalities, and anatomies in the future.

Although the dictionaries we developed in this work are comprehensive enough (at least for the three radiology sub-areas considered in this paper), we intend to further improve the coverage of the candidate generation step that generates the candidate {spatial trigger, radiological entity} pairs by further expanding the terms in the dictionaries that can detect more variation of radiological entities.

The coverage is mainly impacted because of misspellings and the dictionaries lacking less common phrase variations representing findings and anatomies. Some of such phrases include findings such as *gastric distention*, *gross formational abnormalities*, *low attenuation structure*, *signal gap*, and *mesenteric stranding* and anatomical locations such as *portal vein*, *deep venous system*, *mediastinal margin*, *cavernous carotid*, and *antecubital fossa*. This also reflects the challenges involved in creating more robust dictionaries as there are different and many possible ways of expressing radiological entities and we leave this to future work.

8

Application of Rad-SpatialNet for Ischemic Stroke Phenotyping

This chapter describes an ischemic stroke phenotyping application system that is developed leveraging spatial information from radiology reports.

8.1 Introduction

Ischemic stroke (IS) accounts for around 87% of all strokes in the United States (Division for Heart Disease and Stroke Prevention, 2020). Clinical trials and epidemiological studies targeted toward investigating communication, cognitive, and emotional changes after stroke are interested in analyzing specific subsets of patient records pertaining to certain characteristics of IS for treatment

and prognosis research. Radiological findings documented in head computed tomography (CT) and brain magnetic resonance imaging (MRI) reports provide important information to develop IS phenotypes. Understanding and identifying various clinically important information from the report text can facilitate in constructing fine-grained phenotypes. In this work, we propose to utilize spatial information in the reports to construct IS phenotypes. We develop and evaluate a natural language processing (NLP) pipeline for IS phenotyping by using spatial information extracted from the reports. More specifically, we use the spatially-related imaging features and their brain locations as well as the potential diagnoses information to classify the phenotypes.

Effects of stroke in a patient are dependent on the areas of the brain affected (Johns Hopkins Medicine, 2022; Hui et al., 2020). Based on the side and the particular location of the stroke, different body functions are impaired. For example, stroke in the right side of cerebral hemisphere results in *left-sided weakness or paralysis*, *visual*, and *spatial problems*, stroke in the cerebellum manifests in a different set of effects such as *ataxia*, *dizziness*, *nausea*, and *vomiting*, whereas stroke located in the brainstem results in problems associated with *breathing*, *balance*, and *coma*. Moreover, the effects can be further specified based on the particular lobe of the cerebral hemisphere that is affected. For example, *sensation* and *spatial awareness* are impacted with stroke in the parietal lobe whereas *language* and *memory* are impaired with stroke in the temporal lobe. A previous work (Cheng Bastian et al., 2014) has demonstrated that location of stroke infarct influences the functional outcome following an ischemic stroke as measured by modified Rankin Scale, a commonly used scale for rating stroke outcome in clinical trials. Further, a few studies (Shi et al., 2017; Price et al., 2010) have focused on the brain locations affected by stroke for improving treatment of post-stroke depression and predicting post-stroke language outcome. Therefore, categorizing imaging reports according to stroke location–or in other words, constructing phenotypes incorporating the stroke location–holds potential benefits for clinical research studies that focus on targeted treatment based on the specific brain region affected.

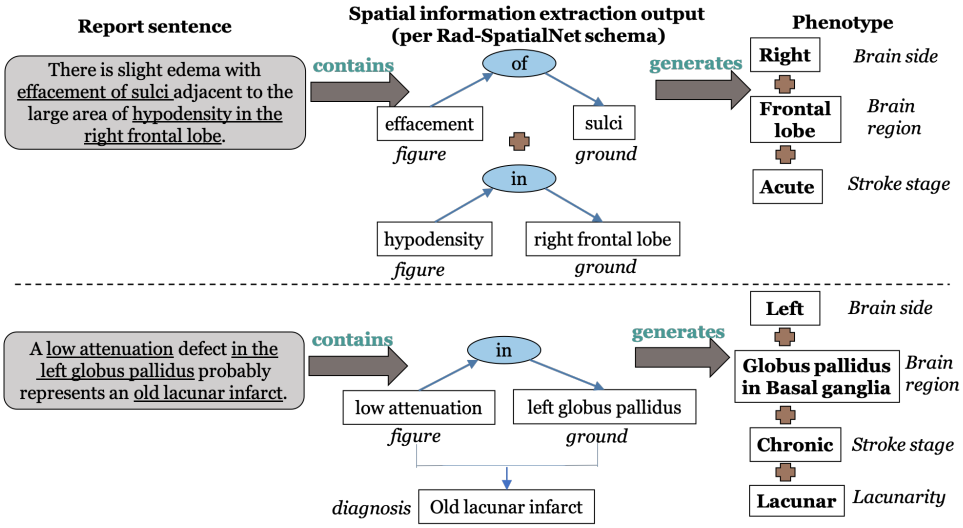

Figure 8.1: Examples of stroke phenotypes using spatial relations from reports. Blue ovals contain spatial triggers.

We construct the IS phenotypes by using the brain location information in the reports both directly and indirectly. Direct use refers to including the side and the specific brain region affected by stroke in the phenotypes. Indirect use of location includes deriving other crucial information such as stroke stage based on the particular brain region a certain imaging feature is detected. Besides these, we also use the IS-related potential diagnoses information directly in the phenotypes to extract the stroke stage in cases when it is included as part of the diagnosis phrase (e.g., *subacute* stage in the diagnosis phrase '*subacute infarction*').

Consider the two examples shown in Figure 8.1 from head CT reports. The first sentence captures information corresponding to mass effect like *sulcal effacement* along with imaging feature such as *cortical hypodensity* that helps to indicate that an *infarction* is *acute*. The second sentence detects an area of *low attenuation* in the left side of *globus pallidus*, part of *basal ganglia*. The sentence also describes that this finding indicates that the infarct is *lacunar* and thus *chronic*. Therefore,

for the first example, we see that spatial relations between imaging features and brain locations (as indicated by phrases like '*effacement of sulci*' and '*hypodensity in the right frontal lobe*') encode important radiological information that facilitates in determining the diagnosis (i.e., *infarction*) and its stage (i.e., *acute*). Also, note that although *acute* is not mentioned explicitly in this sentence, identifying the spatial relations help in inferring that the stroke is *acute*. Thus, spatial relations present in imaging reports can directly be utilized for constructing stroke phenotypes containing fine-grained location information along with additional derived information like stroke stage. We, therefore, use our proposed spatial representation schema–Rad-SpatialNet (refer Section 3.2.1 in Chapter 3) to extract spatial information from reports which can subsequently be used for extracting important IS phenotypes.

Prior studies have attempted to extract IS-related information from radiology reports. Wheater et al. (2019) developed brain imaging phenotypes, however, these phenotypes lacked specificity in the brain location information and were classified as only cortical or deep. Other works identified reports with acute IS (Ong et al., 2020; Kim et al., 2019) and silent brain infarcts (Fu et al., 2019). However, these studies focused on limited information like classifying reports based on presence/absence of IS, acuity, and middle cerebral artery (MCA) territory involvement. Alternatively, we aim to construct specific stroke phenotypes containing more granular information for each stroke affected brain area and this makes the task more complex compared to performing binary classification of the reports. We illustrate the granularity and complexity of our phenotypes in Figure 8.2. Note that the phenotypes consider information at the level of both side and region of the brain affected. Thus we see the stage is *acute* for right cerebellar hemisphere and *chronic* for the left side.

Therefore, using spatial information from the reports forms an intuitive way to extract such fine-grained information for constructing the phenotypes. We define the fine-grained stroke phenotypes described above with input from radiology experts. For automatic labeling of the reports with the relevant phenotypes, we first identify the spatial relations using a transformer-based model

Sentences from a report

There is an **acute infarction** in the **lateral** **aspect** of the **right cerebellar hemisphere**. There are several small **acute infarctions** in the **right midbrain**. There is **encephalomalacia** and **gliosis** in the **inferior** **left cerebellar hemisphere**.

Brain side	Brain region	Stroke stage	Lacunarity
Right	Cerebellar hemisphere	Acute	No
Left	Cerebellar hemisphere	Chronic	No
Right	Midbrain	Acute	No

Phenotypes containing brain region-specific information

Figure 8.2: Granular phenotypes considered in this work (shown for a sample report).

(BERT (Devlin et al., 2019)) for each report. We then apply rules based on domain knowledge on the extracted spatial information to classify the phenotypes. Finally, we evaluate our system by comparing the automatically generated phenotypes with the gold phenotypes for a set of head CT and brain MRI reports. The main contributions of this chapter include:

- Classify fine-grained ischemic stroke phenotypes by applying simple domain rules on top of spatial information extracted from neuroradiology reports.
- Phenotypes contain information targeted at the level of a specific side and region of the brain affected.

8.2 Dataset

We select a set of 150 MIMIC reports (containing a mix of brain MRIs and head CTs) to classify the ischemic stroke phenotypes. These 150 reports contain at least one of the ICD-9 ischemic stroke-relaed diagnosis codes from 433.01, 433.11, 433.21, 433.31, 433.81, 433.91, 434.01, 434.11,

434.91, and 436. We refer to this phenotyping dataset as Rad-IS-P. To train our spatial information extraction (IE) model, we use 400 MIMIC-III (Johnson et al., 2016) radiology reports (consisting of chest X-rays, brain MRIs, and babygrams) annotated following Rad-SpatialNet schema (refer Section 3.2.1 in Chapter 3). Since we extract stroke phenotypes from both types of neuroradiology reports, i.e. MRIs and CTs, we annotated a few (15) head CT reports following the same schema to add to the training data for our spatial IE system. Thus, we use the combined set of 415 reports for training the IE model. We refer to this dataset as Rad-Spatial-IE.

8.3 Dataset Annotation

Each MRI and CT report is annotated with important IS features as validated by a practicing radiologist. These features are identified based on both their clinical importance as well as taking into account the types of information covered in Rad-SpatialNet schema. The pre-defined features are described as follows:

1. Brain side - the laterality of the brain that is affected
2. Brain region - refers to the specific brain area affected due to reduced blood and oxygen supply
3. Stroke stage - three main stages used to describe the CT manifestations of stroke: acute, subacute, and chronic (as described in Birenbaum et al. (2011)). Additionally, some reports document the stage information as acute/subacute, so we also consider acute/subacute separately
4. Lacunarity - whether infarct is lacunar or not. Lacunar infarcts are usually small noncortical infarcts (diameter of 0.2 to 15 mm) and are caused by occlusion of a small perforating artery

Table 8.1: Annotated phenotypes per brain region.

Brain region affected	Frequency	Brain region affected	Frequency
Cerebral hemisphere	26	Basal ganglia	38
Cerebral hemisphere - Frontal lobe	61	Thalamus	6
Cerebral hemisphere - Occipital lobe	30	Cerebral peduncle	2
Cerebral hemisphere - Parietal lobe	46	Internal/External capsule	8
Cerebral hemisphere - Temporal lobe	29	Corona radiata	4
Cerebellum	35	Insula	15
Brainstem	9	Watershed	4

Multiple combinations of these four features can be present in a report. In such cases, we label each report with a maximum of five combinations of brain side, region, stroke stage, and lacunarity. For the example in Figure 8.2, the resulting feature combinations used for annotating the report are – 1. right, cerebellum, acute, not lacunar, 2. left, cerebellum, chronic, not lacunar, and 3. right, brainstem (midbrain), acute, not lacunar. Another point to note is that if the stroke stage is directly available as part of the spatial information extracted from the report, we use that information to annotate the report, otherwise the stage annotation is determined based on certain additional conditions/domain constraints applied over the extracted spatial information. For example, in the sentence "*There are several small acute infarctions in the right midbrain*" in Figure 8.2, *acute* was directly available as part of the Figure frame element *acute infarctions* identified in context to the spatial trigger *in*. However, in the last sentence, the stage is annotated as *chronic* because of the presence of terms like *encephalomalacia* and *gliosis*. Using this annotation scheme, the RAD-IS-P dataset was annotated with the stroke phenotypes by a radiologist. A brief statistics of the brain region-wise phenotype annotations are shown in Table 8.1.

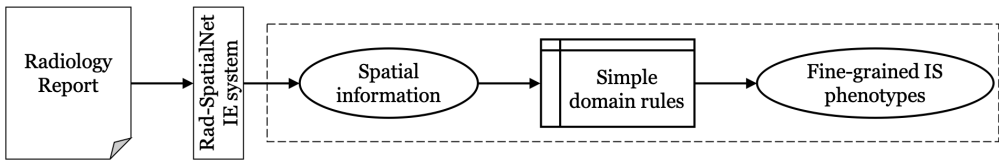

Figure 8.3: Pipeline for ischemic stroke (IS) phenotype classification. Dashed box indicates the main contribution of this work. IE - information extraction.

8.3.1 Phenotyping Method

We use the output of a spatial information extraction (IE) system (information represented following the Rad-SpatialNet schema) to classify the granular ischemic stroke phenotypes. A set of simple domain rules are applied on the output of the IE system for classifying the phenotypes. An overview of our approach is shown in Figure 8.3. We describe the sequential stages of our phenotype extraction system in the following sections.

8.3.1.1 Spatial information extraction

We use an existing BERT-based sequence labeling system for extracting the spatial information from the reports (described in Section 4.1.2 in Chapter 4). This includes identifying the spatial triggers in a sentence followed by identifying the associated frame elements for each extracted trigger. Both spatial trigger and frame element extraction are framed as sequence labeling task. The frame elements identified by the BERT system for each of the spatial triggers in a sample head CT report sentence are illustrated in Figure 8.4. Specifically, in this work, we re-train the BERT-based frame element extractor using the Rad-Spatial-IE data with updated annotation spans for a few frame elements as described below.

