Text Mining with Deep Learning for Secondary Use In Radiology

Wilson Lau

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:
Meliha Yetisgen, Chair
Martin Gunn
Thomas Payne

Program Authorized to Offer Degree:
Biomedical and Health Informatics

University of Washington

Abstract

Text Mining with Deep Learning for Secondary Use In Radiology

Wilson Lau

Chair of the Supervisory Committee:
Associate Professor Meliha Yetisgen
Biomedical Informatics and Medical Education

For more than a decade, electronic health records (EHR) have been used extensively in biomedical research. However, structured data, such as diagnoses and procedural codes do not necessarily capture the most precise medical conditions. Certain patient information, such as signs and symptoms, adherence to medication, social history and clinician recommendations, largely exist in unstructured clinical notes. Computational methods using natural language processing (NLP) techniques offer alternative ways to extract information from clinical notes by analyzing syntactic structure and semantics of words and phrases in unstructured text. In the domain of medical imaging, the radiology report is the main communication channel between radiologists and physicians. It contains a diverse and rich set of information about the imaging test, findings, diagnoses, and recommendations for further follow-up tests. While there has been some limited exploration of structured radiology reports that capture finding details, radiologists' findings today are predominantly documented in unstructured text. Since imaging tests are commonly used for cancer screening and diagnosis, extracting the findings associated with lesions and medical problems could facilitate many secondary use applications, including clinical decision-support systems, diagnostic surveillance of medical problems, and tracking follow-up recommendations. When clinical important findings are observed in the images, radiologists may recommend further imaging tests to the referring physicians. It is vital that these results, particularly if they are

unexpected, are not lost to follow-up. One study showed that approximately 16% of women with abnormal mammograms were diagnosed with breast cancers in 6 months. Extracting these follow-up recommendations and clinical findings (lesions and medication problems), provides supporting evidence for clinicians to determine their course of action.

In this dissertation, we focused on extracting information from radiology reports using state-of-the-art deep learning methods, through multiple research studies. One of the main goals is to deliver an open-source high performance extraction framework. In the first study, 685,912 radiologist recommendations and associated entities (reason, test, time frame) were extracted from 3 million radiology reports using recurrent neural network models. The extraction models achieved 0.93 F1 for recommendation sentences, 0.65 F1 for reason, 0.73 F1 for test, and 0.84 F1 for time frame.

In the second study, we explored using the latest pre-trained language model, BERT, to automatically classify radiology protocols. An in-domain BERT model pre-trained on the radiology corpus was shown to outperform out-of-domain BERT model and statistical ngram models based on Support Vector Machine (SVM), Gradient Boosting Machine (GBM), and Random Forest (RF). The intrinsic imbalanced nature of the dataset was tackled by using a knowledge distillation approach, which boosted the classification performance on the minority classes.

In the third study, the classification framework using the BERT model was further expanded to extract two clinical findings (Medical Problem and Lesion) from computed tomography (CT) radiology reports. Each finding was represented by an event comprising trigger and arguments. A corpus of 500 CT reports were annotated and a general-purpose deep learning framework was developed to extract the finding entities and relations from the reports. The entity extraction results showed that the in-domain BERT model pre-trained on the 3 million radiology reports (obtained from the first study) achieved an overall F1 score of 85.5%, while the recurrent neural model achieved 83.1%. The best end-to-end event extraction results achieved an overall F1 of 92.9% for triggers and 75.0%-84.8% for arguments. To assess model generalizability, we used an external validation set randomly sampled from the MIMIC Chest X-ray (MIMIC-CXR) database. The extraction performance on this validation set was 95.6% for finding triggers and 79.1%-89.7% for arguments, demonstrating that the model generalized well to the cross-institutional data with different imaging modality.

The general-purpose deep learning extraction framework processed annotated data directly from the BRAT rapid annotation tool and can be readily used to train entity and relation extraction models for other annotated corpora. Both the event extraction framework and the extracted MIMIC-CXR clinical findings will be shared with the research community.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table Number Page

LIST OF ABBREVIATIONS

BAN: Born-Again Neural Network

BERT: Bidirectional Encoder Representations from Transformers

CADE: Computer-aided detection

CADX: Computer-aided diagnosis

CRF: Conditional random field

CT: Computed tomography

EHR: Electronic health record

FDA: United States Food and Drug Administration

GBM: Gradient boosting machine

GRU: Gated Recurrent Unit

HAN: Hierarchical Attention Networks

HITECH: Health Information Technology for Economic and Clinical Health Act

IE: Information Extraction

LSTM: Long-Short term memory

MSE: Mean squared error

MIMIC: Medical Information Mart for Intensive Care

MIMIC-CXR: MIMIC Chest X-ray

MRI: Magnetic resonance imaging

NER: Named entity recognition

NLM: National Library of Medicine

NLG: Natural language generation

NLP: Natural language processing

RE: Relation extraction

RF: Random forest

RNN: Recurrent Neural Network

SMOTE: Synthetic Minority Oversample Technique

SVM: Support vector machine

TF-IDF: Term frequency-inverse document frequency

UMLS: Unified Medical Language System

ACKNOWLEDGMENTS

Finally at the end of the tunnel, I am fortunate to have many people lightening me up in my darkest times. Without them, I would not be able to complete this journey. I would like to thank my committee chair, Meliha Yetisgen, for her support, patience and guidance. Without her, my papers would not have been published. I would like to thank Martin Gunn and Thomas Payne for their clinical advice and support in my dissertation work. Many thanks to the radiology IT team who gracefully provided the infrastructure for me to run experiments. Those resources are the building blocks of my scientific research.

I also like to thank the colleagues in the BioNLP group for the helpful discussion, and all the friends and mentors in the Biomedical and Health Informatics department.

Most of all, I would like to thank my wife, Twiggy, who shared my burden and motivated me to continue without looking back, my two sons Logan and Luca who consistently kept me entertained and pulled me away from work for healthy work-life balance.

Although not directly related to my research, I like to acknowledge the pioneers in NLP research. Their vision and research germinated and enabled what we can do today.

I was truly in awe to learn that the first parser implemented by Dr. Naomi Sager and her team in assembly language took 20 minutes to parse a single sentence in 1959. Half a century later today a GTX 680 GPU can parse 250 sentences per second. That is 0.004 second per sentence. It would be interesting to see what the future holds for us in NLP research, and how much of it can be used to improve the quality of patient care.

DEDICATION

To my dear wife, Twiggy and my sons, Logan and Luca

Chapter 1

INTRODUCTION

In 2009, the Health Information Technology for Economic and Clinical Health Act (HITECH Act) was put in place to promote the nation wide adoption of electronic health records [1]. The Electronic Health Record (EHR) Mandate included financial incentive for healthcare providers to adopt EHR systems and convert medical paper charts to digital records. One of the key criteria for qualifying the incentives was to show "meaningful use" of the EHR system . While some requirements for "meaningful use" include concrete actionable descriptions, like "Use computerized order entry for medication orders", some overarching goals, such as "Improve the quality, safety, efficiency of health care, and reduce health disparities" or "Improve coordination of healthcare", entail analysis of large amount of patient data [2]. On the one hand, most patient data are still in unstructured format, without well-defined data points, making analysis challenging and difficult, on the other, converting sequence of utterances and discourse from free text into structured formats greatly reduces the expressiveness of communication. Often, only relevant and useful information would be extracted from clinical text for secondary use driven by the application [3]. Secondary use of EHRs applies to using personal health information (PHI) outside of direct health care delivery, including analysis, research, teaching, quality and safety improvement [4]. As a result, the demand for natural language processing (NLP) in the clinical domain has motivated a body of research in clinical Information Extraction (IE), forming a mainstream interest in bioNLP.

Clinical notes are common forms of documentation within a medical institution. They include physician's progress notes describing patient status over the course of patient care, nurse triage notes briefly recording patient disposition, or discharge summaries capturing patient present condition, significant findings, treatments, and any other information necessary to hand over care to the after care providers. Each type of note contains a different set of health information related to the patient at different stages of care. This dissertation specifically focused on extracting and classifying information in the radiology domain for secondary use application. Through multiple research studies, different neural architectures were explored with extensive experimentation and error analysis. The approach and techniques employed in these studies can be leveraged for applications in other clinical domains. One of the main objectives is to create a high performance extraction framework that can be used as a tool to extract information from unstructured text, even without any deep learning coding experience. The goal is to help radiologists and other physicians obtain more evidence to support their courses of action, a step closer to achieving more ”meaningful use” of EHRs.

1.1 Problem description

This thesis explored different neural architectures and machine learning methods to classify and extract radiological information, utilizing EHR data, most of which are unstructured narratives in radiology reports. Three different research studies were conducted, with specific research goals. However, each study expanded on the previous one and leveraged the data and modeling technique as a new baseline.

1.1.1 Extraction and Analysis of Clinically Important Follow-up Recommendations in a Large Radiology Dataset

Radiologists document the important observations that warrant further clinical follow-up when creating imaging reports. These important recommendations are made by radiologists to suggest that further investigation should be considered in order to avoid any potential adverse outcome to the patient. Unfortunately, radiologists may not necessarily phone the referring physicians and explain the findings and recommendations. The American College of Radiology Practice Parameter for the Communication of Diagnostic Imaging Findings [5] states that ”effective communication is a critical component of diagnostic imaging. Quality patient care can only be achieved when study results are conveyed in a timely fashion to those responsible for treatment decisions.” However, the document further emphasizes that ”there is a reciprocal duty of information exchange. The referring physician or other relevant health care provider also shares in the responsibility for obtaining results of imaging studies ordered and acting on them in an appropriate manner.” While an official interpretation (final radiology report) must be provided by the radiologist to the referring physician, verbal communication by phone is not mandated. In 2005, the Joint Commission set a National Patient Safety Goal (NPSG.02.03.01) to ensure the timely reporting of critical test and diagnostic procedure results to licensed caregivers [6]. However, non life-threatening test results and follow-up recommendations can still be communicated in written reports.

During clinical visits, patients can be transferred from one department to another during the course of care, such as from emergency department to outpatient services. The high patient turnover and short stay can cause follow-up information being missed [7]. Even if the test results are forwarded to the referring providers, their demanding workload and time spent with other patients can delay reviewing the results. One survey of 262 providers showed that physicians spent on average 74 minutes per clinical day managing test results. 83% of them reported at least one delay in reviewing test results over a period of two months [8]. Delay in communication and loss to follow-up can result in adverse outcomes, particularly if the findings are incidental and unexpected. One study showed that approximately 16% of women with abnormal mammograms were diagnosed with breast cancers at the 6-month follow-up [9].