Spatial trigger (lexical unit)	Frame elements
Spatial Frame - 1	
in	Figure (*areas of restricted diffusion*) Ground (*vascular territory*) Hedge (*suggesting*) Diagnosis (*thromboembolic ischemic changes*)
Spatial Frame - 2	
of	Figure (*vascular territory*) Ground (*right MCA*)
Spatial Frame - 3	
on	Figure (*hyperintense foci*) Ground (*right occipital lobe, right basal ganglia*) Hedge (*suggesting*) Diagnosis (*thromboembolic ischemic changes*)
Spatial Frame - 4	
on	Figure (*hyperintense foci*) Ground (*right temporal lobe*) Relative Position (*distally*) Hedge (*suggesting*) Diagnosis (*thromboembolic ischemic changes*)

Figure 8.4: Spatial frames extracted for a sample sentence–*There are areas of restricted diffusion **in** the vascular territory **of** the right MCA, also some scattered hyperintense foci noted **on** the right occipital lobe, right basal ganglia and distally **on** the right temporal lobe suggesting thromboembolic ischemic changes.*

Updates to Rad-SpatialNet for Ground and Diagnosis frame elements Note that for each anatomical location phrase labeled as Ground element in the Rad-SpatialNet schema, the associated laterality terms such as 'left', 'right', and 'bilateral' were annotated as elements in context to that anatomical radiological entity. Similarly, for some of the potential diagnoses labeled as Diagnosis element, the associated temporal descriptors such as 'acute', 'evolving', and 'chronic' were also annotated as elements in context to the diagnosis radiological entity. Thus, the laterality and the temporal descriptor terms were not part of the Ground and Diagnosis frame elements respectively (in turn not directly connected to the spatial triggers) and thus were not identified by

the spatial frame element extraction system. However, considering the need to capture laterality and diagnosis temporality information for our phenotyping task, we updated the mention spans of the Ground and Diagnosis elements in the sentences to support this work. Consider the following examples:

1. Include the laterality of the anatomical location

   Rad-SpatialNet – *There is hypodensity in the left basal ganglia.*

   This chapter – *There is hypodensity in the left basal ganglia.*

2. Include laterality and location descriptor whose span falls in between a laterality phrase and the anatomy phrase

   Rad-SpatialNet – *A small area of white matter hyperintensity in the right frontal subcortical region.*

   This chapter – *A small area of white matter hyperintensity in the right frontal subcortical region.*

3. Include the temporality of the potential diagnosis

   Rad-SpatialNet – *Hypoattenuation in the right frontoparietal distribution consistent with acute infarction.*

   This chapter – *Hypoattenuation in the right frontoparietal distribution consistent with acute infarction.*

In the first example we see that '*left*' has been included in the Ground element, and in the second example both '*right*' and '*frontal*' are included in the Ground element span. In the third example,'*acute*' is included in the Diagnosis element span. The spatial trigger (lexical unit for a spatial frame) is '*in*' for all the examples.

8.3.1.2 Automatic IS phenotype extraction

For each report, we use rules on top of the output of the BERT-based element extractor to automatically classify the phenotypes. We combine the spatial frames identified by the element extractor at the report level. We also keep a track of all the spatial frames predicted by the BERT extractor for each sentence in a sequential order (the order in which the spatial triggers appear in a sentence). This helps to combine the frames when the Ground element associated to a trigger is same as the Figure element of the next trigger. For example, in "*acute infarction in the lateral aspect of right cerebellum*", IS-related finding (*infarction*) is connected to the corresponding location (*right cerebellum*) through the common frame element *aspect* of the two spatial frames with triggers *in* and *of* appearing sequentially in the sentence.

For each spatial trigger identified in a sentence, the following steps are performed:

1. First, the spatial triggers and the frame elements relevant to ischemic stroke are filtered. For this, we check if any of the Figure/Diagnosis element spans detected in relation to a trigger is IS-related. If one of the pre-defined IS-related imaging finding keywords (as shown in the first two rows of Table 8.2) is present in any of the element spans, the following steps are performed.

2. For extracting the brain side, we check for the presence of any laterality-related term in the predicted Ground element span (e.g., *left* for left, and *both*, *bilateral* for bilateral). Additionally, if the Ground elements are *thalami* and *capsules*, we assign the side as *bilateral*. In other cases, *unspecified* is assigned. Moreover, in cases (e.g., *infarction involving left frontal and parietal lobes*) when the same laterality is linked to multiple regions, each region is assigned the laterality separately. Here, *left* is assigned to both *frontal* and *parietal* lobes although *left* does not appear in the Ground span *parietal lobes*.

3. For identifying the brain region, the presence of any of the keywords developed for each of the pre-defined brain areas are checked in the detected Ground element span (e.g., keywords for mapping the brain region as 'Basal ganglia' are *basal ganglia*, *caudate*, *caudate nucleus*, *caudate head*, *caudate nucleus head*, *putamen*, *globus pallidus*, and *lentiform nucleus*). These keywords are built with domain expert input. Additionally, for Ground element spans involving two lobes, we assign both the cerebral lobes (e.g., *frontal* and *parietal* lobes are assigned for Ground element span – *frontoparietal*).

4. For identifying the stroke stage for each pair of brain region and side, two sequential steps are involved. First, we check for the presence of any stage-related term directly in the predicted Figure/Diagnosis element span. Since the term *acute* is also contained in *subacute*, we priortize the search for subacute over acute. If not found, domain constraints are applied over the predicted spatial frame elements (this step also takes into account the other spatial relationships predicted in the same report in connection to the same brain region). If the stage is not determined by these two steps, we assign the label – *Can't determine*.

5. Similarly, for identifying the lacunarity for each pair of brain region and side, we check for the presence of lacunar-specific terms in the Figure/Diagnosis element span. We assign a binary lacunarity label – *Yes* if lacunar and *No* otherwise.

The keywords developed for IS-related imaging findings as well as for identifying the stroke stage and lacunarity from the frame element spans are shown in Table 8.2. These keywords as well as the domain constraints for inferring the stage are developed in collaboration with the radiologist who created the gold phenotypes. A few predominant constraints are demonstrated in Table 8.3.

Table 8.2: Keywords for identifying IS finding, IS stage, and lacunarity from the frame element spans to classify the phenotypes.

Item	Keywords
IS-related imaging finding (CT)	hypodensity, hypodensities, hyperdensity, hyperdensities, hypodense, hypoattenuation, hypo-attenuation, low attenuation, low-attenuation, hypoattenuating, hypo-attenuating, low attenuating, low-attenuating, decreased attenuation, lacune, infarct, lesion
IS-related imaging finding (MRI)	restricted diffusion, slow diffusion, susceptibility artifact, signal, infarct
IS stage - Subacute	sub-acute, subacute, sub acute, evolving
IS stage - Acute	acute
IS stage - Chronic	encephalomalacia, gliosis, known, old, previous, prior
Lacunarity	lacune, lacunar

Table 8.3: Domain constraints applied on BERT predicted spatial frame elements to determine ischemic stroke stage.

Modality	Acute	Chronic
CT	(hypodensity/hypoattenuation in cortical/subcortical region) AND (hyperdense MCA OR hyperdensity in basilar artery OR loss of gray-white matter differentiation OR sulcal effacement)	(hypodensity/hypoattenuation in cortical/subcortical region AND (prominence of ventricles/sulci OR atrophy)) OR gliosis/encephalomalacia
MRI	(slow diffusion/restricted diffusion in cortical/subcortical region) OR (loss of flow void in MCA/basilar artery)	facilitated diffusion in cortical/subcortical region OR gliosis/encephalomalacia OR dilation of ventricles

8.4 Experimental Settings and Evaluation

We use the $\text{BERT}_{\text{LARGE}}$ model for fine-tuning the spatial information extraction task by initializing the model parameters obtained after pre-training BERT on MIMIC-III clinical notes for 300, 000 steps (Si et al., 2019). For extracting the spatial triggers from the Rad-IS-P data, we use the trained model from Table 4.4 described in Section 4.1.4 in Chapter 4. However, for extracting the frame elements, we re-train the BERT-based element extractor on the Rad-Spatial-IE dataset

using the updated gold spans of Ground and Diagnosis frame elements for capturing the laterality and temporality information, respectively. We perform 10-fold cross-validation for evaluating the performance of the element extractor model. For each of the 10 iterations, we split the reports in Rad-Spatial-IE such that reports in 8 folds are used for training and 1 fold each are used for validation and testing. The model is fine-tuned by setting the maximum sequence length at 128, learning rate at $2e-5$, and number of training epochs at 4. We use cased version of the models. Among the 10 versions of the trained model checkpoints (generated for 10 folds of the dataset), we select the version based on the highest F1 measure on the validation set to predict the spatial frame elements from the Rad-IS-P data used for phenotype classification. Additionally, to provide a sense of the performance of the spatial information extraction system on stroke-related reports (that are more representative of the ones used for phenotyping), we annotated a random set of 20 reports from the Rad-IS-P dataset according to the Rad-SpatialNet schema and evaluated the system's performance on these 20 reports. For our phenotyping task, we report the precision, recall, and F1 measures of the phenotype extraction system based on various meaningful subsets or combinations of stroke features described in Section 8.3.

8.5 Results

The average precision, recall, and F1 scores of extracting spatial triggers from the Rad-Spatial-IE data are 86.14%, 79.55%, and 82.66, respectively. For the 20 stroke reports (selected from the Rad-IS-P data), the precision, recall, and F1 values for spatial trigger extraction are 93.70%, 76.28%, and 84.10, respectively. These predicted triggers are used further by the element extractor model in the end-to-end evaluation (shown under the 'Predicted spatial triggers' column in Table 8.4). Table 8.4 also highlights the average 10-fold CV performance measures of the BERT-based element extractor using the gold spatial triggers. The frame elements Associated Process and Reason have

Table 8.4: 10 fold CV results on Rad-Spatial-IE for BERT-based spatial frame element extraction model using gold and predicted spatial triggers. P - Precision, R - Recall.

Main Frame Elements	Gold spatial triggers			Predicted spatial triggers		
	P(%)	R(%)	F1	P(%)	R(%)	F1
Figure	81.39	84.26	82.77	67.53	71.08	69.14
Ground	92.01	93.41	92.69	70.87	80.13	75.09
Hedge	75.51	83.08	78.91	68.94	74.05	71.19
Diagnosis	54.73	78.41	64.06	48.49	67.67	55.95
Relative Position	87.47	81.01	83.54	60.13	66.35	62.17
Distance	75.83	80.83	75.53	73.63	80.00	74.25
Position Status	68.59	66.20	66.97	61.42	64.45	61.55
Overall	82.60	85.31	83.92	66.95	73.17	69.81

very low performance scores as they occur very rarely in the whole dataset and also not used for phenotyping. We additionally illustrate the overall precision, recall, and F1 measures (considering all the spatial frame elements) of the frame element extractor on the 20 stroke report subset in Table 8.5.

The results of our phenotype extraction system are shown in Table 8.6. We calculate the performance metrics of the system based on different combinations of the features (i.e., brain region, side, stroke stage, and lacunarity) that are potentially useful for clinical research studies. The precision, recall, and F1 values are calculated by comparing the distinct combinations of the features per report identified by the system to those of the gold annotated ones. This gives an idea about how well the system performs in classifying various subsets of meaningful features. Since stroke stage and lacunarity are associated with a specific brain region and side pair, we report the performance of the system including the stage and lacunarity features along with brain region and side in the last four rows of the table. Note that for stroke stage, we show the results both by considering various stage types and also by grouping the three stage types–acute, subacute, and acute/subacute together.

Table 8.5: BERT-based spatial frame element extractor's performance on 20 stroke reports (taken from Rad-IS-P). P - Precision, R - Recall.

Spatial triggers used	Overall P (%)	Overall R (%)	Overall F1
Gold annotated triggers	72.80	80.87	76.62
Predicted triggers	65.71	73.48	69.38

Table 8.6: Phenotype extraction results. BR - brain region, CS - corresponding side, SS - stroke stage, SS_CO - SS with coarse types (*acute/chronic*), LC - lacunarity.

Phenotype variant	Example	Precision(%)	Recall(%)	F1
BR	*cerebellum*	73.58	89.62	80.81
BR + CS	*cerebellum, left*	68.34	85.47	75.95
BR + SS_CO	*cerebellum, chronic*	55.53	82.0	66.22
BR + CS + SS_CO	*cerebellum, left, chronic*	49.67	74.11	59.48
BR + CS + SS	*cerebral hemisphere - frontal lobe, bilateral, subacute*	46.32	56.96	51.09
BR + CS + LC	*basal ganglia, bilateral, yes*	62.53	77.2	69.09
BR + CS + SS_CO + LC	*basal ganglia, bilateral, chronic, yes*	48.59	72.49	58.18

8.6 Discussion

This work focuses on identifying complex ischemic stroke phenotypes mainly from the perspective of the stroke location (brain region and side). We utilize the output of a spatial information extraction (IE) system (developed in our previous work) and apply simple neuroradiology-specific rules to classify these phenotypes. Note that the phenotypes we tackle in this work consider information at the level of specific brain area that is affected by stroke. Thus, this involves identification of information related to a stroke affected region in the brain from the report text. Our Rad-SpatialNet schema allows for easy identification of such related information as this captures the spatial relations between imaging findings and brain locations as well as the associated potential diagnoses. This becomes even more useful when the same report contains infarcts of different stages in different brain locations. Figure 8.2 illustrates an example where three different brain regions are affected and the stroke stage varies according to the region and its laterality.

We observe that applying simple domain rules that are mainly based on keyword search and a small set of constraints over the output of the spatial IE system results in satisfactory performance in classifying complex stroke phenotypes. This highlights both the information coverage of the Rad-SpatialNet schema and the sufficiently promising performance of the spatial IE system. Another point to note is that the information covered through Rad-SpatialNet are generic enough to extend our phenotype classification approach to other types of diseases/conditions beyond neuroradiology domains.

We briefly discuss the errors of the phenotype extraction system here. Most of the errors related to missing the brain region (referring to the recall of 89.62% in Table 8.6) is because of the Ground elements that are not predicted by the spatial IE system. There are also a very few cases where spatial triggers are not present explicitly (e.g., *left cerebellar infarct*). The existing Rad-SpatialNet schema doesnot capture such implicit relations and thus such regions are missed. Some of the errors related to stroke stage classification (when all the stage types are considered) is due to the ambiguity involved in distinguishing the acute and the subacute stages. Oftentimes, it becomes difficult to assess the stroke timing based on the report content (one of the major reasons for low recall for BR + CS + SS shown in Table 8.6). A small number of errors also occur when only acute and chronic stage information is considered because the output of the spatial IE system sometimes missed the specific stage-related term (e.g., *evolving*, *chronic*) in the predicted Diagnosis/Figure element span. Moreover, the report does not contain other spatial relations to satisfy the domain constraints for stage inference. Another reason of stage-related errors is when the stage information is mentioned in a following sentence in the report that does not contain any spatial relations (e.g., '*These lesions suggest old infarction*'). Lacunar-related errors happen mainly because their inferences sometimes depend on the specific sizes mentioned in the sentence (e.g., *lesion of 7 mm in diameter*) that are currently not captured in the Rad-SpatialNet schema. Taking into account a few limitations as described here in the Rad-SpatialNet schema, we aim to emphasize that there are rare instances of such scenarios

overall across reports and we intend to further incorporate these information in the Rad-SpatialNet in our future work. We also see that the precision values are low, and one of the main reasons is that many of the stroke locations are referenced multiple times in a report and are expressed differently or with varying levels of specificity. For example, *left frontal lobe* is mentioned in the report's Findings section, whereas *left MCA* is mentioned in the Impressions section. This results in generating some false positive brain regions (e.g., *parietal* and *insula* here) as MCA (middle cerebral artery) maps to parts of *frontal* and *parietal lobes* as well as *insula* (the brain regions where MCA supplies blood to). The performance of our phenotype extraction system reflects the challenging nature of this complex phenotyping task and we aim to improve its performance and evaluate on an augmented dataset in a later work.

However, the phenotyping results suggest that the Rad-SpatialNet schema that we used in this work is robust enough considering the complexity of the phenotypes. We want to highlight that the current Rad-SpatialNet schema can be leveraged further to classify more granular aspects of the stroke location. Specifically, the RelativePosition frame element (e.g., *superior*, *inferior*) can be used to classify the subregions of a brain region like *cerebellum*. For instance, in the sentences of the same report–"*New acute infarction involving the superior left cerebellar hemisphere*" and "*Encephalomalacia and gliosis are again seen in the inferior left cerebellar hemisphere*", the stroke stage is *acute* in case of left cerebellum (superior) and *chronic* for left cerebellum (inferior). Thus, spatial information documented in the reports when extracted with detailed contextual information facilitates the classification of fine-grained phenotypes.

9

Application of Radiology Information for Automated Tracking

This chapter describes a natural language processing system that tracks the same radiological findings and the same medical devices across radiology reports of a patient over time.

9.1 Introduction

Radiology reports contain rich descriptions of clinically important findings and medical devices. Oftentimes, these findings and devices are referred to multiple times in a single report and are also referred to across different reports of a patient. Radiologists make such references in multiple reports mainly to highlight any longitudinal changes of a particular finding (e.g., change in a *tumor* at

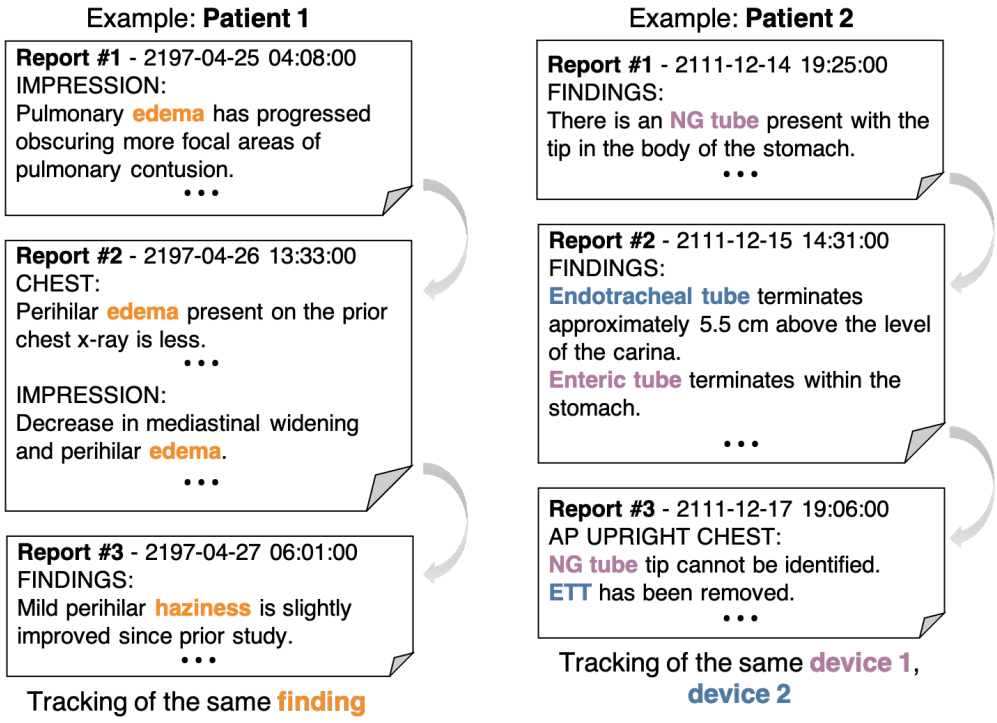

Figure 9.1: Examples of tracking the same finding (*edema*) and the same devices (*NG tube* and *Endotracheal tube*) across multiple reports.

a certain location) and also to describe any interval changes in a device position (e.g., change in the position of an *endotracheal tube* inserted in a patient with respect to an anatomical location). Although extracting important information (e.g., findings, anatomical locations) from radiology reports has been widely studied (Hassanpour & Langlotz, 2016; Steinkamp et al., 2019; Syeda-Mahmood et al., 2020; Sugimoto et al., 2021), tracking (or identifying the coreferences) of radiological findings across reports is unexplored. Automated tracking of findings and devices across a patient's radiology reports holds potential to reduce physician burden in making patient-related decisions as well as to facilitate various retrospective clinical research studies.

Tracking the same finding or device across reports is a challenging problem as it relies on radiology

domain knowledge and requires understanding the linguistic variations used by radiologists as well as understanding both linguistic and domain-specific context across different reports. This is illustrated through an example in Figure 9.1, where, for Patient 1, the *perihilar edema* described in one of the subsequent reports of this patient is referencing to the *pulmonary edema* mentioned in a previous report, and is again described through a different expression, *perihilar haziness*, in a later report. Here, these three finding entities–*pulmonary edema*, *perihilar edema*, and *perihilar haziness* are describing the progress of the same finding for this patient. Similarly, for Patient 2, *NG tube* and *Enteric tube* are discussing the same device, whereas *Endotracheal tube* and *ETT* are describing the change in the status of another device. Thus, we see that there is a strong reliance on domain language knowledge and context information to identify the co-referring expressions of the same findings or devices across reports.

In this chapter, we introduce an annotated dataset to track the same radiological findings and medical devices across reports. We sample a total of 60 patients from the publicly available MIMIC-III clinical database (Johnson et al., 2016), with an average of 10.6 reports per patient. The reports include a variety of imaging modalities covering different human anatomies. Our tracking dataset comprises of a total of 5872 mentions with 2292 mention chains. A chain here represents all the mentions across reports of a patient that refer to the same finding or device entity. We represent the tracking task with enough specificity to capture the clinical granularities that are critical to treatment planning. For example, a *fracture* detected at the right frontal lobe of the skull is different from a *fracture* detected at the left temporal lobe, and, therefore, these two fractures will be placed in two different mention chains. More details are explained in the annotation guideline (Sections 9.3.1 and 9.3.2). Instructions to access the annotated dataset are available at GitHub*. We employ two baseline methods–a rule-based system and a transformer language-based system, BERT (Devlin et al., 2019), to automatically identify the cross-report coreferences. Finally, we evaluate the

*https://github.com/krobertslab/datasets/tree/master/rad-tracking

performance of the systems using standard coreference metrics.

9.2 Dataset

We sample 60 patients from MIMIC-III for creating this annotated tracking dataset with a total of 638 reports. The average number of reports per patient is 10.6, with the maximum being 33. The reports consist of various imaging modalities including X-ray, computed tomography (CT), CT angiography, magnetic resonance imaging (MRI), and ultrasound as well as different body organs such as chest, head, neck, foot, hip, liver, and kidney. The five most frequent modality types are chest X-ray, CT head, CT abdomen, CT C-spine, and abdomen X-ray. The average length of a report in the collection is 244.7 tokens, with the highest being 1490 tokens. Our dataset includes sufficient radiological linguistic variation as the reports belong to different imaging modalities and describe the imaging interpretation of various anatomies. For annotation, 60 patients are split among three annotators with medical background where each report is annotated by two annotators. We are currently in the process of reconciling the annotations.

9.3 Annotation Process

We annotate the finding and medical device instances that refer to the same finding/device across reports for a specific patient. Finding here refers to a radiographic finding described in a report. This includes clinical findings (e.g., pneumonia) and imaging observations (e.g., enhancements such as lesion and foci). Device refers to any medical device including tubes and catheters (e.g., endotracheal tube, central venous catheter). We use the Brat tool (Stenetorp et al., 2012b) for annotation. The reports of a patient are sorted chronologically using the CHARTTIME attribute of the MIMIC table. Since this is a patient-level annotation, we examine all the sequentially arranged reports for a patient to identify the finding/device instances that correspond to the same finding/device. We

assign the same mention identifier to all the entities/mentions across reports that represent the course of a specific finding or device.

9.3.1 Identify references of the same finding

The course of a finding can be roughly represented as – (1) initial detection/diagnosis, (2) improved, worsened, etc., and (3) no longer detected. We came up with the following general rules to track a particular finding:

- Identify the first time a finding is detected
- Identify all the other references of the same finding in the subsequent reports highlighting any change in the characteristics of a finding (e.g, a finding may become large, may improve when compared to a previous study etc.)
- Identify all the references until the last report for a patient is reached or if the finding has been resolved

In certain cases, the corresponding location information of a finding serves as a clue in identifying the same reference of a finding across reports. Let us consider the following two examples for a patient:

- Report 1: Questionable ***aneurysm*** at right posterior communicating artery.
- Report 4: Small ***aneurysm*** of size 2.5 mm arises at the origin of posterior communicating artery.

We see that both aneurysms in the two reports are referring to the same aneurysm and hence will be assigned the same mention identifier (belong to the same mention chain). Note that the location *posterior communicating artery* provides a clue that the aneurysms in these reports are discussing about the same finding.

The findings are tracked at the level of the exact anatomical location. This is described through the following points:

1. If the same finding is detected at a different body location or has moved to a different location, we assign different mention identifiers to these findings. For example, ***opacity*** in ***right lower lobe*** is a different finding than an ***opacity*** in the ***left lower lobe***. So different mention IDs will be assigned to these two opacities and are hence part of two different mention chains.

2. We also differentiate findings based on the hierarchical structure of the anatomies. Thus, a ***left frontotemporal fracture*** and a ***skull fracture*** are placed in two different mention chains as the frontotemporal region is a sub-part of the skull.

3. We separate findings based on their laterality information. For example, ***left pleural effusion*** and ***bilateral effusions*** are placed in different mention chains as bilateral indicates that the effusion is also present on the right side.

4. Common finding terms such as ***normal*** and ***unremarkable*** are tracked separately based on the anatomical location or the observation described. For example, the same finding *normal* is placed in separate mention chains corresponding to the two descriptions–'*The appendix is normal.*' and '*Heart size normal.*', as the former is describing about appendix and the latter is about heart size.

5. Again, if the same finding has re-appeared after a period, a different mention identifier is assigned (e.g., *tumor* re-appearing after a few years).

9.3.2 Identify references of the same device

Similarly, for tracking the medical devices across reports, the same mention identifier is assigned to entities of a device that represent a specific device. The course can be represented as – (1) insertion,

(2) device position status – normal/abnormal, and (3) removal. The following are general rules to track a particular device:

- Identify the first time a device is inserted or placed
- Identify all the other references of the same device in the subsequent reports. This will mainly include updates related to the previously inserted device (e.g., any change in its location or update in the status of its position such as stable, good, satisfactory, unchanged, etc.)
- Continue identifying all the references until the last report for a patient is reached or if the device has been removed

If the same device is re-inserted after a period, we assign a different mention identifier to that device since this indicates a new use of a device. Consider the following sentences from four different reports:

- Report 1: Right ***IJ central venous catheter*** in place and the tip is in distal SVC.
- Report 2: Right ***internal jugular central venous catheter*** in stable position and emanating in middle SVC.
- Report 3: Right ***internal jugular central venous line*** remains in position.
- Report 5: Right ***line*** is terminating in SVC.

Note that all the device mentions in these reports (indicated in bold) are the different variations that are used to refer to the same device and all these mentions are annotated as part of the same mention chain.

9.3.3 Challenges

The challenges involved in creating this dataset broadly fall under two categories – dependence on context both within and across reports and extensive reliance on radiology domain knowledge. Oftentimes, understanding the context is crucial in correctly annotating the same references of a finding. Table 9.1 illustrates a scenario where contextual information documented in a long report helps in identifying the coreferences of a finding – subarachnoid hemorrhage. For the first occurrence of hemorrhage, note that linking the *right frontal* location mentioned a few sentences above to the expression–*occupying the immediately subjacent sulci* in the same sentence where hemorrhage occurs indicates that the hemorrhage is associated with right side of the brain. Again, the second occurrence of hemorrhage is associated with the left side as indicated by the location *left frontotemporal* in the same sentence. And the third occurrence is associated with both sides (left and right). Thus, these three instances of hemorrhage belong to three different mention coreference chains.

There is also a tremendous dependence on domain knowledge. Table 9.2 shows a few example sentence pairs (same/across reports) where domain knowledge of different levels are required for annotation. The first example is simple, where *dissociation* and *disruption* are synonymous terms and can be easily identified as coreferences. The second pair is relatively difficult, requiring basic clinical knowledge, with *swelling* and *edema* referring to the same finding entity. The third pair is at a moderate difficulty level, where *atelectasis* and *collapse* belong to the same mention chain and *pneumonia* and *consolidation* belong to another mention chain. The fourth example requires a deeper knowledge where *vascular markings* and *interstitial edema* refer to the same finding.

9.3.4 Statistics

Some basic statistics of our annotated dataset are shown in Table 9.3. We highlight the five most frequent finding and device mentions in Table 9.4 (note that *"tip"* is a mention that is often

Table 9.1: An example radiology report snippet illustrating the dependence of context for tracking subarachnoid hemorrhage. Findings are in orange, anatomical locations are in green, and the descriptions serving as cues to identify the same finding are **bolded**.

CT HEAD W/O CONTRAST Findings: ..... The right frontal fracture is associated with a focal lentiform extra-axial hematoma measuring roughly 8 mm in thickness and 3.5 cm in maximal transverse dimension. This demonstrates a relatively low-attenuation portion, anteriorly, which may represent acute, non-clotted blood. This collection may be bounded by the coronal suture, and therefore lie in the epidural space. There is moderate subarachnoid hemorrhage occupying the immediately subjacent sulci, which are slightly flattened, due to the mass effect of the hematoma. ..... No significant extra-axial hematoma is identified at the corresponding left frontotemporal fracture site, though there is subarachnoid hemorrhage in the sulci in this region. ..... Impression: ..... Associated subarachnoid hemorrhage at sites described above, with possible small associated right frontal and left frontotemporal contusions .....

documented while referring different medical devices). In terms of inter-annotator agreement, the overall F1 agreement for annotating the mention spans (considering exact span match) is 0.55. The disagreements are mainly related to selecting certain modifier terms describing a radiological findings (e.g., selection of the span *"free intraabdominal air"* by one annotator and only *"air"* by another). For coreference resolution, we calculate the inter-annotator agreement using MUC and CoNLL F1 metrics, and the values are 45.24 and 42.1, respectively.