Failure to follow up test results not only can compromise patient care and cause negative impact on patient health but also entail medical malpractice and financial consequences. In 1997, a claims survey of malpractice data collected from insurance companies showed that ”failure to communicate results of radiologic examinations” was the second most common cause of medical malpractice lawsuit in the United States [10]. In 2013, another study involving 8401 radiologists in 47 states revealed similar results [11]. Error in diagnosis remained the top most common malpractice claim while inadequate communication remained the second. 31% of the radiologists had at least one claim in their career. It is concerning to see the similar severity of inadequate communication in the past twenty years and improvement was greatly desired. Identifying follow-up recommendations in radiology reports systematically could potentially augment existing channels of clinical information for preventing delays in diagnosis.

Yetisgen-Yildiz et al. created a corpus of 800 de-identified radiology reports collected from Harborview Medical Center [12, 13]. To identify recommendation sentences, they developed a Maximum Entropy classifier trained with a very rich set of linguistic features including ngrams, UMLS concepts, syntactic, temporal and structural features, achieving an extraction performance of 87% F1. This dissertation work extended their dataset of 800 reports with a much larger set of 3,301,748 radiology reports collected from two different institutions, including the University of Washington Medical Center (1,903,772 reports) and Harborview Medical Center (1,397,976 reports) from year 2008 to 2018. In addition to follow up sentences extraction, three associated entities (reason, time frame, test) were also extracted. This dissertation study explored using recurrent neural networks for both recommendations and entities extraction. Specifically, the recommendations were extracted by Hierarchical Attention Networks (HAN) [14] and entities were extracted by bi-directional Long-Short term memory (LSTM) with conditional random field (CRF). Based on the extraction results, a follow-up analysis was conducted to investigate the adherence of follow-up encounters by imaging modalities.

1.1.2 Automatic Assignment of Radiology Examination Protocols

Radiologists are constantly juggling between tasks. A Medscape survey in 2015 reported that 49% of radiologists had burnout symptoms. Their burnout rate was ranked the 7th highest among all physicians [15]. One of the risk factors was constant interruption in their complex working environments. Besides image interpretation, other responsibilities include phone calls with referring physicians (as described in previous section), answering pages, consultation, teaching and protocol assignment. The interruptions not only have negative impact on radiology report turn-around times [16] but can also lead to diagnostic discrepancy [17]. One study investigated the time spent on different tasks by a primary neuroradiology fellow in a 48 hours time period. About 38% of the time was spent on non-image interpretative tasks, in which 6% was on study protocoling [18].

In medical imaging, an examination protocol is referred to as a sequence of steps that determine how an imaging test is administered. This could involve selecting the optimal spatial orientation, resolution, parameter setting for image acquisition and reconstruction. In a nonemergent setting, after receiving the suggested examination from the ordering provider, the radiologist will look at the patient's clinical information such as diagnosis and clinical history before selecting the appropriate protocol for advanced imaging studies (e.g. computed tomography, magnetic resonance imaging, nuclear medicine examinations) in order to answer the clinical question that the study has been ordered to answer. The manual protocoling process can be time-consuming, repetitive, and may delay performing timely imaging, and result in unnecessary variability in the techniques used for image acquisition [19].

Generally, protocols are differentiated by the anatomic region of interest, the administration of intravenous, and/or oral contrast, or no contrast. For cancer screening and diagnosis, computed tomography (CT) scans are often used to study different parts of the body. Chest, Abdomen and Pelvis (CAP) are the common body regions since they cover major internal organs, such as liver, pancreas, bowel, kidneys, bladder, lungs, and heart. Although protocoling can be time-consuming, some common imaging examinations are fairly simple and repetitive, making this task a good candidate for automation.

In this study, a machine learning approach using pre-trained language model was investigated to automate protocol assignment of 35,085 radiology body CT examinations. As noted, CT examinations involving certain body regions, i.e. Chest, Abdomen and Pelvis (CAP), are very common, and therefore, make up the majority of cases. To tackle the imbalanced nature of the dataset, a novel approach using knowledge distillation was used to augment the minority instances.

1.1.3 Extraction of clinical findings from radiology reports

Radiology reports are the official interpretation of the imaging tests from radiologists. They are also the principal means of communication and documentation. In fact, radiology reports contain a diverse and rich set of information, including findings, overall impression and recommendations for further diagnostic tests. As discussed in the previous study, some CT exams are very common, specifically for certain body locations. According to RNSA and ACR [20], CT imaging is one of the fastest and most accurate tools for examination because it provides detailed, cross-sectional views of all types of tissue. It is also the best way to detect cancers in the chest, abdomen and pelvis. Not only does it confirm the presence of a tumor, but also identify the precise location in the organs such as lung, liver, kidney, pancreas, as well as the measurement of the tumor size and metastasis in nearby tissues. CT is also commonly used to evaluate blood clots (pulmonary embolism) and other medical problems. It is often used in the Emergency Department to quickly assess injuries. Extracting clinical findings from CT radiology reports provides great opportunities to improve clinical care and decision support. However, the heterogeneous writing style, use of abbreviation, presence of hedging statements in radiology reports poses some challenges to the extraction task. To fully capture the clinical finding details, we introduced a new event-based annotation schema focused on two clinical findings: Lesion and Medical Problem. Each finding event consisted of relevant arguments to capture the detail of the finding. A new corpus of 500 CT reports was annotated using this new schema.

In the first study, we extracted the recommendation associated entities from the multi-institutional dataset using a recurrent neural network model. In the second study, we explored using the state-of-the-art pre-trained language model, BERT [21] in CT exam protocol classification. This third study leveraged what we learned from the previous two studies. Specifically, we used the model from the first study as a baseline and incorporated the multi-institutional dataset. We further explored the BERT model employed in the second study, and presented a new BERT model pre-trained on the multi-institutional dataset. A new deep learning framework was developed to fine-tune the same BERT model to extract both the entities and relations in clinical finding events. To assess the generalizability of the extraction model, we extracted all the clinical findings from the MIMIC Chest X-ray radiology reports [22] and evaluated the extraction performance.

1.2 Contributions and objectives

The objectives of this dissertation are extracting and classifying information in the radiology domain for secondary use, employing start-of-the-art neural NLP approaches. Three research studies were conducted. Each study expanded on the previous one and leveraged the knowledge and new techniques in machine learning and IE. The studies challenged previous baseline methods, employed novel neural approaches to achieve higher performance. This work makes the following contributions: (1) extracting recommendations and related entities from over 3 million radiology reports in UW medical institutions, (2) a detailed event-based annotated corpus for extracting clinical finding in radiology reports, (3) a high performance deep learning extraction framework that can be trained to predict entities and relations from unstructured text, (4) a new approach to automatically classify protocols for CT examinations and handle data imbalance using knowledge distillation.

This dissertation adopts recent advancements in artificial intelligence to unlock new opportunities to effectively extract information from radiology reports for secondary use. The extracted clinical findings and recommendations can complement existing structured elements in EHR to enable better secondary research.

1.3 Outline for readers

The structure of this dissertation describes the different projects in distinct chapters. Chapter 2 starts with a general background on NLP in the clinical domain and reviews the IE research in the radiology domain. Chapter 3, 4, and 5 describe the three research studies in detail with their own specific literature reviews. A summary of each chapter is described as follows:

Chapter 3 : this chapter introduces the multi-institutional radiology corpus, and presents two neural approaches to extract follow-up recommendations and associated entities.

Chapter 4 : this chapter explores using the pre-trained language model, BERT, to classify CT protocols, and investigates different machine learning approaches to handle data imbalance.

Chapter 5 : this chapter introduces a new annotation schema for 2 specific clinical finding (Lesion, Medical Problem) and a new corpus of 500 CT reports annotated with the schema. This work capitalizes what we have learned from the previous two studies and introduces a new BERT model to extract clinical findings from both CT reports and chest X-ray reports.

Chapter 6 : finally, we conclude this dissertation work by providing insights into future research opportunities using the extracted information.

Chapter 2
BACKGROUND

This chapter presents a brief overview of how natural language processing has been applied in the clinical domain. In addition, we review the past research in radiology information extraction and how the research shifted from rule-based, statistical approaches to the more recent artificial intelligence based neural models.

2.1 Natural Language Processing in the Clinical Domain

NLP research in the clinical domain dates back to the 1960 [23]. Recognising a sentence as a string of language structure, a context-free grammar can be defined recursively by grammatical formulas of substrings. The Linguistic String Project (LSP), initiated by Dr. Naomi Sager, introduced a parsing program to process a sentence from left-to-right [24, 25]. The parser was later adapted to include medical lexicons and dictionaries [26]. The Medical Language Processor (MLP) transformed unstructured clinical documents into XML representation of medical concepts by extracting symptoms, drug dosage and possible side effects of prescriptions [27]. The MLP system laid a foundation work in syntactic parsing of clinical text. To improve computer "understanding" of clinical semantics, in 1986, the National Library of Medicine (NLM) began the development of the Unified Medical Language System (UMLS) [28], an effort to disambiguate medical concepts from diverse machine-readable sources, and to distribute useful information to research communities. The UMLS became the semantic backbone of multiple notable clinical NLP systems, including MetaMap, a freely available processing pipeline that automatically identifies UMLS concepts from unstructured narratives [29], clinical Text Analysis and Knowledge Extraction System (cTAKES) from Mayo Clinic [30], and the open source Health Information Text Extraction system (HITEx) developed by the Brigham and Women's Hospital and Harvard Medical School [31], which combines the language analysis capabilities from the General Architecture for Text Engineering framework (GATE) [32] and the domain knowledge from UMLS.

2.1.1 NLP challenges and Open datasets

To motivate public interest in advancing clinical NLP research, different academic organizations and conferences promoted community challenges with openly accessible datasets. Informatics for Integrating the Biology and the Bedside (i2b2) has been organizing NLP challenges on different types of clinical information extraction since 2006. These challenges included private health information de-identification [33], medical concept extraction [34], temporal information extraction [35], as well as medication information extraction [36]. Besides i2b2, other conferences and workshops also hosted clinical NLP challenges. The SemEval (Semantic Evaluation) has been hosting challenges since 1998. One example is the 2017 temporal evaluation which aimed to predict future medication conditions based on the existing ones [37]. The Text Analysis Conference (TAC) also hosted and provided annotated corpus for adverse drug reaction extraction [38]. Funded by the National Institutes of Health (NIH), the collaboration between the Clinical E-Science Framework (CLEF) and the Shared Annotated Resources (ShARe) hosted the 2013 challenge targeting extraction of disorders and acronyms/abbreviations [39]. In the following year, they released the ShARe corpus for identifying and mapping of diseases and disorders in clinical reports to UMLS concepts [40]. The same corpus was also used in the 2014 challenge for identification of disorder related attributes [41]. The Association for Computational Linguistics (ACL) has organized BioNLP workshops since 2008. Although not hosted in the format of community challenges, the conference had a specific focus each year and often included sessions for clinical language processing [42].