We provide more insights about our annotated corpus through Figures 9.2, 9.3, and 9.4. Figure 9.2 illustrates the number of different reports that are included while annotating the mention

Table 9.2: Examples denoting reliance on domain knowledge for annotation.

Difficulty	Example pairs
Simple (synonymous)	Some degree of dissociation as well as lateral displacement of the ossicular chain; Complex fracture of the left temporal bone with evidence of lateral displacement and disruption of the left ossicular chain
Simple (clinical knowledge)	There is a left parietovertex soft tissue swelling; There is extensive left supra- and periorbital soft tissue edema
Moderate (clinical knowledge)	There is new patchy opacity at the left lung base, which may represent resolving postoperative atelectasis with effusion, but *pneumonia* cannot be excluded; New retrocardiac collapse/*consolidation* and bilateral effusions
Complex (clinical knowledge)	There is mild prominence of the pulmonary vascular markings without overt evidence for failure; In the interval, there is increased interstitial edema and small-moderate bilateral pleural effusions.

chains. Each stack in a bar highlights the proportion of mention chains according to their lengths (i.e., # mentions in a chain). It is interesting to note that there are more mention chains of length 2 where only a single report contains both the mentions than when the two mentions are present in two different reports (represented in blue). Also, the number of chains of lengths 3 and 4 are the highest when the chains contain mentions from only two reports.

We show the distribution of temporal distances (in weeks) between co-referring mentions in two sequentially ordered reports of a patient in Figure 9.3. Overall, more radiological findings are co-referred than medical devices in radiology reports. We observe that the majority of the coreferences between two consecutive reports occurred within an interval of 2 weeks whereas the maximum interval was found to be 2.15 years.

We also illustrate the overlap of imaging modalities of the reports in the annotated mention

Table 9.3: Dataset statistics.

Item	Count
Avg no. of reports per patient	10.6
Total reports	638
Avg no. of tokens per report	244.7
Min no. of mention chain per patient	8
Max no. of mention chain per patient	110
Total mention chains	2292
Total singleton mention chains	1102
Longest chain length	53
Avg chain length (excluding singletons)	4
Avg no. of tokens per mention	1.44
Total entities (radiological finding)	4978
Total entities (medical device)	894

chains using the UpSet visualization technique (Lex et al., 2014) in Figure 9.4. While a majority of the mention chains contain mentions described in either only X-ray or CT reports, we do find the inclusion of mentions described in multiple modalities. Among two-modality combinations, (X-ray, CT), (CT, CTA), (CT, MR), and (CT, Ultrasound) are the most frequent co-occurring modalities. Among three-modality combinations, (X-ray, CT, CTA) is the most frequent. We also see a very small percent of mention chains spanning four modalities.

9.4 Methods

We frame the tracking task as a cross document coreference resolution (CDCR) problem. We apply two baseline methods for automatically identifying the coreferences of findings and devices across all radiology reports of a patient. First is a simple string matching-based baseline, whereas in the second we employ a BERT-based classification approach to predict the mention chains. Since CDCR is our main focus, we use the gold mentions.

Table 9.4: Top five frequent mentions in the dataset.

Finding	Count	Device	Count
effusion	398	tip	144
pneumothorax	238	ng tube	103
fracture	229	endotracheal tube	101
opacity	180	chest tube	42
atelectasis	176	swan-ganz catheter	36

9.4.1 Rule-based

We perform sentence segmentation and word tokenization using NLTK. We combine all entities or mentions at a patient level. Then all possible mention pairs are generated. If the lower-cased version of the two mention strings in a pair match, we consider that these two mentions will belong to the same chain. All these mention pairs are then combined to construct the chain.

9.4.2 BERT-based

In this approach, given a mention pair, we use BERT as a binary classifier to predict whether the two mentions are coreferences. Specifically, we apply BERT in a sentence pair classification setting where information about the two mentions are combined to form the input sequence. Later, the output generated by BERT for all mention pairs corresponding to a patient is combined to predict the final mention chains. We describe the details in the following sub-sections.

9.4.2.1 Pre-processing

First, we generate all possible mention pair combinations for each patient. We then generate positive and negative pair instances for fine-tuning BERT using gold mention chain information. Since there is imbalance in the number of positive and negative instances (negative instances being

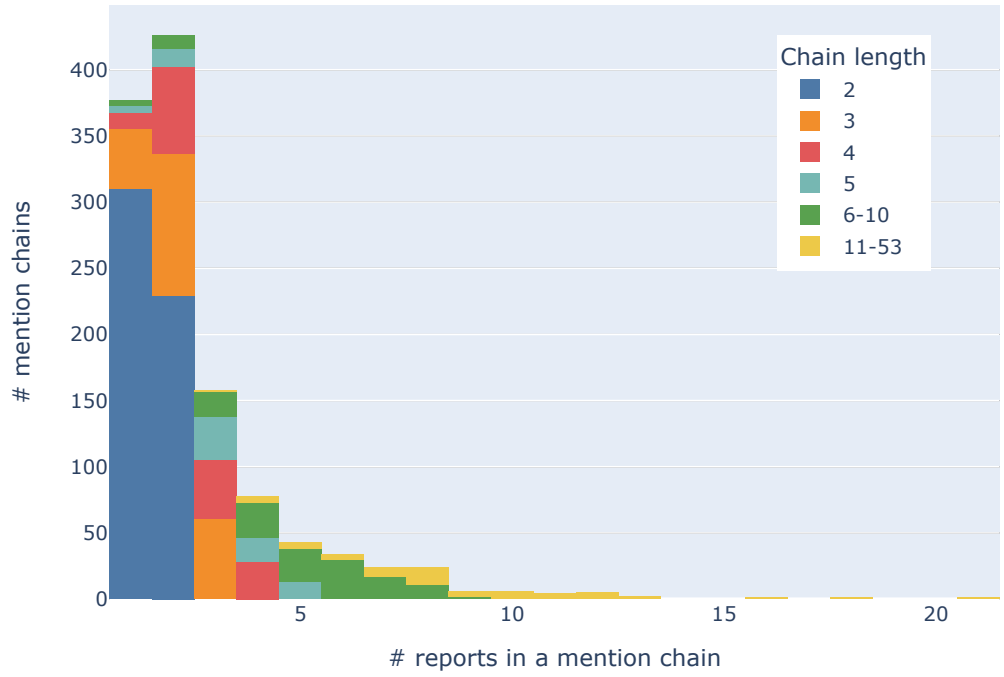

Figure 9.2: Coverage of reports in mention chains. The x-axis indicates the number of different reports of a patient covered in a mention chain whereas the y-axis indicates the actual number of mention chains.

25 times as many positive instances), we randomly sample negative instances such that there are equal instances of positive and negative pairs.

While forming the input sequence to BERT, we provide additional contextual information associated with the two mentions besides the mention spans. We incorporate anatomy and radiology modifier information surrounding a mention span in the sequence. This is grounded on the point that two finding mentions with the same name (e.g., fracture) are placed in separate chains based on their different anatomical locations (e.g., skull vs hip) or different associated modifiers (e.g., right vs left). For this, we leverage the Stanza python library (Qi et al., 2020) and use the clinical model package for identifying the observation, anatomy, and their corresponding identifiers. Specifically,

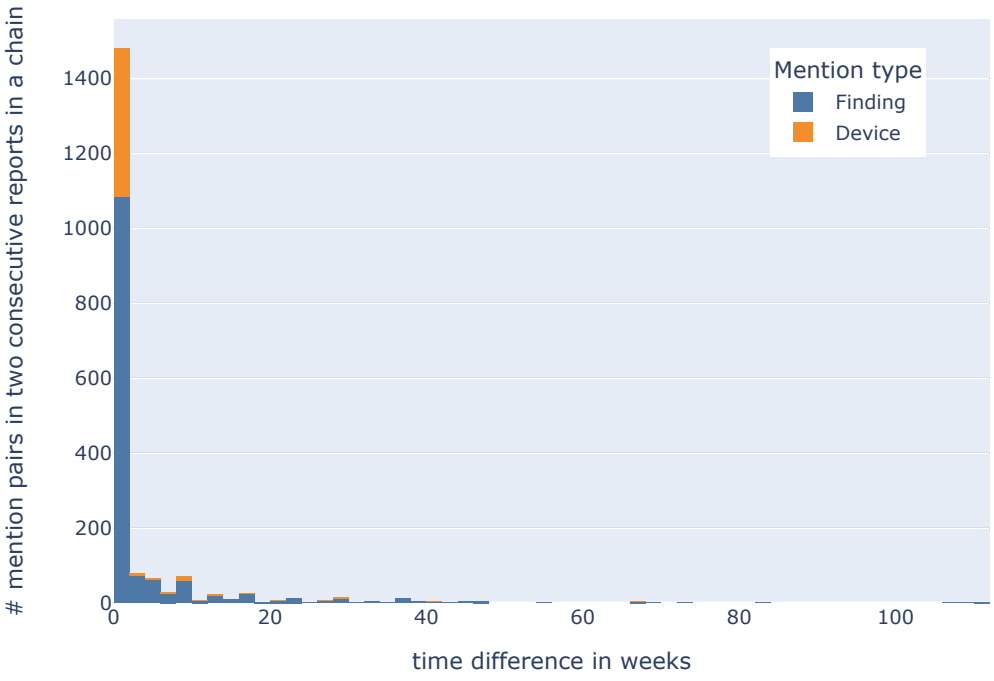

Figure 9.3: Time difference between two mentions annotated in two consecutive reports in a chain. Each bin denotes an interval of two weeks.

we apply the radiology named entity recognition (NER) model (Zhang et al., 2021) that was trained on radiology reports from three hospitals using a bi-directional LSTM character-level language model. We feed in the pretokenized text generated from NLTK to the Stanza NER pipeline.

9.4.2.2 Fine-tuning BERT

We fine-tune a $\text{BERT}_{\text{LARGE}}$ model to classify whether the two mentions in a pair are co-referring. We initialize the model parameters obtained by pre-training BERT on MIMIC-III clinical notes (Si et al., 2019). We frame our mention pair classification problem as a text pair classification task. First, we use only the mention spans of the two mentions to construct the BERT input as: [CLS]$m1$[SEP]$m2$[SEP], where $m1$ and $m2$ are the spans of the two mentions in a pair. Next, to provide additional information

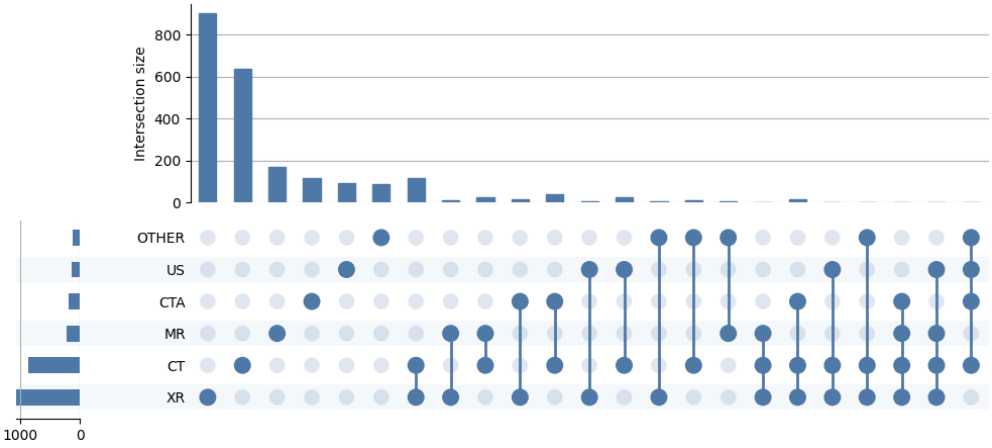

Figure 9.4: Distribution of imaging modalities in mention chains. XR - X-ray, CT - Computed Tomography, MR - Magnetic Resonance, CTA - CT Angiography, US - Ultrasound, OTHER - other modalities.

to the BERT model about both the mentions in a pair, we encode the anatomies as well as the anatomy and observation modifiers predicted by Stanza in the sentences containing the mentions. Following the standard BERT input format used in text pair classification configuration, we separate the information corresponding to the two mentions using the special [SEP] token, where anatomy and modifier information of each mention are delimited by a comma. Specifically, we include the Stanza-generated anatomy and modifiers in the left and right of a mention with an window size 5 in the order they appear in a sentence. We construct the BERT input sequence as follows for a mention pair:

[CLS]$m1$, $anty_i(m1)$, ...$anty_n(m1)$, $mod_i(m1)$, ...$mod_n(m1)$[SEP]$m2$, $anty_i(m2)$, ...$anty_n(m2)$, $mod_i(m2)$, ...$mod_n(m2)$[SEP]

Here, $anty_i(m1)$ refers to all the anatomy terms surrounding mention $m1$. Similarly, $mod_i(m2)$ refers to all the modifier terms surrounding mention $m2$.

The output corresponding to the [CLS] token is used to classify if the two mentions are co-

referring. The BERT classifier output is then processed to generate the mention chains. All the pairs for which BERT predicted as coreference positive are merged to form the coreference chains. Further, the predicted chain information is converted to CoNLL format for evaluation.

9.5 Evaluation

We evaluate the methods using gold mentions. We perform 5-fold cross validation to evaluate the performance of the BERT-based approach for CDCR. For each of the 5 iterations, our dataset of 60 patients are split into training, validation, and test sets in the ratio of 60, 20, and 20 %, respectively. The BERT classifier is applied to all possible mention pairs in the test sets. We report the results using the CR evaluation metrics–MUC, B^3, $CEAF_e$, the average F1 of these metrics i.e., CoNLL F1, and BLANC. MUC (Vilain et al., 1995) is a link-based evaluation metric that is based on the minimum number of coreference links required to translate from gold to predicted mention chains. B^3 (Bagga & Baldwin, 1998) is a mention-based metric where the evaluation uses the recall or precision of the individual mentions. For each mention in the gold chains, B^3 recall considers the fraction of the correct mentions that are included in the predicted chain containing that mention. The main assumption of CEAF (Luo, 2005) is that each gold chain should be mapped to only one response chain, and vice versa. BLANC (Recasens & Hovy, 2011; Luo et al., 2014) is another link-based metric where the recall and precision are calculated by averaging the recall and precision of coreference and non-coreference links.

We use the $BERT_{LARGE}$ cased model to classify the mention pairs. The model is pre-trained on MIMIC-III notes for 320K steps. We set the maximum sequence length at 128, learning rate at $2e$-5, and the number of training epochs at 4.

Table 9.5: 5-fold CV results of $\text{BERT}_{\text{LARGE}}$ models for classifying if two mentions in a pair are coreferring. P - Precision, R - Recall, Acc - Accuracy.

Model	P (%)	R (%)	F1	Acc
Mentions	44.83	85.89	58.91	95.53
+ Context	52.76	86.3	65.49	96.61

9.6 Results

We show the results of our BERT classification models in Table 9.5. We illustrate a few sample errors of the BERT classifiers in Table 9.6. In most of the false positive cases, we observe that the mention strings are the same and better learning of more broad context is required. The false negative errors indicate the need to incorporate more domain-specific knowledge. We then use the output of the BERT models to perform coreference resolution across reports. The cross-report coreference resolution results of the string matching baseline as well as both the BERT variants are in Table 9.7. We use the gold mention spans in this evaluation. Although the BERT classifier that uses context performs better than the one that uses only mention spans (as per the performance measures in Table 9.5), we see that the CDCR performance of the latter is better for all metrics. We also observe that the recall values of MUC, B^3, and BLANC are higher for the BERT (mentions) model than the string-matching method that has better precision values (the case is reverse for CEAF_e).

9.7 Discussion

We create an annotated cross-document coreference resolution (CDCR) dataset in the radiology domain to track the same radiological findings and medical devices across all reports of a patient and apply BERT-based baseline method to perform CDCR. The task of CDCR is relatively under-explored in the clinical domain, and in this work we propose a sufficiently large dataset with an

Table 9.6: Common error types of BERT classification models. FP - False Positive, FN - False Negative.

Mention pairs	Corresponding sentences	Category	Reason
NG tube; NG tube	**Report-5** Compared with prior radiograph, an NG tube has been withdrawn and there is significant dilatation of the colon lying just below the right hemidiaphragm; **Report-10** An *NG tube* terminates with its tip in the stomach	FP	More deep understanding of context is required (e.g., *"withdrawn"* in Report-5 indicates that the NG tube in Report-10 is different from the first one). Sufficient contextual information is not incorporated into the models
thrombus; thrombus	**Report-3** The grayscale ultrasound of the veins of the upper extremities demonstrated filling defect in the right cephalic vein at the level of the antecubital fossa consistent with *thrombus*; **Report-13** No intraluminal *thrombus* is identified		
collapse; atelectases	**Report-1** There is increased retrocardiac density, consistent with left lower lobe *collapse* and/or consolidation; **Report-20** There is cardiomegaly with *atelectases* in the left upper lobe as well as atelectasis in the left lower lobe.	FN	More domain knowledge understanding is required to link the correlated findings
hemorrhage; hematoma	**Report-1** There is no intraparenchymal *hemorrhage* identified; **Report-6** There is a small left frontal subdural *hematoma*, slightly larger than prior CT studies		

average of 10.6 reports per patient (compared to previous 3 notes per patient in Wright-Bettner et al. (2019)). Additionally, this is the first CDCR dataset in radiology.

The results in Tables 9.5 and 9.7 indicate that there is enough scope for performance improvement. A brief analysis of the output from the BERT classifiers suggests that incorporating rich radiology-specific domain knowledge will be useful in improving CDCR systems. For example, there is potential in encoding knowledge about relations between different human anatomies, knowledge about clinical correlation between various radiological findings (e.g., 'consolidation' and 'pneumonia'), and information about findings that are more often coreferred across different imaging modalities. Another promising avenue is allowing the model to learn more broad cross-report context (e.g.,

Table 9.7: CDCR performances. Precision - P %, Recall - R %. 5-fold cross validation results are reported for BERT models.

Methods	MUC			B^3			$CEAF_e$			CoNLL	BLANC		
	P	R	F1	P	R	F1	P	R	F1	F1	P	R	F1
String match	80.18	70.36	74.95	83.52	70.75	76.6	61.88	76.46	68.4	73.32	78.52	71.6	74.44
BERT (mentions)	68.56	95.31	79.65	40.84	86.57	55.23	76.53	23.46	35.84	56.91	53.39	77.18	50.29
BERT (mentions + context)	67.46	92.58	77.87	32.72	85.7	46.48	76.12	18.97	29.99	51.45	49.79	67.92	39.13

by leveraging certain language patterns in the reports suggesting any potential coreference such as 'compared to previous study'). We also intend to investigate the impact of BERT classifier output on the various CDCR evaluation metrics in detail.

An interesting method to explore for CDCR model development using this annotated dataset is by adopting the recently proposed cross-document language modeling technique that uses a new pre-training approach that has shown to be effective for several multi-document downstream tasks including CDCR and multihop question answering (Cattan et al., 2021a). The pre-training technique considers two main ideas: pre-training over sets of multiple related documents and usage of dynamic global attention pattern over masked tokens. This pre-training approach can be used to develop a CDCR system similar to the CDCR pairwise scoring framework proposed in a recent work (Caciularu et al., 2021). Here, we can feed the whole radiology reports corresponding to the two mentions in a pair into the CDLM rather than feeding only the local context of the mentions (e.g., surrounding words of a mention). Future work can focus on building an end-to-end CDCR system where the predicted mention spans are used to infer the mention chains instead of the gold mentions, although this relies on a robust extraction system to identify the radiological entities accurately (which is oftentimes challenged by the presence of different modifier terms described in conjunction with the main finding terms). From the clinical application perspective, this dataset can be extended to cancer domain that demands long-term tracking of findings (e.g., tumor, cyst).

10

Conclusion

This dissertation has ranged from representing detailed spatial information in radiology reports to building natural language processing methods for automatic identification of such information, and applying this extracted information for important clinical use cases. This concluding chapter summarizes the key findings and lays out the limitations and future work of these studies as a whole.

10.1 Key Findings

Our proposed spatial representation frameworks capture different spatial information of clinical importance in radiology reports. The basic schema, Rad-SpRL, captures four important spatial roles in the context of a spatial indicator (that denotes the presence of a spatial relation between clinical findings and body locations) including probable diagnoses. The advanced schema, Rad-

SpatialNet, captures a wider variety of spatial and contextual information such as the relative position of a finding with respect to a location and the size of a finding.

We employed transformer language models based on both sequence labeling and question answering for extracting the spatial information. For Rad-SpRL, the sequence labeling models based on BERT and XLNet achieve satisfactory performance with the highest average F1 measure of 91.29 for extracting the indicators and 85.7, 89.3, 79.0, and 78.6 for identifying TRAJECTOR, LANDMARK, DIAGNOSIS, and HEDGE roles, respectively using the predicted INDICATORS. For Rad-SpatialNet, the performance of BERT-based sequence labeling models is decent, with F1 of 77.89 for spatial trigger extraction and an overall F1 of 81.61 and 66.25 across all frame elements using gold and predicted spatial triggers, respectively. We also frame the problem of extracting fine-grained radiology spatial information (annotated as per Rad-SpatialNet representation) as two-turn question answering (QA). This approach outperforms traditional transformer-based sequence labeling in extracting both spatial triggers and their corresponding spatial frame elements. The average F1 score for identifying spatial triggers is 90.07 and the average F1 scores for identifying important frame elements like Figure and Ground are 78.13 and 83.77, respectively. Here we see promising improvements of 12, 13, and 12 points in the average F1 scores for identifying the spatial triggers, Figure, and Ground frame elements, respectively when compared to a traditional sequence tagging method. This demonstrates the advantages that QA provides over sequence labeling for information extraction (IE). This is the first work to employ a multi-turn QA approach for granular IE both in the radiology as well as spatial IE domains.

Second, after text is extracted from radiology reports it needs to be grounded in formal terminologies or ontologies to support reasoning. Our findings here include a system that mapped radiology concepts to RadLex. We constructed a manually annotated normalization corpus in the domain of radiology and this is the first attempt to normalize diverse radiological entities to RadLex concepts. We proposed two BERT-based models where we configured BERT for the normalization task as a

re-ranker as well as a span detector. We obtain satisfactory results by fine-tuning the BERT models on our annotated dataset with the span detector model achieving an accuracy of 78.44% in cross validation.

Next, in order to alleviate the reliance on human annotations for creating training data, we proposed a weak supervision approach to automatically create radiology training data for spatial IE. This is based on *data programming* that uses rules (or labeling functions) relying on domain-specific dictionaries and radiology language characteristics to generate weak labels. These weak labels were then used to fine-tune a pre-trained BERT model. Our weakly supervised BERT model provide satisfactory results in extracting spatial relations without using any manual annotations for training (spatial trigger F1: 72.89, relation F1: 52.47). To our knowledge, this is the first work to automatically create detailed weak labels corresponding to crucial radiological information. Our data programming approach is 1) adaptable as the labeling functions can be updated with relatively little manual effort to incorporate more variations in radiology language reporting formats and 2) generalizable as these functions can be applied across multiple radiology sub-domains in most cases. We demonstrate that a weakly supervised model performs sufficiently well in identifying a variety of relations from radiology text without manual annotations, while exceeding state-of-the-art results when annotated data is available.

Finally, we use the clincial information extracted from the reports to develop phenotyping and tracking applications. First, we use the output of our spatial IE system based on Rad-SpatialNet to classify complex ischemic stroke phenotypes. We demonstrate that a generalizable and fine-grained representation schema like Rad-SpatialNet could be utilized for determining detailed phenotypes that often requires information about various related radiological entities (such as findings, brain locations, and diagnoses). Our phenotypes are based on specific brain regions affected by stroke. We show that satisfactory results can be achieved by applying simple domain rules on top of the IE system's output to classify the phenotypes. Second, we develop an automated tracking system

that can track the same radiological findings and medical devices across multiple reports of a patient over time. For this, we constructed a manually annotated cross-document coreference resolution (CDCR) dataset. We applied two baseline methods to automatically identify the cross-report coreferences. The performance of these methods are low to moderate, highlighting the challenging nature of this task and our dataset.

10.2 Limitations and Future Work

The task of relation extraction can be broken down broadly into three categories - easy, medium, and hard based on its difficulty level. Our work for spatial IE covers the intra-sentence relations that fall into the medium category. The easy category includes intra-noun phrase relations (e.g., 'lung cancer') which are already captured to some extent in existing ontologies such as Unified Medical Language System (UMLS). We do not cover such relations in the Rad-SpatialNet schema, but included these cases when we extended the schema to the ophthalmology domain. The hard category consists of inter-sentence relations, which is not covered in this work. Let's consider the following example:

There is profoundly reduced blood flow in the temporal lobe. This finding suggests ***infarction****.*

Here we see that the potential diagnosis 'infarction' is documented in a sentence that follows the sentence containing the spatial relation between 'reduced blood flow' and 'temporal lobe'. This is a relatively hard NLP problem and forms an interesting direction for future research.

More complex scenarios where two spatial relations are spatially related can also be handled in the future. Let's take the example below:

There is effacement of sulci ***adjacent to*** *the hypodensity in subcortical region.*

Here, 'adjacent to' is a spatial expression that connects two spatial relations–'effacement of sulci' and ' hypodensity in subcortical region'. Oftentimes, the same radiological findings are mentioned

multiple times in the same report, e.g., once in the 'Findings' section and again in the 'Impressions/Conclusions' section. We do not map or link these same entities as part of Rad-SpatialNet. This can also be included in the future.

We notice that one of the main reasons behind the low performance of our information extraction methods for identifying some frame elements (e.g., Reason, Density descriptor, etc.) is that those elements exist very infrequently in our annotated dataset. More instances of these frame elements can be annotated in the future. Rad-SpatialNet also holds potential to extend to other domains like pathology that could aim to extract spatial information from pathology reports. There is also a need to evaluate the generalizability of our proposed sequence labeling and question answering-based extraction methods on multi-institutional datasets and also across reports belonging to other imaging modalities (e.g., ultrasound, computed tomography angiography, etc.).

For the downstream applications, there is a scope for developing various disease-specific phenotyping applications. For example, the spatial information from the radiology reports could potentially be used to classify important phenotypes for Alzheimer's disease. For instance, if certain radiological findings like 'volume loss' is present in the 'precuneus' part of 'medial parietal lobe', this would imply that the Alzheimer's disease is at its early stage. Thus important phenotypes based on disease stage could be classified using the spatial information from the reports. The fine-grained spatial information extracted by our proposed methods could also serve as fine-grained labels for the corresponding medical images. These labels can be used to train deep image classification models in the future. For automated tracking of radiological findings, our constructed annotated dataset can be augmented with more cancer-specific findings that demand long-term tracking and, therefore, could be leveraged for longitudinal tracking of cancer or tumor-related findings. Another potential research avenue is to develop more advanced tracking methods that can incorporate more domain knowledge understanding and broad cross-report context understanding.

A

Labeling functions developed for our weak supervision approach to identify spatial information

Table A.1: Heuristics used in the labeling functions to identify the spatial frame elements. FE - Frame Element. LF - Labeling Function. RadEnt - Radiological Entity. SpTrg - Spatial Trigger.