One of the largest publicly accessible clinical datasets is the MIMIC-III dataset, which contains 7 years of de-identified patient data from intensive care [43]. The dataset consists of medications, laboratory results, clinical notes, demographics, and billing information. It facilitated a large body of biomedical research, from a longitudinal study in disease detection [44] to pre-training a deep learning language model [45]. Another dataset released by the same authors and mostly relevant to this dissertation is the MIMIC Chest X-ray (MIMIC-CXR) dataset [22], which consists of 227,835 imaging studies for 65,379 patients hospitalized in intensive care unites, with accompanying de-identified radiology reports. NLP researchers have used the radiology reports in this dataset to predict chest related diseases, such pulmonary edema [46].

2.2 IE in the Radiology Domain

Radiology reports contain findings and recommendations documented by radiologists. Extracting this information into structured representation can harness their potential to improve clinical care and facilitate secondary use, such as generating alerts for follow up examination [47], or identifying patients with pulmonary nodules [48] and pulmonary embolism [49].

Generally the extraction involved identifying clinical entities using named entity recognition (NER), and additionally identifying the relations among the entities using relation extraction (RE). NER is considered a sequence labelling task in IE. The goal is to correctly locate and identify mentions of pre-defined concept labels in unstructured text. It can be achieved by using some common tagging format, such as BIO (beginning, inside, outside). For a simple example, the text sequence *John Smith* can be labelled as B-Person, I-Person which signifies the beginning token and inside token of a Person entity. Early research efforts on radiology IE employed rule-based approaches. The notable MedLEE system developed by the Columbia University incorporated comprehensive syntactic and semantic grammars to extract information from chest X-ray reports [50, 51]. The conceptual model comprised 350 semantic grammar rules, 1,720 single-word lexicons, and 1,400 multi-word phrases. It took half a person-year to develop the semantic grammars [52, 53]. Sevenster et al. used MedLEE to identify and correlate the finding observation and body location entities. However, the major drawback was that the extraction recall was less than 46% due to the lack of comprehensive lexicons and grammatical rules [54]. Domain adaptability is a major problem for rule-based and lexicon-based approaches as these methods require expert intervention to upkeep the logic of the rules and the dictionaries, which are often tailored to a specific problem or domain.

To overcome the limitations of rule-based systems, more contemporary radiology extraction work has used statistical machine learning approaches to extract finding information. Statistical machine learning incorporates numeric features derived from input observations and makes probabilistic decisions based on the feature weights. Hassanpour et al. compared three different NER approaches in the extraction of anatomy, observation, and uncertainty from Chest CT reports [55]. For the rule-based approach, they leveraged cTAKES NER module with a custom dictionary extracted from RadLex terms [56]. For the other two statistical methods (Conditional Markov Model and Conditional Random Field), they used linguistic features such as part of speech, word stems, n-grams, orthographical shape of words, negation as well as RadLex semantic classes. Both models achieved a very similar F1 score of 85% whereas the rule-based method achieved 58%. Further analysis showed that the RadLex terms had higher feature weights in the model and boosted the F1 score another 15% higher when they were included in the feature set. Yim et al. employed maximum entropy model to extract relations between tumor references and attributes from radiology reports of hepatocellular carcinoma patients [57]. The feature set consisted of n-grams, part of speech tags, dependency tree, UMLS concepts and custom linguistic rules. They achieved 87% in entity extraction and 74% in relation extraction. However, one challenge with the statistical machine learning approach is that engineering the optimal set of features require substantial data preprocessing.

Recent IE research in radiology employed neural network modeling to learn the optimal features from high-dimensional data points. To capture the long distance dependencies in text sequence, one popular architecture is Bi-directional Long Short-Term Memory (BiLSTM). Cornegruta et al. extracted 4 different entities (body location, clinical finding, descriptor and medical device) from an annotated corpus of 2,000 radiology reports using

BiLSTM [58]. Despite the promising results using BiLSTM, such models often demand a large collection of training data to learn the context of words. The same medical concept can often be described by words in different inflected or synonymous forms. For instance, the words "renal", "nephric" and "kidney" refer to similar anatomical location despite their different morphological structures. Furthermore, distributed word embeddings, such as the Global Vectors for Word Representation (GloVe) [59], are not capable of representing words that are absent in the training corpus, or represent the same word differently based on the context. For example, the two instances of the word "back" would be given the same embedding despite their polysemous meanings in these two sentences: "the patient will be *back* for contrast study", "Clinical history: Low *back* pain with history of compression". If words could be not represented based on their context, the limited knowledge encoded by these embeddings can result in sub-optimal performance in the NLP tasks.

State-of-the-art neural language models, such as Bidirectional Encoder Representations from Transformers (BERT) [21], and Generative Pre-trained Transformer 3 (GPT-3) [60], utilized layers of multi-head self-attention architecture and pre-training to develop deep representation of words. Provided that the model is sufficiently pre-trained on unlabeled data in the target domain, the expressive contextual representations can be transferred to specific prediction tasks, including IE. This approach is particularly advantageous when the target data is scarce. Sugimoto et al. extracted 7 different clinical entities from a corpus of 540 Japanese CT radiology reports using a pre-trained Japanese BERT model [61]. Zhang et al. fine-tuned a BERT model to extract both breast cancer entities and relations from a corpus of 600 Chinese clinical notes (100 radiology reports) [62]. Both studies demonstrated that the BERT model outperformed the BiLSTM model. This dissertation explored using the pre-trained language model, BERT, to classify CT examination protocols and extract radiology clinical findings.

Chapter 3

EXTRACTION AND ANALYSIS OF CLINICALLY IMPORTANT FOLLOW-UP RECOMMENDATIONS IN A LARGE RADIOLOGY DATASET

This chapter describes a study using recurrent neural models to extract follow-up recommendations and associated entities from a large collection of UW radiology reports. Using the extracted information, an analysis was conducted to understand whether follow-up recommendations actually occurred.

3.1 Introduction

Depending on circumstances after the imaging tests, radiologists may recommend further investigation to clarify the diagnosis. If a finding is not expected, such as tumors, diagnostic surveillance could be recommended to monitor the progression and clarify significance of the finding. These recommendations are made to inform ordering providers about the clinical significance of the findings and to ensure further investigation is considered to avoid possible adverse outcomes. Despite the importance of follow-up recommendations in a radiology report, follow-up encounters do not always happen. One reason is that the recommendations are not explicitly highlighted in a report and therefore can be overlooked. Moreover, the radiologists' busy work schedule may prevent them from communicating to the physicians verbally on the phone [8]. Patients can also be transferred between facilities, which could cause miscommunication and delay of clinical intervention [7]. Lost to follow-up not only can result in adverse outcomes [63], but potentially cause legal and financial consequences [10]. In prior work, Yetisgen-Yildiz et al. created a corpus of 800 de-identified radiology reports collected from Harborview Medical Center [12, 13]. Their Maximum Entropy classifier achieved 87% F1 score in identifying follow-up recommendations. This dissertation study extended their dataset and included a larger set of 3.3 millions radiology reports with different imaging modalities. In addition, we extracted the entities associated with the recommendation using a BiLSTM-CRF model. We conducted a retrospective analysis on the extracted recommendations and entities to understand the adherence of follow-up recommendations across the imaging modalities.

In this study, we define a follow-up recommendation as a statement made by the radiologist in a given radiology report. The recommendation is to advise the ordering provider to further evaluate an imaging finding by other imaging tests. Figure 3.1 presents a radiology report with such a follow-up recommendation. In our annotation, we first labeled sentences containing a recommendation. For each identified recommendation, we also annotated the spans that describe (1) the reason for follow-up, (2) recommended test, and (3) time frame. In Figure 3.1, the recommendation sentence is “Given family history, would recommend repeat ultrasound in 4-5 weeks to evaluate fetal growth and complete anatomic survey”, reason is “to evaluate fetal growth and complete anatomic survey”, recommended test is “ultrasound”, and time frame is “4-5 weeks”.

IMPRESSION

Singleton pregnancy.Size consistent with dates. Anatomic survey limited by maternal body habitus and fetal position. Inadequate views of fetal heart and spine. **Given family history, would recommend repeat ultrasound in 4-5 weeks to evaluate fetal growth and complete anatomic survey. If unable to visualize fetal heart at that time, consider fetal echo.**

Figure 3.1: Example radiology report with recommendation information annotations.

3.2 Related Work

Prior follow-up recommendation detection research were primarily based on rule-based approaches. Dutta et al. [64], Dang et al.[65], Mabotuwana et al. [66] used lexicons pattern matching to identify the recommendations. Chapman et al. [67] and Johnson et al. [68] added negation detection using the heuristic algorithm, ConText. More recent work was based on statistical models. Carrodeguas et al. created a corpus of 1000 randomly selected ultrasound, radiography, CT, and MRI reports. The extraction performance of three statistical models were 75% F1 (random forest), 83% F1 (logistic regression), and 85% F1 (support vector machine) respectively [69]. Yetisgen-Yildiz et al. developed a maximum entropy classifier and achieved a F1 score of 87% based on a very rich set of features including ngrams, UMLS concepts, syntactic, temporal as well as structural features [13]. Their work in particular is most relevant to our study and provides a baseline for our work.

3.3 Methods

This section describes the datasets in this study and our neural modeling approach. We conducted the same data imbalance experiments employed by Yetisgen-Yildiz et al [13] to explore the optimal ratio of negative sentences over positive sentences.

3.3.1 Data

This work used two different corpora. The first one was the pilot corpus from Yetisgen-Yildiz et al [13], which consisted of 800 de-identified radiology reports extracted from the radiology information system of our institution. The reports represented four different imaging modalities, including radiography computer tomography (CT), ultrasound, and magnetic resonance imaging (MRI). The distribution of the reports is listed in Table 3.1.

Imaging modality	**Number of reports**
Computer tomography	486
Radiograph	259
Magnetic resonance imaging	45
Ultrasound	10
Total	800

Table 3.1: Distribution of reports in pilot corpus.

The annotation was performed by one radiologist and one internal medicine specialist. They independently went through each of the 800 reports and highlighted the boundary of the sentences that contained follow-up recommendations. Out of 18,748 sentences in 800 reports, the radiologist annotated 118 sentences and the internal medicine specialist annotated 114 sentences as recommendation. They agreed on 113 of the sentences annotated as recommendation. The inter-rater agreement measured in terms of F-score was 97.4%.

The second corpus was a much larger set of 3,301,748 radiology reports from two different institutions including the University of Washington Medical Center (1,903,772 reports) and Harborview Medical Center (1,397,976 reports) from year 2008 to 2018. Table 3.2 shows the distribution of radiology reports by modality in this larger dataset.

Imaging modality	**Number of reports**
Angiography	53,658
Computed Tomography	706,908
Fluoroscopy	1,072
Magnetic Resonance Imaging	243,833
Mammogram	157,374
Nuclear Medicine	58,350
Portable Radiography	310,311
Positron emission tomography	1,799
Ultrasound	351,761
X-Ray	1,416,682
Total	3,301,748

Table 3.2: Distribution of reports in multi-institutional radiology corpus

Two levels of annotations were performed on this dataset. First, one radiologist and one neurologist highlighted the boundary of the sentences that contained follow-up recommendations. Then one neurologist and one medical school student tagged three different entities in the highlighted recommendation sentences: (1) Test: the imaging test or clinical exam that is recommended for follow-up, e.g., screening breast MRI or CT, (2) Time frame: the recommended time frame for the recommended follow-up test or exam, e.g., 1-3 weeks, and (3) Reason: the reason for the critical follow-up recommendation, e.g., to assess the actual risk of Down's Syndrome. Since only 15% of radiology reports in the corpus contained recommendations, to actively select reports with recommendations for annotations, a high recall and low precision classifier was used [12]. By having the classifier predict reports with potential recommendations, the annotators only needed to correct the false positives and thereby expedited the annotation process. The annotator agreement for the recommendation sentences was 0.59 F1 score, and was subsequently improved after multiple meetings of disagreement resolution and revision of annotation guidelines. At the entity level, the agreement was 0.78 F1 for reason, 0.88 F1 for test, and 0.84 F1 for time frame.

Note that the annotation in both corpora was performed by Yetisgen-Yildiz et al. in the prior work [12, 13]. It is however important to present the annotation details in order to better understand the gold standard used in this study. Our final annotated corpus contained 597 positive instances of recommendation sentences and 11787 sentences without recommendation from 567 radiology reports, taken from both datasets. At the entity level, there were 735 test, 173 time frame and 545 reason entities in the final corpus.

3.3.2 Approach

Recommendations extraction

To extract the recommendation sentences from the dataset , the reports were first chunked into sentences using the NLTK[1] sentence tokenizer. As shown in Table 3.3, some imaging modalities generally have more sentences per report than others.

Imaging Modality	**Number of sentences**	**Average number of sentences per report**
Angiography	1,504,939	28.05
Computed Tomography	18,109,590	25.62
Fluoroscopy	13,452	12.55
Magnetic Resonance Imaging	5,688,512	23.33
Mammogram	2,016,911	12.82
Nuclear Medicine	1,144,518	19.62
Portable Radiography	2,055,534	6.62
Positron emission tomography	41,423	23.03
Ultrasound	6,841,966	19.45
X-Ray	10,008,031	7.06

Table 3.3: Distribution of sentences by image modality in the multi-institutional radiology corpus

We defined our recommendation extraction task as a binary classification problem at the sentence level. We implemented our sentence classifier based on Hierarchical Attention Networks (HAN) [14]. HAN is a neural model that employs a stacked recurrent neural network architecture. In particular, the weights of the hidden layers for each word are aggregated

[1] `https://www.nltk.org`

by an attention mechanism to form a sentence vector. The importance of each word in association with the outcome label (binary value indicating the presence of recommendation) is represented by the attention weight vector that can be learned by a layer of bidirectional Gated Recurrent Unit (GRU). The attention weight vector is computed through a softmax function of the input context vector and a single hidden layer. Intuitively, the attention vector represents how important the word is in determining the outcome label. The sentence vector which is made up of these word attentions are then passed to another similar attention mechanism where the importance of sentences can also be learned by another layer of bidirectional GRU (sentence encoder). The bidirectional nature of the encoders allows the contextual information in the input to be read in both directions and summarized. The hierarchical architecture allows the model to learn the context of a document by summarizing the context of its sentences, each of which in turn was summarized by its own words. The ability to selectively learn from local segments of text to predict the outcome labels is a unique characteristic of attention mechanism in deep learning. This network model has been proven to be more effective [70] than conventional statistical machine learning approaches in extracting information from pathology reports. Since radiology recommendations follow similar hierarchical structure which consist of multiple sentences made up of multiple words, the HAN model is suitable for our recommendation classification task. Figure 3.2 shows how a sentence is being classified by the HAN model. During the model inference, each sentence would be predicted individually to determine the presence of recommendation. Consecutive sentences both of which contained positive prediction would constitute a single recommendation.

Hyperparameter optimization: We pretrained our word embeddings using Word2Vec [2] on the entire radiology dataset. Based on our preliminary experiments, taking into account the limitation of hardware resources, we identified the range for each hyperparameter in the search space: Word2Vec embedding dimension (100-300); number of bidirectional GRU unit

[2] `code.google.com/p/word2vec`

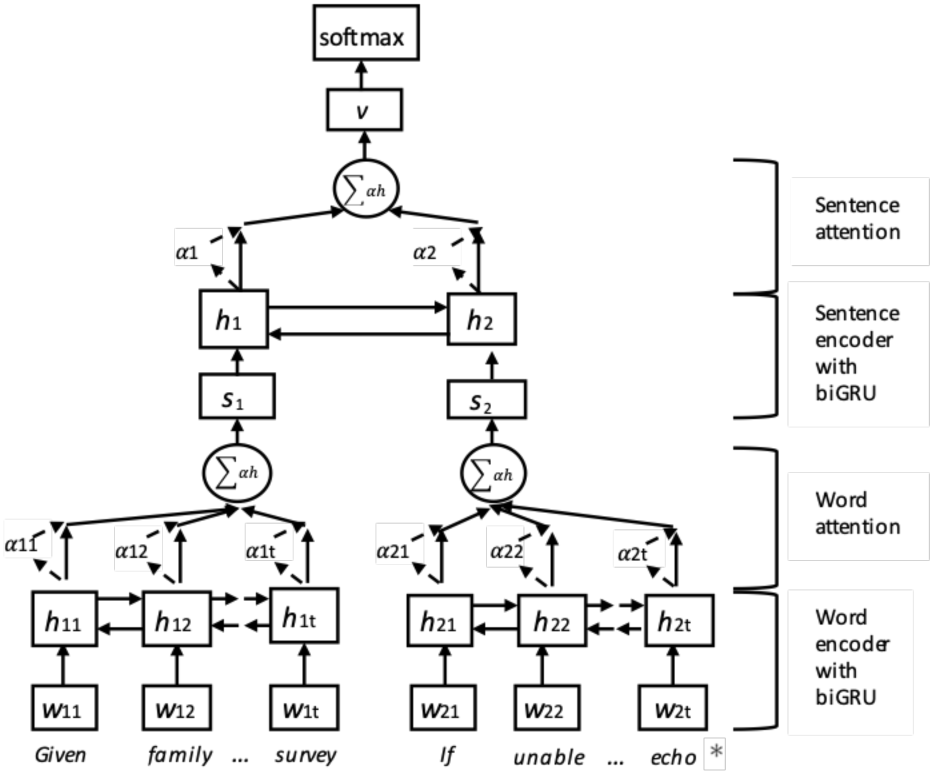


Figure 3.2: Architecture of the HAN model. *See Figure 3.1 for complete sentence

on word encoder (100 - 500); number of bidirectional GRU unit on sentence encoder (100 - 500); drop out (0.3 - 0.5). We have also experimented with both Adam optimizer and stochastic gradient descent (SGD). Table 3.4 shows our best hyperparameter configuration.

Parameter	**Value**
word2vec embedding dimension	300
number of bidirectional GRU unit on word encoder	300
number of bidirectional GRU unit on sentence encoder	300
drop out	0.4
optimizer	Adam

Table 3.4: HAN hyperparameter configuration

We used 0.8/0.2(train/validation) split and applied early stopping with the validation set to avoid overfitting. The patience level was set to 10 epochs. On each epoch, we evaluated the model based on the predicted F1 score on the validation set. The training would stop when no improvement was shown in the last 10 epochs.

Entities extraction

We used NeuroNER [71] for the implementation of the BiLSTM-CRF model to process the annotated files in BRAT standoff format [3]. The core of NeuroNER consists of two stacked layers of recurrent neural networks. The first layer is the Character-enhanced token-embedding layer in which the embedding of each word token is learned by a BiLSTM from its character embedding. The resulting token embedding is then concatenated with our pretrained Word2Vec word embeddings to form an enhanced token embedding. These token embeddings are then processed by another layer of BiLSTM, the Label prediction layer, to learn the context of the sequence. Finally, the output states are sent to a feed-forward layer, the Label sequence optimization layer, to determine the predicted entity for each token

[3] https://brat.nlplab.org/standoff.html

with the highest confidence. The character embedding captures the morphological features of word tokens and allows the model to learn morphemes, acronyms and out-of-vocabulary tokens. It provides another level of word presentation that is not captured by sampling word co-occurrence as in Word2Vec and GloVe. We used BIOES annotation (Begin, Inside, Outside, End, Single) to tag each token in the sequence and performed 5-fold cross validation on the training corpus. We used the same Word2Vec embeddings trained with the multi-institutional radiology corpus of 3.3 million radiology reports as in the recommendations extraction.