Spatial FE	**Heuristics**

Figure (LF 1)	anatomies to ignore = [*side*, *region*, *portion*, *part*, *territory*, *fragment*, *margin*, *site*, *aspect*, *division*, *area*, *branch*];
	anatomy related terms = all anatomies - anatomies to ignore
	RadEnt is neither relative position nor position status nor hedge AND
	• IF SpTrg is any of [*with*|*without*|*show(s)*|*demonstrate(s)*|*is*|*are*|*reveal(s)*] AND RadEnt lies to the right of SpTrg AND RadEnt is finding
	• ELSE IF RadEnt lies to the left of SpTrg AND any of [*with tip*|*with its tip*|*with the tip*] does not follow RadEnt AND any of [*tip of* |*tip of the*] does not precede RadEnt AND
	  • IF no preposition-containing hedge term between SpTrg and RadEnt AND the other trigger term between SpTrg and RadEnt is *'of'* AND RadEnt is *'tip'*
	  • ELSE IF no preposition-containing hedge term between SpTrg and RadEnt AND no additional spatial trigger between SpTrg and RadEnt AND RadEnt is not an anatomy-related term
	  • ELSE IF 0 word in between SpTrg and RadEnt AND RadEnt is not an anatomy-related term

Figure (LF 2)	• RadEnt lies to the left of SpTrg AND RadEnt belongs to acronyms dictionary
Figure (LF 3)	specific terms = [*collapsed*, *engorged*, *widened*, *calcified*, *unfolded*, *occluded*, *prominent*, *inflated*, *hypoinflated*, *hyperinflated*, *aerated*, *hyperaerated*, *hypoaerated*, *narrowed*] • RadEnt lies to the right of SpTrg AND SpTrg is any of [*is*|*are*] AND RadEnt belongs to specific terms list
Ground (LF 1)	• SpTrg is any of [*with*|*without*|*show(s)*|*demonstrate(s)*|*is*|*are*|*reveal(s)*] AND RadEnt lies directly adjacent to the left of the SpTrg • For other SpTrgs, RadEnt lies to the right of SpTrg AND there is 0-2 words in between SpTrg and RadEnt AND RadEnt is anatomy
Ground (LF 2)	RadEnt lies to the right of SpTrg AND • 0 word in between SpTrg and RadEnt AND RadEnt is anatomy • For greater than 0 word in between SpTrg and RadEnt, no other trigger in between AND RadEnt is anatomy

Ground (LF 3)	*Regular expressions used for matching specific anatomy patterns*: anatomy with segment = "[A-Z][0-9]{1,2}[\-|\/][A-Z]?[0-9]{1,2}" anatomy with segment without hyphen = "[A-Z][0-9]{1,2}" anatomy segment body = "[A-Z][0-9]{1}\s{1,}body" RadEnt lies to the right of SpTrg AND • 0 word in between SpTrg and RadEnt AND RadEnt matches any of the three anatomy patterns • For greater than 0 word in between SpTrg and RadEnt, no other trigger in between AND RadEnt matches any of the three anatomy patterns
Ground (LF 4)	specific diseases = [*hmd*, *hyaline membrane disease*, *ards*, *acute respiratory distress syndrome*, *rds*, *respiratory distress syndrome*] • RadEnt lies to the right of SpTrg AND RadEnt belongs to acronyms dictionary except for the terms in specific diseases list
Diagnosis (LF 1)	RadEnt is finding AND text span to the right of RadEnt is '.' AND • IF preposition-containing hedge term between SpTrg and RadEnt • ELSE IF a hedge term present to the left of the RadEnt with window length 4 and no additional spatial trigger between SpTrg and RadEnt

Diagnosis (LF 2)	left window = [*represent*, *suggest*, *indicat*, *consistent with*]; right window = [*ruled out*, *excluded*, *vs*, *versus*] • any item in left window list present to the left of RadEnt with window length 4 AND RadEnt is finding • any item in right window list present to the right of RadEnt with window length 4 AND RadEnt is finding
Diagnosis (LF 3)	specific diseases = [*hmd*, *hyaline membrane disease*, *ards*, *acute respiratory distress syndrome*, *rds*, *respiratory distress syndrome*] • RadEnt lies to the right of SpTrg AND RadEnt belongs to specific diseases list AND text span to the right of RadEnt is '.'
Diagnosis (LF 4)	• RadEnt lies to the right of SpTrg AND no additional spatial trigger or hedge term to the right of RadEnt AND RadEnt is finding
Diagnosis (LF 5)	RadEnt lies to the right of SpTrg AND RadEnt is finding • IF preposition-containing hedge term between SpTrg and RadEnt • ELSE IF a hedge term present between SpTrg and RadEnt and no additional spatial trigger between SpTrg and RadEnt
Hedge (LF 1)	• RadEnt lies to the right of SpTrg AND RadEnt is a hedging-related term AND a finding term is present to the right of the RadEnt

Hedge (LF 2)	RadEnt lies to the left of SpTrg AND RadEnt is a hedging-related term AND • 0 word in between SpTrg and RadEnt OR a finding term is present between SpTrg and RadEnt

Distance (LF 1)	device related = [*tube*, *catheter*, *ett*, *tip*, *port*, *lead*, *device*, *drain*, *screw*]
	Regular expressions used for matching distance-related entities:
	distance first pattern = (\d+\.( )?\d+|\d+( )?\.\d+|\.\d+|\d+) *([\-]\(mm|cm|millimeter(s)?|centimeter(s)?\)(?![a-z/]));
	distance second pattern = (\d+\.( )?\d+|\d+( )?\.\d+|\.\d+|\d+) *(\-? *(mm|cm|millimeter(s)?|centimeter(s)?\)(?![a-z/]));
	distance third pattern = \b(few|some)\b\s{1,}\b(mm|mms|millimeter|millimeters|cm|cms|centimeter|centimeters)\b
	• any of the terms in device related list is present in the sentence containing RadEnt AND RadEnt matches any of the three distance patterns

Position Status (LF 1)	device related = [*tube*, *catheter*, *ett*, *tip*, *port*, *lead*, *device*, *drain*, *screw*] • RadEnt is a position status-related term AND any of the terms in device related list is present in the sentence containing RadEnt
Relative Position (LF 1)	• RadEnt is a relative position-related term AND the next or preceding word of RadEnt is contained in any of the terms in anatomies dictionary

Reason (LF 1)	• RadEnt is finding AND any reason associated hedge term is present to the left of RadEnt with window length 4
Associated Process (LF 1)	• 0 word in between SpTrg and RadEnt AND RadEnt is an associated process-related term • For greater than 0 word in between SpTrg and RadEnt, no other trigger in between AND RadEnt is an associated process-related term

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467*.

Alex, B., Grover, C., Tobin, R., Sudlow, C., Mair, G., & Whiteley, W. (2019). Text mining brain imaging reports. *Journal of Biomedical Semantics*, 10(1), 23.

Annarumma, M., Withey, S. J., Bakewell, R. J., Pesce, E., Goh, V., & Montana, G. (2019). Automated Triaging of Adult Chest Radiographs with Deep Artificial Neural Networks. *Radiology*, 291(1), 196–202.

Apostolova, E., Tomuro, N., Mongkolwat, P., & Demner-Fushman, D. (2012). Domain Adaptation of Coreference Resolution for Radiology Reports. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing* (pp. 118–121).

Aronson, A. R. & Lang, F.-M. (2010). An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3), 229–236.

Badene, S., Thompson, K., Lorré, J.-P., & Asher, N. (2019). Data Programming for Learning Discourse Structure. In *Proceedings of the 57th Annual Meeting of the Association for*

Computational Linguistics (pp. 640–645).

Bagga, A. & Baldwin, B. (1998). Algorithms for Scoring Coreference Chains. In *In The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference* (pp. 563–566).

Baker, C. F. (2014). FrameNet : A Knowledge Base for Natural Language Processing. In *Proceedings ofFrame Semantics in NLP: A Workshop in Honor ofChuck Fillmore*, number 1968 (pp. 1–5).

Banerjee, I., Li, K., Seneviratne, M., Ferrari, M., Seto, T., Brooks, J. D., Rubin, D. L., & Hernandez-Boussard, T. (2019). Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. *JAMIA Open*, 2(1), 150–159.

Banerjee, P., Pal, K. K., Devarakonda, M., & Baral, C. (2020). Knowledge Guided Named Entity Recognition for BioMedical Text. *arXiv:1911.03869 [cs]*.

Barhom, S., Shwartz, V., Eirew, A., Bugert, M., Reimers, N., & Dagan, I. (2019). Revisiting Joint Modeling of Cross-document Entity and Event Coreference Resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4179–4189).

Bastianelli, E., Croce, D., Basili, R., & Nardi, D. (2013). UNITOR-HMM-TK: Structured Kernel-based learning for Spatial Role Labeling. In *Second Joint Conference on Lexical*

*and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 573–579).

Baughman, D. M., Su, G. L., Tsui, I., Lee, C. S., & Lee, A. Y. (2017). Validation of the Total Visual Acuity Extraction Algorithm (TOVA) for Automated Extraction of Visual Acuity Data From Free Text, Unstructured Clinical Records. *Translational Vision Science & Technology*, 6(2), 2.

Birchall, D. (2015). Spatial ability in radiologists: A necessary prerequisite? *British Journal of Radiology*, 88(1049), 6–8.

Birenbaum, D., Bancroft, L. W., & Felsberg, G. J. (2011). Imaging in Acute Stroke. *Western Journal of Emergency Medicine*, 12(1), 67–76.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): Integrating biomedical terminology. *Nucleic Acids Res*, 32(Database issue), D267–D270.

Bozkurt, S., Alkim, E., Banerjee, I., & Rubin, D. L. (2019). Automated Detection of Measurements and Their Descriptors in Radiology Reports Using a Hybrid Natural Language Processing Algorithm. *Journal of Digital Imaging*, 32(4), 544–553.

Bradshaw, T., Weisman, A., Perlman, S., & Cho, S. (2020). Automatic image classification using labels from radiology text reports: Predicting Deauville scores. *Journal of Nuclear Medicine*, 61(supplement 1), 1410–1410.

Brady, A. P., Bello, J. A., Derchi, L. E., Fuchsjäger, M., Goergen, S., Krestin, G. P., Lee, E. J. Y., Levin, D. C., Pressacco, J., Rao, V. M., Slavotinek, J., Visser, J. J., Walker, R. E. A., & Brink, J. A. (2021). Radiology in the Era of Value-based Healthcare: A Multi-Society Expert Statement from the ACR, CAR, ESR, IS3R, RANZCR, and RSNA. *Radiology*, 298(3), 486–491.

Brown, E. G., Wood, L., & Wood, S. (1999). The Medical Dictionary for Regulatory Activities (MedDRA). *Drug-Safety*, 20(2), 109–117.

Bugert, M., Reimers, N., & Gurevych, I. (2021). Generalizing Cross-Document Event Coreference Resolution Across Multiple Corpora. *arXiv:2011.12249 [cs]*.

Bustos, A., Pertusa, A., Salinas, J.-M., & de la Iglesia-Vayá, M. (2019). PadChest: A large chest x-ray image dataset with multi-label annotated reports. *arXiv preprint arXiv:1901.07441*.

Caciularu, A., Cohan, A., Beltagy, I., Peters, M., Cattan, A., & Dagan, I. (2021). CDLM: Cross-Document Language Modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021* (pp. 2648–2662).

Callahan, A., Fries, J. A., Ré, C., Huddleston, J. I., Giori, N. J., Delp, S., & Shah, N. H. (2019). Medical device surveillance with electronic health records. *npj Digital Medicine*, 2(1), 1–10.

Candemir, S., Rajaraman, S., Thoma, G., & Antani, S. (2018). Deep learning for grading cardiomegaly severity in chest x-rays: An investigation. In *IEEE Life Sciences Conference (LSC)* (pp. 109–113).: IEEE.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., & Dagan, I. (2021a). Cross-document Coreference Resolution over Predicted Mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 5100–5107).

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., & Dagan, I. (2021b). Realistic Evaluation Principles for Cross-document Coreference Resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics* (pp. 143–151).

Cattan, A., Johnson, S., Weld, D., Dagan, I., Beltagy, I., Downey, D., & Hope, T. (2021c). SciCo: Hierarchical Cross-Document Coreference for Scientific Concepts. *arXiv:2104.08809 [cs]*.

Chang, A., Savva, M., & Manning, C. D. (2014). Learning Spatial Knowledge for Text to 3D Scene Generation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2028–2038).: Association for Computational Linguistics.

Chang, E., Demberg, V., & Marin, A. (2021). Jointly Improving Language Understanding and Generation with Quality-Weighted Weak Supervision of Automatic Labeling. In *EACL*.

Cheng Bastian, Forkert Nils Daniel, Zavaglia Melissa, Hilgetag Claus C., Golsari Amir, Siemonsen Susanne, Fiehler Jens, Pedraza Salvador, Puig Josep, Cho Tae-Hee, Alawneh Josef, Baron

Jean-Claude, Ostergaard Leif, Gerloff Christian, & Thomalla Götz (2014). Influence of Stroke Infarct Location on Functional Outcome Measured by the Modified Rankin Scale. *Stroke*, 45(6), 1695–1702.

Collell, G. & Moens, M.-F. (2018). Learning Representations Specialized in Spatial Knowledge: Leveraging Language and Vision. *Transactions of the Association for Computational Linguistics*, 6, 133–144.

Cornegruta, S., Bakewell, R., Withey, S., & Montana, G. (2016). Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. In *Proceedings of the Seventh International Workshop on Health Text Mining and Information Analysis* (pp. 17–27).

Corry, C. (2011). The future of recruitment and selection in radiology. Is there a role for assessment of basic visuospatial skills? *Clinical Radiology*, 66(5), 481–483.

Coyne, B. & Sproat, R. (2001). WordsEye: An automatic text-to-scene conversion system. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques* (pp. 487–496).

Coyne, B., Sproat, R., & Hirschberg, J. (2010). Spatial relations in text-to-scene conversion. In *Computational Models of Spatial Language Interpretation, Workshop at Spatial Cognition*, volume 620 (pp. 9–16).

Cusick, M., Adekkanattu, P., Campion, T. R., Sholle, E. T., Myers, A., Banerjee, S., Alexopoulos, G., Wang, Y., & Pathak, J. (2021). Using weak supervision and deep learning to classify clinical notes for identification of current suicidal ideation. *Journal of Psychiatric Research*, 136, 95–102.

Daniels, Z. A. & Metaxas, D. N. (2019). Exploiting Visual and Report-Based Information for Chest X-Ray Analysis by Jointly Learning Visual Classifiers and Topic Models. In *IEEE 16th International Symposium on Biomedical Imaging (ISBI).*

Datta, S., Godfrey-Stovall, J., & Roberts, K. (2021a). RadLex Normalization in Radiology Reports. *AMIA Annual Symposium Proceedings*, 2020, 338–347.

Datta, S., Khanpara, S., Riascos, R. F., & Roberts, K. (2021b). Leveraging Spatial Information in Radiology Reports for Ischemic Stroke Phenotyping. *AMIA Summits on Translational Science Proceedings*, 2021, 170–179.

Datta, S., Lam, H. C., Pajouhi, A., Mogalla, S., & Roberts, K. (2022). A Cross-document Coreference Dataset for Longitudinal Tracking across Radiology Reports. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3686–3695).

Datta, S. & Roberts, K. (2020). A Hybrid Deep Learning Approach for Spatial Trigger Extraction from Radiology Reports. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2020, 50–55.

Datta, S. & Roberts, K. (2022). Fine-grained spatial information extraction in radiology as two-turn question answering. *International Journal of Medical Informatics*, 158, 104628.