3.4 Results

Recommendations extraction

We merged annotations from the pilot corpus and the multi-institutional radiology corpus to create one gold standard corpus that contains 693 positive sentences and 30429 negative sentences from a total of 1367 radiology reports. Following the same imbalance experiements by Yetisgen-Yildiz et al. [12], a series of experiments were designed to determine the optimal ratio of positive and negative instances. Let P the set of positive training sentences and N be the set of negative sentences. For each k (k=1,...,n), we trained a classifier where the cardinality of N was equal to k times the cardinality of P. We performed 5-fold cross-validation at each value of K to obtain the average performance scores. We achieved the best 5-fold cross validation results at K=32 with 0.94 precision, 0.92 recall, and 0.93 F1 score (true positive: 635, true negative: 11755, false positive: 39, false negative: 58), as shown in Table 3.5 and Figure 3.3. The performance was better than Yetisgen-Yildiz et al. [12] (0.66 precision, 0.88 recall, 0.76 F1 score using Max-Ent classifier with extensive feature engineering).

K	**TP**	**TN**	**FP**	**FN**	**Precision**	**Recall**	**F1**	**Accuracy**
1	672	11602	192	21	0.778	0.970	0.863	0.983
3	676	11618	176	17	0.793	0.975	0.875	0.985
5	664	11682	113	29	0.855	0.958	0.903	0.989
7	665	11680	114	28	0.854	0.960	0.904	0.989
9	661	11691	104	32	0.864	0.954	0.907	0.989
11	660	11710	84	33	0.887	0.952	0.919	0.991
13	650	11718	76	43	0.895	0.938	0.916	0.99
15	649	11724	70	44	0.903	0.937	0.919	0.991
17	648	11737	57	45	0.919	0.935	0.927	0.992
19	650	11730	64	43	0.910	0.938	0.924	0.991
21	646	11739	55	47	0.922	0.932	0.927	0.992
23	638	11746	48	55	0.930	0.921	0.925	0.992
25	631	11755	40	62	0.940	0.911	0.925	0.992
27	626	11757	38	67	0.943	0.903	0.923	0.992
29	633	11757	37	60	0.945	0.913	0.928	0.992
31	627	11761	34	66	0.949	0.905	0.926	0.992
32	635	11755	39	58	**0.942**	**0.916**	**0.929**	0.992
33	632	11757	37	61	0.945	0.912	0.928	0.992
34	614	11762	32	79	0.950	0.886	0.917	0.991
35	638	11745	49	55	0.929	0.921	0.925	0.992
36	621	11759	36	72	0.945	0.896	0.920	0.991
37	623	11753	41	70	0.938	0.899	0.918	0.991
38	619	11763	31	74	0.952	0.893	0.922	0.992
39	623	11763	31	70	0.953	0.899	0.925	0.992
40	624	11762	32	69	0.951	0.900	0.925	0.992
41	603	11765	29	90	0.954	0.870	0.910	0.990
42	627	11759	35	66	0.947	0.905	0.925	0.992
43	622	11759	35	71	0.947	0.898	0.921	0.992

Table 3.5: Performance evaluation. K: class ratio, TP, true positive; TN, true negative; FP, false positive; FN, false negative; The highest precision, recall, and F1 score values are in bold (k = 32).

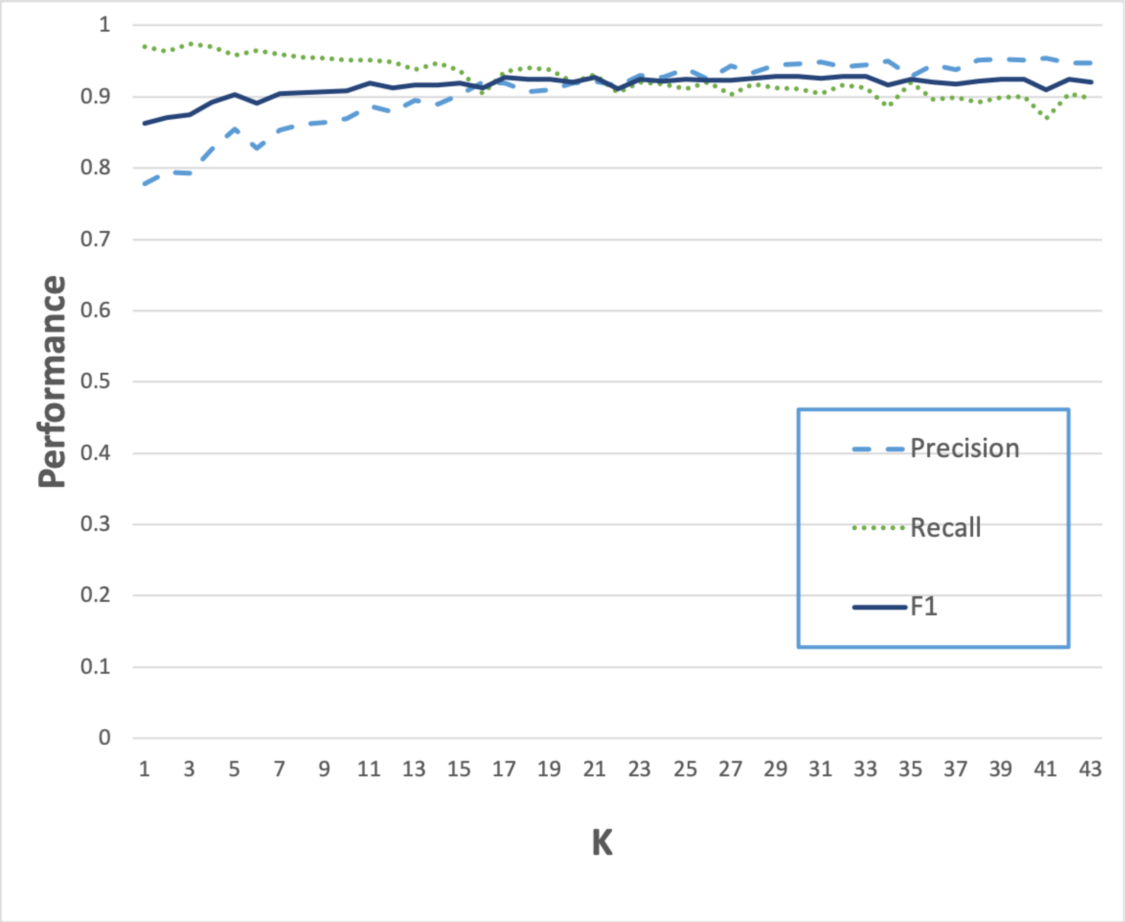


Figure 3.3: Precision, recall, F1 score curves. k: class ratio.

Entities extraction

Table 3.6 shows the token-based 5-fold cross validation results on the three entities.

Entity	**Precision**	**Recall**	**F1**
Reason	68.53	62.05	65.10
Test	74.20	71.48	72.71
Time frame	83.38	85.05	84.16

Table 3.6: Token level entity extraction 5-fold cross-validation results in %

3.4.1 Extraction of multi-institutional radiology corpus

We used the trained recommendations extraction model to identify recommendations from the multi-institutional radiology corpus. The corpus consisted of 47,424,876 sentences. A total of 685,912 recommendations were extracted. The distribution by modality is shown in Table 3.7. An example of recommendations in each modality was presented in Table 3.8.

Imaging Modality	**# of recommendations**	**# of reports with recommendations (%)**
Angiography	8455	7234 (13.48%)
Computed Tomography	193414	140066 (19.81%)
Fluoroscopy	103	100 (9.33%)
Magnetic Resonance Imaging	60954	34928 (14.32%)
Mammogram	210828	154255 (98.02%)
Nuclear Medicine	10141	7426 (12.73%)
Portable Radiography	13519	12951 (4.17%)
Positron emission tomography	472	336 (18.68%)
Ultrasound	109166	90266 (25.66%)
X-Ray	78860	75909 (5.36%)

Table 3.7: Number of predicted recommendations by modality

Imaging Modality	**Example recommendation sentences**
Angiography	The patient will be followed up in the VIR clinic in approximately 2-3 weeks.
Computed Tomography	For a low risk patient, CT follow-up is recommended in 6 to 12 months. In the high risk patient, follow up is recommended at 3 to 6 months.
Fluoroscopy	Further evaluation with endoscopy is recommended.
Magnetic Resonance Imaging	BI-RADS category 6. Take appropriate action. MRI would be the best modality to assess response to neoadjuvant therapy.
Mammogram	Normal interval follow-up is recommended in 12 months.
Nuclear Medicine	Follow up nuclear medicine whole body scan is recommended in approximately 7 to 10 days after discharge.
Portable Radiography	A lateral radiograph or CT of the chest is recommended for further evaluation of this nodule.
Positron emission tomography	Follow up examination could be performed in 2 to 3 months to re-evaluate these lesions on PET.
Ultrasound	Recommend follow-up pelvic ultrasound in 2-3 months to evaluate for change.
X-Ray	Evaluation with weight bearing views is recommended.

Table 3.8: Example recommendation sentences extracted from the dataset for each modality

15.9% (523,471 reports) of the entire dataset contained recommendations. As can be observed from Table 3.7, 98.02% of mammograms included a follow-up examination. For other modalities, percentages of reports with recommendations varied between 4.17% (portable radiography) and 25.66% (ultrasound). To evaluate the performance of our recommendation extraction model, we randomly selected 40 recommendations for top 5 modalities with highest recommendation percentages: mammograms (98.02%), ultrasound (25.66%), computed tomography (19.81%), positron emission tomography (18.68%), and Magnetic Resonance Imaging (14.32%) and manually validated their correctness. We identified 185 out of 200 of those recommendations as true positives which resulted a precision value (0.925) on the target dataset similar to our 5-fold cross validation result (0.94) on the annotated set.

We then used the trained NER model to extract the entities within the predicted recommendation sentences. Not all recommendation sentences included reason, test, or time frame information. For instance, the example recommendation of Fluoroscopy presented in Table 3.8 does not have time frame entity. From 685,912 recommendations, the NER model extracted 250,840 (36.57%) reason, 528,040 (76.98%) test, and 216,128 (31.51%) time frame entities. Table 3.9 shows the distribution of predicted entities by modality.

Imaging Modality	**Reason**	**Test**	**Time frame**
Angiography	7,732	8,421	4,474
Computed Tomography	191,453	221,941	25,440
Fluoroscopy	159	125	7
Magnetic Resonance Imaging	41,136	68,452	20,679
Mammogram	24,998	250,605	162,421
Nuclear Medicine	11,895	12,476	974
Portable Radiography	15,292	15,725	367
Positron emission tomography	449	525	12
Ultrasound	82,371	134,233	36,827
X-Ray	73,383	65,115	2,894

Table 3.9: Number of predicted entities by modality

3.5 Analysis of Follow-up recommendations adherence rate

To understand the follow-up status of each identified recommendation, we performed a longitudinal analysis on the multi-institutional radiology dataset based on the information extracted by the NLP methods. Specifically, for each patient's timeline, we identified all reports with follow-up recommendations. The timestamps of the reports represented the timestamps of the recommendations. For each identified recommendation in the patient's timeline, we checked whether a radiology test with the same modality actually occurred to roughly estimate the percentage of patients who stayed within the network of two hospitals in our dataset. Table 3.10 presents the results of this initial analysis.

Imaging Modality	**# of reports with follow-up recommendation**	**No following tests of same modality**	**Had following tests of same modality**
Angiography	7234	2972 (41.08%)	4262 (58.92%)
Computed Tomography	140066	43698 (31.20%)	96368 (68.80%)
Fluoroscopy	100	84 (84.00%)	16 (16.00%)
Magnetic Resonance Imaging	34928	15791 (45.21%)	19137 (54.79%)
Mammogram	154255	45357 (29.40%)	108898 (70.60%)
Nuclear Medicine	7426	4131 (55.63%)	3295 (44.37%)
Portable Radiography	12951	3629 (28.02%)	9322 (71.98%)
Positron emission tomography	336	282 (83.93%)	54 (16.07%)
Ultrasound	90266	35067 (38.85%)	55199 (61.15%)
X-Ray	75909	22952 (30.24%)	52957 (69.76%)

Table 3.10: Number of patients who did / didn't have follow-up tests

We further analysed the patient adherence to follow-up recommendations, using the extracted time frame entities. Since the time-frame entities were free-text spans, we normalized the values by using the Stanford temporal tagger (SUTime) [72]. SUTime normalizes the temporal phrases into a value (e.g., 3 months = P3M, 1 year = P1Y). Then using the normalized time frame value for follow-up, we projected the next imaging test date for the patient. If the recommended time consists of a range such as “6 to 12 months”, we used the end of the range to project the next visit for recommended imaging test. Because some projected dates are outside of the collected time range of the dataset, we considered those radiology encounters censored (18,338 records). Furthermore, a report could contain multiple follow-up recommendations (122,256 records). If the patient did not have any one of the follow-up encounters as recommended in the report, we considered no follow-up for that report. If the patient was late to any one of the recommended follow-up encounters in the report, we considered late follow-up for that report. Table 3.11 shows the number of patients who did not have a follow-up encounter as recommended by radiologist as well as those who had a follow-up earlier or later than the recommended time.

Imaging Modality	**Reports with recommendation and projected time frame**	**No follow-up**	**Early follow-up**	**Late follow-up**
Angiography	2075	759 (36.58%)	393 (18.94%)	923 (44.48%)
CT	14506	5516 (38.03%)	4716 (32.51%)	4274 (29.46%)
Fluoroscopy	5	3 (60.00%)	0 (0%)	2 (40.00%)
MRI	8708	3393 (38.96%)	1736 (19.94%)	3579 (41.10%)
Mammogram	121716	27689 (22.75%)	19935 (16.38%)	74092 (60.87%)
NM	349	143 (40.97%)	124 (35.53%)	82 (23.50%)
Portable Radiography	222	113 (50.90%)	62 (27.93%)	47 (21.17%)
PET	7	6 (85.71%)	0 (0%)	1 (14.29%)
Ultrasound	21083	8599 (40.79%)	5060 (24.00%)	7424 (35.21%)
X-Ray	976	354 (36.27%)	233 (23.87%)	389 (39.86%)

Table 3.11: Number of patients who had no follow-up / early follow-up / late follow-up

The results shows that mammograms had the highest follow-up rate (77%: 16% early, 61% late follow-up). This is expected as mammograms are commonly used as a screening tool to detect early breast cancer in women and annual exam is recommended for women over age 40. For the other modalities, the follow-up rates varied between 14% (positron emission tomography) and 64% (X-Ray).

3.6 Summary

The main contribution of this study is identifying recommendation information in radiology notes using neural models and subsequently analysing follow-up adherence with the extracted data. We applied the trained models to a multi-institutional dataset of 3.3 million radiology notes and presented our analysis of recommendation follow-up adherence over a period of 10 years. There are several limitations in our analyses. First we assume that the follow-up exams will be performed in the UW network of care facilities, while in reality, some patients could also have imaging tests elsewhere and continue to be followed up from other providers. Second, we assume the recommended imaging test is the same one based on which the recommendation was provided. However, it is entirely possible that a different imaging test is recommended for a different diagnostic purpose. To account for this scenario, a better way is to include the recommended test entity when projecting the next encounter. In the case where a recommended time frame was not provided by the radiologist, one possible solution is to develop a document level classifier to predict the recommended time frame based on the radiology report. Another limitation of this study was the size of the training set for recommendations. Our labeled training corpus consisted of 1367 reports. To achieve good performance, sequential neural approaches require relatively larger dataset than traditional machine learning methods. Although the presented performance results were promising, there was still room for improvement in the extraction. This motivated my later research using the contextual embedding and transformer architecture such as BERT [21]. Through model pre-training on unlabelled in-domain text, the knowledge can be transferred to another task even with limited annotated labels.

Chapter 4

AUTOMATIC ASSIGNMENT OF RADIOLOGY EXAMINATION PROTOCOLS USING PRE-TRAINED LANGUAGE MODELS WITH KNOWLEDGE DISTILLATION

This chapter describes a study using the pre-trained language model, BERT [21] to classify radiology protocols. We investigated the effect of in-domain pre-training in classification performance. We also conducted different re-sampling experiments to handle the data imbalance in the dataset, and employed a new approach called knowledge distillation using augmented data.

4.1 Introduction

Imaging tests are commonly used for diagnosis and screening. They reveal conditions inside the patient's body and offer evidence to answer clinical questions. To be able to collect this evidence, physicians rely on radiologists to design a sequence of imaging scans, with specific technical parameters, such as use of intravenous or oral contrast, number of scanning planes and the orientations. Selecting these technical parameters is a main part of protocoling. Radiologists make the protocol decision based on structured data, such as patient demographics, and the unstructured clinical information in the electronic order for the suggested examination request by the ordering clinician.

This manual protocoling is an important task but can also be time consuming due to the variation in the equipment parameter setting [73], complexity in procedural terminology [74], and lack of best practice standardization [75]. Reviewing each patient record and designing an optimal protocol tailored to each patient's profile can be overwhelming and cumbersome [19]. Apart from protocol assignment, radiologists often need to attend to other responsibilities, including image interpretation, report dictation, clinical consultation, teaching, and communicating test results to referring physicians through phone calls and pages. One study showed that radiologists on average spend 3 hours in a 48-hour period to assign protocols [18]. In addition, on-call radiologists can be disrupted as many as 2 to 3 times by phone calls while interpreting routine CT examinations [76]. The distraction from phone calls not only has negative impact on radiology report turn-around times [17] but can also lead to diagnostic discrepancy [16]. Despite the potential complexity, certain examinations are very common, and protocoling these common examinations is a fairly simple and repetitive task. It is therefore a strong candidate for automation. By applying machine learning techniques to protocoling, radiologists could spend a greater proportion of their time performing interpretive tasks, thereby improving the cost-effectiveness of a radiology practice, reducing interruptions for protocoling, improve interpretation accuracy and shorten report turnaround time.

In this study, we used structured radiology exam meta-data (exam name and code provided by the referring physician) and patient demographics (age and gender) as well as unstructured diagnoses and history information to train our models. Table 4.1 presents an example of the radiology examination data from our dataset. We (1) compared different statistical ML models to the state-of-the-art BERT [21] model for radiology protocol classification task, (2) evaluated the BERT model pre-trained on general domain (BERT$_{base}$) in comparison to a BERT model pre-trained on our radiology corpus (BERT$_{rad}$), and (3) applied deep learning knowledge distillation approach to tackle high data imbalance in our dataset.

4.2 Related Work

Prior studies in automating protocol selection used machine learning approaches. Brown et al. compared three different models to classify MRI protocols, including support vector machine (SVM), gradient boosting machine (GBM), and random forest (RF) [77]. They used bag-of-words approach with unigrams to represent features for the text data and combined

Exam metadata		Demo		Patient history		Protocol
Code	Name	Sex	Age	History	Diagnosis	
CABDWC	CT ABDOMEN W CONTRAST	2	67	heart failure, hepatic vein	concern for liver laceration post procedure, post biopsy, on apixaban	BODY CT Liver 2 phase for hypervascular liver metastases (art venous, no delay)

Table 4.1: Example examination data from our dataset

them with the structured variables (age, sex, location and ordering service). The dataset consisted of 7487 observations. Since each protocol can consist of a sequence of procedures, they trained 41 binary classifiers for each model to predict each procedure in a sequence. The three ML algorithms included in this study demonstrated similar performance. GBM achieved 86% precision and 80% recall. SVM achieved 83% precision and 82% recall, followed by RF with 85% precision and 80% recall. Trivedi et al. used IBM Watson to determine the use of intravenous contrast for musculoskeletal MRI protocols using only clinical text [78]. The dataset consisted of 650 positive and 870 negative labels. Watson achieved over 90% precision and 74% recall. The overall performance is similar to their ensemble model comprising 8 traditional statistical models (SVM, scaled linear discriminant analysis, boosting, bagging, classification and regression tree, RF, Lasso and elastic-net regularized generalized linear model, maximum entropy). Although they claimed that Watson's classifier was based on deep learning, no specific details about the model architecture and hyperparameters were provided by IBM. One study conducted by Kalra et al. is the most similar to our study. They developed two statistical ML models and one deep learning model to automate CT and MRI protocol assignment. The dataset contained 18000 CT and MRI examinations in 108 unique protocols. Similar to our dataset, their protocol frequency distribution is highly imbalanced with the 5 most commonly assigned protocols making up 49% of the entire dataset. They trained a k-nearest neighbor and a random forest classifier using TF-IDF feature vectors on unigrams from clinical text. Interestingly, they excluded structured data elements such as age and gender, which could be strong predictor variables. The performance results from the top two classifiers, RF (80% precision, 82% recall) and DNN (82% precision, 84% recall), were comparable. However, they only reported weighted micro-averages and did not report performance metrics per protocol. Hence, we do not know how the model performed on the minority classes.