Datta, S., Si, Y., Rodriguez, L., Shooshan, S. E., Demner-Fushman, D., & Roberts, K. (2020a). 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.

Datta, S., Ulinski, M., Godfrey-Stovall, J., Khanpara, S., Riascos-Castaneda, R. F., & Roberts, K. (2020b). Rad-SpatialNet: A Frame-based Resource for Fine-Grained Spatial Relations in Radiology Reports. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 2251–2260).

Demner-Fushman, D., Kohli, M. D., Rosenman, M. B., Shooshan, S. E., Rodriguez, L., Antani, S., Thoma, G. R., & McDonald, C. J. (2016). Preparing a collection of radiology examinations for distribution and retrieval. *Journal of the American Medical Informatics Association*, 23(2), 304–310.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4171–4186).: Association for Computational Linguistics.

Division for Heart Disease and Stroke Prevention (2020). Stroke Facts | cdc.gov.

Doğan, R. I., Leaman, R., & Lu, Z. (2014). NCBI Disease Corpus: A Resource for Disease Name Recognition and Concept Normalization. *J Biomed Inform*, 47, 1–10.

Dong, H., Suárez-Paniagua, V., Zhang, H., Wang, M., Whitfield, E., & Wu, H. (2021). Rare Disease Identification from Clinical Notes with Ontologies and Weak Supervision. *arXiv:2105.01995 [cs]*.

Dua, S., Baldini, I., Katz-Rogozhnikov, D. A., van der Veen, E., Britt, A., Mangalath, P., Kleiman, L. B., & Fitz, C. D. V. (2021). Biomedical Corpus Filtering: A Weak Supervision Paradigm With Infused Domain Expertise. In *SDU@AAAI*.

Dunnmon, J. A., Ratner, A. J., Saab, K., Khandwala, N., Markert, M., Sagreiya, H., Goldman, R., Lee-Messer, C., Lungren, M. P., Rubin, D. L., & Ré, C. (2020). Cross-Modal Data Programming Enables Rapid Medical Machine Learning. *Patterns (New York, N.Y.)*, 1(2).

Eyuboglu, S., Angus, G., Patel, B. N., Pareek, A., Davidzon, G., Long, J., Dunnmon, J., & Lungren, M. P. (2021). Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. *Nature Communications*, 12(1), 1880.

Fasola, J. & Mataric, M. J. (2013). Using semantic fields to model dynamic spatial relations in a robot architecture for natural language instruction of service robots. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems* (pp. 143–150).

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., & Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association: JAMIA*, 1(2), 161–174.

Friedman, C., Johnson, S. B., Forman, B., & Starren, J. (1995). Architectural requirements for a multipurpose natural language processor in the clinical environment. *Proceedings of the Annual Symposium on Computer Application in Medical Care*, (pp. 347–351).

Friedman, C., Shagina, L., Lussier, Y., & Hripcsak, G. (2004). Automated Encoding of Clinical Documents Based on Natural Language Processing. *Journal of the American Medical Informatics Association : JAMIA*, 11(5), 392–402.

Fries, J., Wu, S., Ratner, A., & Ré, C. (2017). SwellShark: A Generative Model for Biomedical Named Entity Recognition without Labeled Data. *arXiv:1704.06360 [cs]*.

Fries, J. A., Steinberg, E., Khattar, S., Fleming, S. L., Posada, J., Callahan, A., & Shah, N. H. (2021a). Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1), 2017.

Fries, J. A., Steinberg, E., Khattar, S., Fleming, S. L., Posada, J., Callahan, A., & Shah, N. H. (2021b). Ontology-driven weak supervision for clinical entity classification in electronic health records. *Nature Communications*, 12(1), 2017.

Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K. J., Shen, F., Wang, L., Wang, Y., Wen, A., Zhao, Y., Sohn, S., & Liu, H. (2020). Development of Clinical Concept Extraction Applications: A Methodology Review. *arXiv*, 1910.11377 [cs].

Fu, S., Leung, L. Y., Wang, Y., Raulli, A.-O., Kallmes, D. F., Kinsman, K. A., Nelson, K. B., Clark, M. S., Luetmer, P. H., Kingsbury, P. R., Kent, D. M., & Liu, H. (2019). Natural Language Processing for the Identification of Silent Brain Infarcts From Neuroimaging Reports. *JMIR Medical Informatics*, 7(2).

Garg, R., Oh, E., Naidech, A., Kording, K., & Prabhakaran, S. (2019). Automating Ischemic Stroke Subtype Classification Using Machine Learning and Natural Language Processing. *J Stroke Cerebrovasc*, 28(7), 2045–2051.

Govindarajan, P., Soundarapandian, R. K., Gandomi, A. H., Patan, R., Jayaraman, P., & Manikandan, R. (2020). Classification of stroke disease using machine learning algorithms. *Neural Computing and Applications*, 32(3), 817–828.

Guadarrama, S., Riano, L., Golland, D., Gohring, D., Jia, Y., Klein, D., Abbeel, P., & Darrell, T. (2013). Grounding spatial relations for human-robot interaction. In *IEEE International Conference on Intelligent Robots and Systems* (pp. 1640–1647).

Han, S., Tran, T., Rios, A., & Kavuluru, R. (2017). Team UKNLP: Detecting ADRs, Classifying Medication Intake Messages, and Normalizing ADR Mentions on Twitter. In *SMM4H@AMIA*. .

Hassanpour, S., Bay, G., & Langlotz, C. P. (2017). Characterization of Change and Significance for Clinical Findings in Radiology Reports Through Natural Language Processing. *Journal of Digital Imaging*, 30(3), 314–322.

Hassanpour, S. & Langlotz, C. P. (2016). Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine*, 66, 29–39.

Hayward, W. G. & Tarr, M. J. (1995). Spatial language and spatial representation. *Cognition*, 55(1), 39–84.

Huang, X., Fang, Y., Lu, M., Yao, Y., & Li, M. (2019). An Annotation Model on End-to-End Chest Radiology Reports. In *IEEE Access*, volume 7: IEEE.

Hui, C., Tadi, P., & Patti, L. (2020). Ischemic Stroke. In *StatPearls*. StatPearls Publishing.

Humbert-Droz, M., Mukherjee, P., & Gevaert, O. (2022). Strategies to Address the Lack of Labeled Data for Supervised Machine Learning Training With Electronic Health Records: Case Study for the Extraction of Symptoms From Clinical Notes. *JMIR Medical Informatics*, 10(3), e32903.

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., Seekins, J., Mong, D. A., Halabi, S. S., Sandberg, J. K., Jones, R., Larson, D. B., Langlotz, C. P., Patel, B. N., Lungren, M. P., & Ng, A. Y. (2019). CheXpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison. In *AAAI Conference on Artificial Intelligence*.

Ji, Z., Wei, Q., & Xu, H. (2019). BERT-based Ranking for Biomedical Entity Normalization. *arXiv*, 1908.03548 [cs].

Johns Hopkins Medicine (2022). Effects of Stroke.

Johnson, A. E., Pollard, T. J., Shen, L., wei H. Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., , & Mark, R. G. (2016). MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3, 160035.

Karimi, S., Metke-Jimenez, A., Kemp, M., & Wang, C. (2015). Cadec: A corpus of adverse drug event annotations. *Journal of Biomedical Informatics*, 55, 73–81.

Kergosien, E., Alatrista-Salas, H., Gaio, M., Güttler, F. N., Roche, M., & Teisseire, M. (2015). When textual information becomes spatial information compatible with satellite images. In *2015 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K)*, volume 01 (pp. 301–306).

Kim, C., Zhu, V., Obeid, J., & Lenert, L. (2019). Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. *PloS One*, 14(2), e0212778.

Kordjamshidi, P., Otterlo, M. V., & Moens, M.-F. (2010). Spatial Role Labeling : Task Definition and Annotation Scheme. In *Proceedings of the Language Resources & Evaluation Conference* (pp. 413–420).

Kordjamshidi, P., Rahgooy, T., & Manzoor, U. (2017). Spatial Language Understanding with Multimodal Graphs using Declarative Learning based Programming. In *Proceedings of the 2nd Workshop on Structured Prediction for Natural Language Processing* (pp. 33–43).

Kordjamshidi, P., Roth, D., & Moens, M.-F. (2015). Structured learning for spatial information extraction from biomedical text: Bacteria biotopes. *BMC Bioinformatics*, 16(1), 1–15.

Krasakis, A. M., Kanoulas, E., & Tsatsaronis, G. (2019). Semi-supervised Ensemble Learning with Weak Supervision for Biomedical Relationship Extraction. In *AKBC*.

Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., & Dyer, C. (2016). Neural Architectures for Named Entity Recognition. In *Proceedings of NAACL-HLT* (pp. 260–270).

Langlotz, C. P. (2006). RadLex: a new method for indexing online educational materials. *Radiographics*, 26(6), 1595–1597.

Laparra, E., Bethard, S., & Miller, T. A. (2020). Rethinking domain adaptation for machine learning over clinical language. *JAMIA Open*, 3(2), 146–150.

Leaman, R., Khare, R., & Lu, Z. (2015). Challenges in Clinical Natural Language Processing for Automated Disorder Normalization. *Journal of Biomedical Informatics*, 57, 28–37.

Levin, D. I. & Janiga, N. J. (2021). 2021 Outlook: Diagnostic Imaging Centers and Radiology Practices.

Levy, O., Seo, M., Choi, E., & Zettlemoyer, L. (2017). Zero-Shot Relation Extraction via Reading Comprehension. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)* (pp. 333–342).

Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R., & Pfister, H. (2014). UpSet: Visualization of Intersecting Sets. *IEEE transactions on visualization and computer graphics*, 20(12), 1983–1992.

Li, F., Peng, W., Chen, Y., Wang, Q., Pan, L., Lyu, Y., & Zhu, Y. (2020a). Event Extraction as Multi-turn Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 829–838).

Li, F., Zhang, M., Fu, G., & Ji, D. (2017). A neural joint model for entity and relation extraction from biomedical text. *BMC Bioinformatics*, 18(1), 198.

Li, X., Feng, J., Meng, Y., Han, Q., Wu, F., & Li, J. (2020b). A Unified MRC Framework for Named Entity Recognition. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5849–5859).

Li, X., Yin, F., Sun, Z., Li, X., Yuan, A., Chai, D., Zhou, M., & Li, J. (2019). Entity-Relation Extraction as Multi-Turn Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 1340–1350).

Li, Y., Shetty, P., Liu, L., Zhang, C., & Song, L. (2021). BERTifying the Hidden Markov Model for Multi-Source Weakly Supervised Named Entity Recognition. *ArXiv*.

Limsopatham, N. & Collier, N. (2016). Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In *Proceedings of the 54th Annual Meeting of the ACL* (pp. 1014–1023).

Lison, P., Hubin, A., Barnes, J., & Touileb, S. (2020). Named Entity Recognition without Labelled Data: A Weak Supervision Approach. In *ACL*.

Liu, J., Chen, Y., Liu, K., Bi, W., & Liu, X. (2020). Event Extraction as Machine Reading Comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1641–1651).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.

Luo, X. (2005). On Coreference Resolution Performance Metrics. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing* (pp. 25–32).

Luo, X., Pradhan, S., Recasens, M., & Hovy, E. (2014). An Extension of BLANC to System Mentions. *Proceedings of the conference. Association for Computational Linguistics. Meeting*, 2014, 24–29.

Luo, Y.-F., Sun, W., & Rumshisky, A. (2019a). A Hybrid Normalization Method for Medical Concepts in Clinical Narrative using Semantic Matching. *AMIA Jt Summits Transl Sci Proc*, 2019, 732–740.

Luo, Y.-F., Sun, W., & Rumshisky, A. (2019b). MCN: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics*, 92, 103132.

Mabotuwana, T., Hall, C. S., Hombal, V., Pai, P., Raghavan, U. N., Regis, S., McKee, B., Dalal, S., Wald, C., & Gunn, M. L. (2019). Automated Tracking of Follow-Up Imaging Recommendations. *AJR. American journal of roentgenology*, (pp. 1–8).

Mabotuwana, T., Hall, C. S., Tieder, J., & Gunn, M. L. (2018). Improving Quality of Follow-Up Imaging Recommendations in Radiology. *AMIA Annual Symposium Proceedings*, 2017, 1196–1204.

Majersik Jennifer J, Mowery Danielle, Zhang Mingyuan, Hill Brent, Cannon-Albright Lisa A, & Chapman Wendy (2018). Towards High-Precision Stroke Classification Using Natural Language Processing. *Stroke*, 49(Suppl_1), 92.

Mallory, E. K., de Rochemonteix, M., Ratner, A., Acharya, A., Re, C., Bright, R. A., & Altman, R. B. (2020). Extracting chemical reactions from text using Snorkel. *BMC Bioinformatics*, 21(1), 217.

Mani, I., Doran, C., Harris, D., Hitzeman, J., Quimby, R., Richer, J., Wellner, B., Mardis, S., & Clancy, S. (2010). SpatialML: Annotation scheme, resources, and evaluation. *Language Resources and Evaluation*, 44(3), 263–280.

Mbagwu, M., French, D. D., Gill, M., Mitchell, C., Jackson, K., Kho, A., & Bryar, P. J. (2016). Creation of an Accurate Algorithm to Detect Snellen Best Documented Visual Acuity from Ophthalmology Electronic Health Record Notes. *JMIR Medical Informatics*, 4(2), e14.

Miftahutdinov, Z. & Tutubalina, E. (2019). Deep Neural Models for Medical Concept Normalization in User-Generated Texts. In *Proceedings of the 57th Annual Meeting of the ACL: Student Research Workshop* (pp. 393–399).

Miller, T., Dligach, D., Bethard, S., Lin, C., & Savova, G. (2017). Towards generalizable entity-centric clinical coreference resolution. *Journal of Biomedical Informatics*, 69, 251–258.

Miwa, M. & Bansal, M. (2016). End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1105–1116).

Nogueira, R. & Cho, K. (2019). Passage Re-ranking with BERT. *arXiv*, 1901.04085 [cs].

Ong, C. J., Orfanoudaki, A., Zhang, R., Caprasse, F. P. M., Hutch, M., Ma, L., Fard, D., Balogun, O., Miller, M. I., Minnig, M., Saglam, H., Prescott, B., Greer, D. M., Smirnakis, S., & Bertsimas, D. (2020). Machine learning and natural language processing methods to

identify ischemic stroke, acuity and location from radiology reports. *PLOS ONE*, 15(6), e0234908.