4.3 Methods

4.3.1 Data

Our dataset included 35,085 radiology body CT examinations performed at 7 hospital-based and clinic-based imaging sites between January 2018 and June 2019. The data were extracted from the University of Washington radiology information system. As shown in Table 4.1, each exam is represented with 4 structured data fields including exam meta-data (exam code, protocol code) and patient demographics (age, gender) as well as 2 unstructured fields to capture patient clinical history (history, diagnosis). Table 4.2 describes the word level statistics on the two unstructured fields. Our initial analysis showed that the lengths of the unstructured data were relatively short (average numbers of words for history and diagnosis fields were 8 and 10 with standard deviations 6.57 and 8.6 respectively). 4759 (13.6%) examinations contained no history data and 3 (0.01%) examinations contained no diagnosis data.

	Min	**Max**	**Mean**	**Median**	**Standard deviation**
History	0	47	8	6	6.57
Diagnosis	0	108	10	8	8.6

Table 4.2: Word statistics on unstructured fields.

Due to the different naming and coding of the same protocols in different clinical sites, our radiologists consolidated them into 27 unique “protocol groups”, each uniquely identify the protocol with similar acquisition parameters. Generally, the protocols were categorized by the anatomical region, administration of contrast, number of pre and post-contrast phases and scan range. We excluded 2 groups that had less than 20 examinations in our experiments (CT CA Oral Only and CT Abdomen IV Only). Table 4.3 shows the examination frequency with percentages for each protocol group. As can be observed, the dataset is highly imbalanced, with the first two protocol groups constituting 57% of the entire dataset. The distribution of examination frequency among the groups has a mean of 1299, median of 200 and standard deviation of 2706.

	Protocol group	**Frequency**	**%**
1	CT CAP IV and Oral	11911	33.95%
2	CT Abdomen Pelvis w IV Only	8057	22.96%
3	CT CAP IV Only	3351	9.55%
4	CT Abdomen Pelvis w IV and Oral	2941	8.38%
5	CT Renal Mass	2036	5.80%
6	CT Liver 3 Phase	1652	4.71%
7	CT Abdomen Pelvis No Contrast	931	2.65%
8	CT IVP 50 yrs +	854	2.43%
9	CT CAP Oral Only	531	1.51%
10	CT CAP No Contrast	336	0.96%
11	CT Abd Pel Enterography	297	0.85%
12	CT Liver 4 Phase	252	0.72%
13	CT CA IV Only	226	0.64%
14	CT IVP < 50	220	0.63%
15	CT Pancreas Mass 3 Phase	202	0.58%
16	CT Abdomen No Contrast	195	0.56%
17	CT CA IV and Oral	194	0.55%
18	CT Pelvis IV Only	192	0.55%
19	CT Abdomen IV and Oral	173	0.49%
20	CT Pancreas Mass 2 Phase	143	0.41%
21	CT Abdomen Pelvis w Oral only	132	0.38%
22	CT CA No Contrast	75	0.21%
23	CT Pelvis Cystogram	68	0.19%
24	CT Liver 2 Phase	51	0.15%
25	CT Pelvis IV and Oral	42	0.12%
26	CT CA Oral Only (excluded)	15	0.04%
27	CT Abdomen IV Only (excluded)	8	0.02%

Table 4.3: Distribution of examinations across protocols

4.3.2 Approach

We trained a classifier to automatically assign protocols to computer tomography (CT) examinations. The classifier was implemented by fine-tuning the pre-trained language model, BERT [21]. The fine-tuning process followed Devlin et al.'s suggestion to use a linear layer on top of the BERT model and train with the cross-entropy loss. Because BERT is a language model, we therefore first transformed the structured and unstructured data into the following template: "Exam is <exam code>. Sex is <gender>. Age at Exam <age>. History: <history>. Diagnosis: <diagnosis>" and subsequently classifying it into one of 25 protocol groups listed in Table 4.3. We observed that the mean and median of number of characters in the templated data are 192 and 178. In order to capture context presented in the training instances, we set the maximum sequence length parameter of the BERT model to be 200 with a batch size of 48. We followed the suggestions described in the BERT paper and used the Adam optimizer with a learning rate of 2-e5. We fine-tuned the BERT model for 4 epochs.

Conceptually, BERT learns the relationships between words by randomly masking words in a sequence with a [MASK] token and then trains itself to predict them from the context of the unmasked ones. Additionally, it learns the sentence relationships by training itself to predict if two sentences are adjacent to each other. These two learning tasks allow BERT to self-train and capture the context of language used in an unlabeled corpus before transferring all parameters to down-stream applications. Previous studies showed promising results of using BERT in clinical applications. Examples include chest x-ray reports classification [79], and relation extraction in clinical domain [80].

In this study, we first experimented with the google BERT_{base} model which was originally pre-trained on BookCorpus and English Wikipedia. However, by fully encoding the semantic context in clinical and biomedical text, it has been shown that further training BERT_{base} on MIMIC and PubMed data can boost the performance of named entity recognition in the biomedical domain [81, 45]. Inspired by these studies, we further pre-trained BERT_{base} on our radiology protocol corpus and named it BERT$_{rad}$. We repeated the same experiment with BERT$_{rad}$ using the same hyperparameters listed above. All BERT experiments were implemented with Huggingface's transformer library [82].

Knowledge Distillation

Imbalanced class distribution usually leads to poor classification results on the minority classes [83]. A popular approach in dealing with imbalanced datasets is to use the Synthetic Minority Oversample Technique (SMOTE) which generates new artificial samples for the minority classes by interpolating the nearest neighbors of the existing samples [84]. This method reduces the likelihood of overfitting minority classes commonly observed in random over sampling approach. Because BERT utilizes WordPiece tokenization which incorporates special tokens, such as the classifier token [CLS] and separator token [SEP], synthesizing these input values in vector space using interpolation will lose the context of the tokens in the samples.

Recent studies have successfully demonstrated the possibility of transfering task specific knowledge from the large BERT model to a smaller neural architecture without significantly degrading performance [85, 86], using a technique called knowledge distillation. The process involves training a second model (student) to match the predictions from the first model (teacher). We hypothesized that by transferring knowledge specific to the minority classes from the BERT$_{rad}$ model to a second BERT model, we could improve the classification performance on the minority classes. In particular, we aimed to train a student model that could outperform the teacher with identical neural architecture. Furlanello et al. referred to this approach as Born-Again Neural Network (BAN) [87], which has been shown to produce better results in both single and multi-task settings [88]. During the knowledge distillation process, the raw predictions from the teacher model, known as logits, are being used as "soft labels" for training the student model. The distribution in the logits, even among incorrect predictions, contains information about how the teacher model is generalizing, thereby offering more training signals than one-hot categorical labels [85].

To effectively transfer knowledge about the minority classes to a student model, a large unlabeled dataset is needed to generate enough soft labels from the teacher model. We applied Tang et al.'s data augmentation techniques to synthesize masked data in order to allow the teacher to fully express its knowledge [89]. To augment a given training instance, we randomly sampled a number P from the uniform distribution [0,1]. If P < 0.1, we randomly replaced a word in the history and diagnosis section with the [MASK] token. If P is between 0.1 and 0.2, we randomly replaced a word with another word in the training set that has the same POS tag. Finally, we randomly replaced an n-gram ($n \in [1, 3]$) in the training instance with the [MASK] token. We repeated this augmentation process to generate 30 new instances, without duplication, for each training instance. We evaluated different numbers of augmented instances (25, 30, 35, 40, 50) by running 5-fold cross validation with the augmented data. Our evaluation showed that the experiment with 30 augmented instances achieved the best result. To limit the augmented sample size of the dominant classes, we set a maximum sampling limit of 12000, such that the final sample size of each class after augmentation could not exceed 12000. We then ran inferencing on the augmented dataset using the teacher model BERT$_{rad}$ to generate soft labels for distillation. Finally, we initialized a student BERT$_{rad}$ model with a different random seed and trained it to imitate the teacher by minimizing the mean squared error (MSE) between the student's logits and teacher's logits. At the same time, we allowed the student model to surpass the teacher by training with the true labels by minimizing the cross-entropy loss against the one-hot multi-class labels:

$$L_{distill} = \alpha * L_{cross\_entropy} + (1 - \alpha) * L_{MSE}$$

where α is the ratio of true labels within a single batch of training samples. After each iteration of knowledge distillation, the student model became the teacher for next generation.

To establish the baselines, we trained three separate statistical models that were employed in prior research studies: Support Vector Machine (SVM), Gradient Boosting Machine (GBM) and Random Forest (RF). The feature sets included the unigrams and bigrams of the history and diagnosis notes and were transformed into vector space using TF-IDF before combining with the numeric values in the structured data.

4.4 Results

We used 5-fold cross validation to evaluate the general performance of the models. For each fold, the models were trained on the same training data and evaluated on the same held-out test data. We used precision, recall, and F1-score as metrics to measure the performance. Table 4.4 presents the overall macro-average and weighted micro-average results. The macro-averaged results were the mean of the metrics for each class, where each class was given equal weight. The weighted micro-averaged metrics were the metric averages weighted by the number of true labels in each class. As can be observed, the micro-average results are largely similar due to the bias towards the majority classes. In the macro-average results, among the baselines, RF performed the best with 0.60 F1-Score. Both SVM and GBM produced 0.45 F1-score. The SVM in general produced higher precision and lower recall, when compared to GBM. The classifiers based on BERT models performed better than the SVM, GBM and RF baselines. Furthermore, the in-domain BERT$_{rad}$ produced 0.2 higher macro F1 score than the out-of-domain BERT$_{base}$ model (0.63 versus 0.61).

Model	**Macro average**			**Micro (Weighted) average**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
SVM	0.60	0.42	0.45	0.79	0.80	0.79
GBM	0.46	0.46	0.45	0.80	0.81	0.80
RF	0.63	0.59	0.60	0.83	0.83	0.83
BERT$_{base}$	0.68	0.60	0.61	0.84	0.84	**0.84**
BERT$_{rad}$	0.67	0.62	**0.63**	0.84	0.84	**0.84**

Table 4.4: Comparison of model results.

To mitigate the high data imbalance, two resampling experiments were conducted with the best performing BERT$_{rad}$. Note that the resampling was performed only on the training data while the validation data were kept the same. First, we under-sampled the 2 majority classes by randomly removing some training instances such that their sample sizes matched the size of the third largest protocol group (#3). As shown in Table 4.5, the macro-average F1 dropped 0.24 and the weighted-average F1 dropped 0.2 due to the misclassification of the majority classes given their smaller sample sizes. In the over-sampling experiment, the training instances in the minority classes were randomly replicated so that their sample sizes matched the size of the second largest protocol group (#2). The result shows no performance improvement in the macro-average F1 but degradation in the weighted-average. This can be caused by overfitting the duplicate training samples in the minority classes. Using knowledge distillation, the BAN models{2,3} achieved better macro-average performance than BERT$_{base}$ and BERT$_{rad}$, without any degradation in weighted-average performance. More specifically, the macro-average F1 in generations of student BAN models improved, suggesting that the classifiers achieved better performance in predicting the minority classes. We also observed that the performance saturated after training the second generation of BAN student model. This finding is also observed by Furlanello et al [87].