Pattisapu, N., Anand, V., Patil, S., Palshikar, G., & Varma, V. (2020). Distant supervision for medical concept normalization. *Journal of Biomedical Informatics*, 109, 103522.

Peng, Y., Wang, X., Lu, L., Bagheri, M., Summers, R., & Lu, Z. (2018). NegBio: a high-performance tool for negation and uncertainty detection in radiology reports. In *AMIA Joint Summits on Translational Science Proceedings.*, volume 2018 (pp. 188–196).

Pesce, E., Withey, S. J., Ypsilantis, P.-P., Bakewell, R., Goh, V., & Montana, G. (2019). Learning to detect chest radiographs containing lung nodules using visual attention networks. *Medical Image Analysis*, 53, 26–38.

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the NAACL: Human Language Technologies* (pp. 2227–2237).

Peterson, K. J., Jiang, G., & Liu, H. (2020). A corpus-driven standardization framework for encoding clinical problems with HL7 FHIR. *Journal of Biomedical Informatics*, (pp. 103541).

Petruck, M. R. & Ellsworth, M. (2018). Representing Spatial Relations in FrameNet. In *Proceedings of the First International Workshop on Spatial Language Understanding* (pp. 41–45).

Pons, E., Braun, L. M., Hunink, M. M., & Kors, J. A. (2016). Natural Language Processing in Radiology: A Systematic Review. *Radiology*, 279(2).

Price, C., Seghier, M., & Leff, A. (2010). Predicting Language Outcome and Recovery After Stroke (PLORAS). *Nature reviews. Neurology*, 6(4), 202–210.

Pustejovsky, J. & Moszkowicz, J. L. (2008). Integrating Motion Predicate Classes with Spatial and Temporal Annotations. In *Coling 2008: Companion Volume: Posters* (pp. 95–98).: Coling 2008 Organizing Committee.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108).

Qu, C., Yang, L., Qiu, M., Croft, W. B., Zhang, Y., & Iyyer, M. (2019). BERT with History Answer Embedding for Conversational Question Answering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1133–1136).

Ratner, A., Bach, S. H., Ehrenberg, H., Fries, J., Wu, S., & Ré, C. (2020). Snorkel: Rapid training data creation with weak supervision. *The VLDB Journal*, 29(2), 709–730.

Recasens, M. & Hovy, E. (2011). Blanc: Implementing the rand index for coreference evaluation. *Natural Language Engineering*, 17(4), 485–510.

Rink, B., Roberts, K., Harabagiu, S., Scheuermann, R. H., Toomay, S., Browning, T., Bosler, T., & Peshock, R. (2013). Extracting actionable findings of appendicitis from radiology reports using natural language processing. In *AMIA Joint Summits on Translational Science Proceedings*, volume 2013 (pp. 221).

Roberts, K., Demner-Fushman, D., & Tonning, J. M. (2017). Overview of the TAC 2017 Adverse Reaction Extraction from Drug Labels Track. In *TAC*. .

Roberts, K., Rink, B., Harabagiu, S. M., Scheuermann, R. H., Toomay, S., Browning, T., Bosler, T., & Peshock, R. (2012). A machine learning approach for identifying anatomical locations of actionable findings in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2012 (pp. 779–788).

Roberts, K., Rodriguez, L., Shooshan, S., & Demner-Fushman, D. (2015). Automatic Extraction and Post-coordination of Spatial Relations in Consumer Language. In *AMIA Annual Symposium Proceedings*, volume 2015 (pp. 1083–1092).

Robertson, S. E., Walker, S., Jones, S., Hancock-Beaulieu, M. M., & Gatford, M. (1996). : (pp. 109–126).

Rosse, C. & Mejino, J. L. V. (2008). The Foundational Model of Anatomy Ontology. In A. Burger, D. Davidson, & R. Baldock (Eds.), *Anatomy Ontologies for Bioinformatics: Principles and Practice* (pp. 59–117).

Rubin, D. L., Willrett, D., O'Connor, M. J., Hage, C., Kurtz, C., & Moreira, D. A. (2014). Automated Tracking of Quantitative Assessments of Tumor Burden in Clinical Trials. *Translational Oncology*, 7(1), 23–35.

Safranchik, E., Luo, S., & Bach, S. H. (2020). Weakly Supervised Sequence Tagging from Noisy Rules. In *AAAI*.

Sarker, A., Belousov, M., Friedrichs, J., Hakala, K., Kiritchenko, S., Mehryary, F., Han, S., Tran, T., Rios, A., Kavuluru, R., de Bruijn, B., Ginter, F., Mahata, D., Mohammad, S. M., Nenadic, G., & Gonzalez-Hernandez, G. (2018). Data and systems for medication-related text classification and concept normalization from Twitter: Insights from the Social Media Mining for Health (SMM4H)-2017 shared task. *J Am Med Inform Assoc*, 25(10), 1274–1283.

Sedghi, E., Weber, J. H., Thomo, A., Bibok, M., & Penn, A. M. W. (2015). Mining clinical text for stroke prediction. *Network Modeling Analysis in Health Informatics and Bioinformatics*, 4(1), 16.

Sevenster, M., Van Ommering, R., & Qian, Y. (2012). Automatically correlating clinical findings and body locations in radiology reports using MedLEE. *Journal of Digital Imaging*, 25(2), 240–249.

Shang, J., Liu, L., Ren, X., Gu, X., Ren, T., & Han, J. (2018). Learning Named Entity Tagger using Domain-Specific Dictionary. *arXiv:1809.03599 [cs]*.

Shen, Z., Yi, Y., Bompelli, A., Yu, F., Wang, Y., & Zhang, R. (2021). Extracting Lifestyle Factors for Alzheimer's Disease from Clinical Notes Using Deep Learning with Weak Supervision. *arXiv:2101.09244 [cs]*.

Shi, Y., Zeng, Y., Wu, L., Liu, Z., Zhang, S., Yang, J., & Wu, W. (2017). A Study of the Brain Functional Network of Post-Stroke Depression in Three Different Lesion Locations. *Scientific Reports*, 7(1), 14795.

Shin, H.-C., Roberts, K., Lu, L., Demner-Fushman, D., Yao, J., & Summers, R. M. (2016). Learning to Read Chest X-Rays : Recurrent Neural Cascade Model for Automated Image Annotation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 2497–2506).

Si, Y., Wang, J., Xu, H., & Roberts, K. (2019). Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association*, (pp. 1–8).

Smit, A., Jain, S., Rajpurkar, P., Pareek, A., Ng, A., & Lungren, M. (2020). Combining Automatic Labelers and Expert Annotations for Accurate Radiology Report Labeling Using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1500–1519).

Son, R. Y., Taira, R. K., & Kangarloo, H. (2004). Inter-document coreference resolution of abnormal findings in radiology documents. *Studies in Health Technology and Informatics*,

107(Pt 2), 1388–1392.

Srivastava, R. K., Greff, K., & Schmidhuber, J. (2015). Highway NetWorks: Training Very Deep Networks. In *NIPS'15 Proceedings of the 28th International Conference on Neural Information Processing Systems* (pp. 2377–2385).

Stein, J. D., Rahman, M., Andrews, C., Ehrlich, J. R., Kamat, S., Shah, M., Boese, E. A., Woodward, M. A., Cowall, J., Trager, E. H., Narayanaswamy, P., & Hanauer, D. A. (2019). Evaluation of an Algorithm for Identifying Ocular Conditions in Electronic Health Record Data. *JAMA ophthalmology*, 137(5), 491–497.

Steinkamp, J. M., Chambers, C., Lalevic, D., Zafar, H. M., & Cook, T. S. (2019). Toward Complete Structured Information Extraction from Radiology Reports Using Machine Learning. *Journal of Digital Imaging*, 32(4), 554–564.

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012a). Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Demonstrations at the 13th Conference of the European Chapter of the ACL* (pp. 102–107).

Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. (2012b). Brat: A Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102–107).

Sugimoto, K., Takeda, T., Oh, J.-H., Wada, S., Konishi, S., Yamahata, A., Manabe, S., Tomiyama, N., Matsunaga, T., Nakanishi, K., & Matsumura, Y. (2021). Extracting clinical terms from radiology reports with deep learning. *Journal of Biomedical Informatics*, 116, 103729.

Sun, C., Yang, Z., Wang, L., Zhang, Y., Lin, H., & Wang, J. (2020). Biomedical named entity recognition using BERT in the machine reading comprehension framework. *arXiv:2009.01560 [cs]*.

Sung, S.-F., Lin, C.-Y., & Hu, Y.-H. (2020). EMR-based phenotyping of ischemic stroke using supervised machine learning and text mining techniques. *IEEE Journal of Biomedical and Health Informatics*, (pp. 1–1).

Syeda-Mahmood, T., D, P., Wong, K. C. L., D, P., Wu, J. T., D., M., H, M. P., Jadhav, A., D, P., Boyko, O., & D, M. D. P. (2020). Extracting and Learning Fine-Grained Labels from Chest Radiographs. *arXiv:2011.09517 [cs]*.

Tahmasebi, A. M., Zhu, H., Mankovich, G., Prinsen, P., Klassen, P., Pilato, S., van Ommering, R., Patel, P., Gunn, M. L., & Chang, P. (2019). Automatic Normalization of Anatomical Phrases in Radiology Reports Using Unsupervised Learning. *J Digit Imaging*, 32(1), 6–18.

Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., & Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, 84, 93–102.

Ulinski, M., Coyne, B., & Hirschberg, J. (2019). SpatialNet: A Declarative Resource for Spatial Relations. In *Proceedings of the Combined Workshop on Spatial Language Understanding (SpLU) and Grounded Communication for Robotics (RoboNLP)* (pp. 61–70).: Association for Computational Linguistics.

Vilain, M., Burger, J., Aberdeen, J., Connolly, D., & Hirschman, L. (1995). A model-theoretic coreference scoring scheme. In *MUC*.

Wang, S., Tseng, B., & Hernandez-Boussard, T. (2022). Deep Learning Approaches for Predicting Glaucoma Progression Using Electronic Health Records and Natural Language Processing. *Ophthalmology Science*, 0(0).

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., & Summers, R. M. (2017). ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3462–3471).

Wang, X., Peng, Y., Lu, L., Lu, Z., & Summers, R. M. (2018). TieNet: Text-Image Embedding Network for Common Thorax Disease Classification and Reporting in Chest X-Rays. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 9049–9058).

Wang, X. D., Weber, L., & Leser, U. (2020). Biomedical Event Extraction as Multi-turn Question Answering. In *Proceedings of the 11th International Workshop on Health Text Mining and*

Information Analysis (pp. 88–96).

Wang, Y., Sohn, S., Liu, S., Shen, F., Wang, L., Atkinson, E. J., Amin, S., & Liu, H. (2019a). A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making*, 19(1), 1.

Wang, Y., Sun, L., & Jin, Q. (2019b). Enhanced Diagnosis of Pneumothorax with an Improved Real-time Augmentation for Imbalanced Chest X-rays Data Based on DCNN. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 14(8), 1–1.

Wang, Z., Ng, P., Ma, X., Nallapati, R., & Xiang, B. (2019c). Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. *EMNLP/IJCNLP*.

Wheater, E., Mair, G., Sudlow, C., Alex, B., Grover, C., & Whiteley, W. (2019). A validated natural language processing algorithm for brain imaging phenotypes from radiology reports in UK electronic health records. *BMC Med Inform Dec Mak*, 19(1), 184.

Wood, D., Guilhem, E., Montvila, A., Varsavsky, T., Kiik, M., Siddiqui, J., Kafiabadi, S., Gadapa, N., Busaidi, A. A., Townend, M., Patel, K., Barker, G., Ourselin, S., Lynch, J., Cole, J., & Booth, T. (2020). Automated Labelling using an Attention model for Radiology reports of MRI scans (ALARM). In *Medical Imaging with Deep Learning*.

Woodward, M. A., Maganti, N., Niziol, L. M., Amin, S., Hou, A., & Singh, K. (2021). Development and Validation of a Natural Language Processing Algorithm to Extract

Descriptors of Microbial Keratitis From the Electronic Health Record. *Cornea*, 40(12), 1548–1553.

Wright-Bettner, K., Palmer, M., Savova, G., de Groen, P., & Miller, T. (2019). Cross-document coreference: An approach to capturing coreference without context. In *Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019)* (pp. 1–10).

Yan, K., Peng, Y., Sandfort, V., Bagheri, M., Lu, Z., & Summers, R. M. (2019). Holistic and Comprehensive Annotation of Clinically Significant Findings on Diverse CT Images: Learning from Radiology Reports and Label Ontology. *arXiv:1904.04661 [cs]*.

Yang, P., Fang, H., & Lin, J. (2018). Anserini: Reproducible Ranking Baselines Using Lucene. *J. Data and Information Quality*, 10(4), 16:1–16:20.

Yim, W.-W., Denman, T., Kwan, S. W., & Yetisgen, M. (2016). Tumor information extraction in radiology reports for hepatocellular carcinoma patients. In *AMIA Joint Summits on Translational Science Proceedings*, volume 2016 (pp. 455–64).

Yuan, Y. (2011). Extracting spatial relations from document for geographic information retrieval. In *Proceedings - 2011 19th International Conference on Geoinformatics*: IEEE.

Zech, J. R., Badgeley, M. A., Liu, M., Costa, A. B., Titano, J. J., & Oermann, E. K. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLoS Medicine*, 15(11), 1–17.

Zeng, X., Li, Y., Zhai, Y., & Zhang, Y. (2020). Counterfactual Generator: A Weakly-Supervised Method for Named Entity Recognition. In *EMNLP*.

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., & Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *Journal of the American Medical Informatics Association*, 28(9), 1892–1899.

Zhao, X., Ding, H., & Feng, Z. (2021). GLaRA: Graph-based Labeling Rule Augmentation for Weakly Supervised Named Entity Recognition. In *EACL*.

Zheng, C., Luo, Y., Mercado, C., Sy, L., Jacobsen, S. J., Ackerson, B., Lewin, B., & Tseng, H. F. (2019). Using natural language processing for identification of herpes zoster ophthalmicus cases to support population-based study. *Clinical & Experimental Ophthalmology*, 47(1), 7–14.

Zolnoori, M., Fung, K. W., Patrick, T. B., Fontelo, P., Kharrazi, H., Faiola, A., Wu, Y. S. S., Eldredge, C. E., Luo, J., Conway, M., Zhu, J., Park, S. K., Xu, K., Moayyed, H., & Goudarzvand, S. (2019). A systematic approach for developing a corpus of patient reported adverse drug events: A case study for SSRI and SNRI medications. *Journal of Biomedical Informatics*, 90, 103091.