Model	**Macro average**			**Micro (Weighted) average**		
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
BERT$_{rad}$	0.67	0.62	0.63	0.84	0.84	**0.84**
BERT$_{rad}$ undersample	0.42	0.38	0.39	0.63	0.66	0.64
BERT$_{rad}$ oversample	0.63	0.63	0.63	0.83	0.82	0.82
BAN1	0.68	0.64	0.65	0.84	0.84	**0.84**
BAN2	0.69	0.65	**0.66**	0.84	0.84	**0.84**
BAN3	0.69	0.65	**0.66**	0.84	0.84	**0.84**

Table 4.5: Comparison of resampling results. BAN1,2,3 denotes the 1st, 2nd and 3rd generation of knowledge distillation.

4.5 Analysis of Protocol level classification results

We observed that BERT$_{rad}$ performed better than BERT$_{base}$ in some protocol groups. For example, as shown in Table 4.6 in the protocol group "CT Abd Pel Enterography" (#11), the word hernia, which describes the condition that a tissue pushes through an abdominal opening, appeared in over 79% of the diagnosis fields, while another word CREATININE, a compound that indicates the level of kidney function, appeared in over 73% of the history fields. These two medical terms are not commonly seen in the general corpora. By pre-training on the radiology corpus, BERT$_{rad}$ was able to learn better contextual representation and outperformed BERT$_{base}$ by 0.07 F1 in that protocol group. We observed similar improvement in protocol groups (#19) and (#21).

Protocol group	**Exam count**	**SVM**	**GBM**	**RF**	**BERT$_{base}$**	**BERT$_{rad}$**	**BAN1**	**BAN2**	**BAN3**
11. CT Abd Pel Enterography	59	0.54	0.41	0.5	0.53	0.6	**0.63**	0.61	0.62
19. CT Abdomen IV and Oral	35	0.06	0.03	0.36	0.41	0.45	**0.48**	**0.48**	**0.48**
21. CT Abdomen Pelvis w Oral only	26	0	0.05	0.26	0.23	0.37	**0.39**	0.38	0.37

Table 4.6: Comparison between BERT$_{base}$ and BERT$_{rad}$

Table 4.7 shows that the minority groups classification were improved by the BAN models.

One interesting observation, shown in Table 4.8, was BERT$_{base}$ model's substantially low F1-score of 0.16 for group "CT IVP $<$ 50" (#14) when compared to the F1-scores (SVM: 0.61, GBM: 0.73, RF: 0.88) of statistical baselines. Further investigation showed that 87% of the false negatives for "CT IVP $<$ 50" (#14) were misclassified to "CT IVP 50 yrs +" (#8) by BERT$_{base}$. The main difference between these two protocol groups is the age of patient, and the age feature by itself offered high information gain to allow RF to learn a more robust model. On the other hand, the smaller sample size of protocol group #14

Protocol group	**Exam count**	**SVM**	**GBM**	**RF**	**BERT$_{base}$**	**BERT$_{rad}$**	**BAN1**	**BAN2**	**BAN3**
9. CT CAP Oral Only	107	0.29	0.56	0.31	0.58	0.59	0.59	**0.61**	**0.61**
15. CT Pancreas Mass 3 Phase	41	0.46	0.36	0.58	0.64	0.62	0.63	**0.67**	0.65
19. CT Abdomen IV and Oral	35	0.06	0.03	0.36	0.41	0.45	**0.48**	**0.48**	**0.48**
23. CT Pelvis Cystogram	14	0.69	0.27	0.95	0.95	0.95	0.95	0.95	**0.96**

Table 4.7: Improvement on the minority groups

limited the BERT$_{base}$ model to learn to differentiate from protocol group #8. However, data augmentation in the knowledge distillation process eventually supplied additional training signals for the model to generalize, leading to the similar performance levels as RF.

Protocol group	**Exam count**	**SVM**	**GBM**	**RF**	**BERT$_{base}$**	**BERT$_{rad}$**	**BAN1**	**BAN2**	**BAN3**
8. CT IVP 50 yrs +	171	0.81	0.9	0.92	0.84	0.84	0.91	**0.93**	0.92
14. CT IVP < 50	44	0.61	0.73	**0.88**	0.16	0.24	0.78	**0.88**	0.87

Table 4.8: Misclassifcation by BERT$_{base}$

One protocol group "CT Liver 2 Phase" (#24) in particular was difficult to predict by any models, as shown in Table 4.9. The error analysis showed that the models misclassified some 24 cases to "CT Liver 3 Phase" (#6) because of similar patient diagnosis and history.

Protocol group	**Exam count**	**SVM**	**GBM**	**RF**	**BERT$_{base}$**	**BERT$_{rad}$**	**BAN1**	**BAN2**	**BAN3**
24. CT Liver 2 Phase	10	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00

Table 4.9: Poor classification performance on one protocol group

Table 4.10 presents one of these cases. While these were the correct protocol assignments in clinical practice, because #24 only constituted 0.15% of the training data and was 30 times less than #6, there were not enough data to train the models to differentiate #24 from #6. Additionally, we found that some #24 cases were misclassified to "CT CAP IV and Oral" (#1) because of the exact same history and diagnosis found in #24. Without any additional clinical information to help differentiate the two protocol assignments, the models simply inferred to the group that was more dominant in the training data.

Protocol group	**History**	**Diagnosis**
6. CT Liver 3 Phase	Last creatine level:CREATININE 0.92	ABDOMEN W/CONTRAST; 6MO REPEAT F/U FOR HCC SURVEILLANCE, S/P LIVER TRANSPLANT
24. CT Liver 2 Phase	Last creatine level:CREATININE 0.81	ABDOMEN W/CONTRAST; TO EVALUATE SIZE OF PSEUDOCYST, S/P LIVER TRANSPLANT

Table 4.10: Examinations in two different protocol groups with similar history and diagnosis.

4.6 Summary

This study explores using the state-of-the-art pre-trained language model, BERT, to classify radiology examination protocols. The results shows that overall pre-trained language models perform better than traditional n-gram models. The in-domain pre-training allows the model to develop better contextual representations. Additionally, the results demonstrate that knowledge distillation with augmented data improves overall classification performance for most of the under-represented groups. In the next study, we will apply the in-domain pre-training of BERT to a much larger scale and develop a more robust IE deep learning pipeline to extract different clinical findings from radiology reports.

Chapter 5

EVENT-BASED CLINICAL FINDINGS EXTRACTION FROM RADIOLOGY REPORTS WITH PRE-TRAINED LANGUAGE MODEL

This chapter describes a study using pre-trained language models to extract two clinical findings (Lesion and Medical Problem) from radiology reports. A new corpus consisting of 500 computed tomography (CT) radiology reports were annotated using an event-based schema to capture fine-grained details of both clinical findings. In addition, a general-purpose deep learning framework was developed to fine-tune a BERT model in a multi-task fashion. We pre-trained a new BERT$_{rad}$ model on the 3.3 million multi-institutional corpus from our first study [90] (Chapter 3) . To demonstrate the generalizability of the model with cross-institutional data and imaging modality, we fine-tuned the BERT$_{rad}$ with the 500 CT corpus and used it to extract all the clinical findings from the chest X-ray reports from the MIMIC Chest X-ray (MIMIC-CXR) database [22].

5.1 Introduction

Radiology reports remain the primary channel of communication for radiologists to documents their findings in imaging tests. Extracting clinical findings from these unstructured narratives facilitates many secondary use applications, including clinical decision-support systems [91], diagnostic surveillance of medical problems [92], identification of patient cohorts with specific phenotypes [48], and simplification of report language for patients [93]. To support various types of applications in large scale, a detailed semantic representation of the findings is needed to capture the important information in the findings, such as anatomy, assertion, characteristics and size.

This chapter discusses a novel event-based annotation schema that focuses on two clinical findings: Lesion and Medical Problem. A lesion finding is defined as an abnormal space occupying mass that is observable on the images. Lesions included primary tumors, metastases, benign tumors, abscesses, nodules, other masses. A medical problem finding is a pathological process describing other types of clinical problems, for example cirrhosis, air-trapping, atherosclerosis, effusions. A new corpus of 500 computed tomography (CT) randomly sampled from the UW clinical data warehouse was annotated. The gold standard included 2,344 Lesion and 8,065 Medical Problem finding events. To extract the finding events, we developed a deep learning extraction framework that fine-tuned a single BERT model. We explored different contextualized embeddings through pre-training on different clinical text sources and introduced a new BERT model that was pre-trained with the multi-institutional corpus in the first study, coverage a wide range of modalities (Table 3.2). To assess the generalizability of the event extraction model, we annotated a subset of the MIMC-CXR radiology reports. The extraction model achieved comparable performance on the MIMIC-CXR and UW datasets, despite the differences between the datasets. The extracted MIMIC-CXR clinical findings, the annotation guidelines and the event extraction framework are made available to the public.

5.2 Related Work

5.2.1 Clinical finding entity extraction from radiology reports

Numerous studies have applied rule-based patterns to extract clinical entities specific to certain diseases, including appendicitis [94], adrenal abnormalities [95], osteoporosis [96], and pneumonia [92]. Other studies employed statistical machine learning approaches. Hassanpour et al. extracted anatomy, observations, modifiers, uncertain expressions using Conditional Markov Model and Conditional Random Field from a corpus of 150 chest CT reports [55]. Cheng et al. combined rule-based and statistical methods to identify tumor status, magnitude of change and significance of change from a corpus of 778 MRI reports. Recent research employed neural network approaches. Cornegruta et al. extracted 4 different entities (body location, clinical finding, descriptor and medical device) with a corpus of 2000 radiology reports using BiLSTM network [58]. Most state-of-the-art neural modeling however used the pre-trained language model, BERT. Sugimoto et al. extracted 7 different clinical terms from a corpus of 540 Japanese CT radiology reports using a pre-trained Japanese BERT model [61]. Miao et al. extracted entities associated with the Breast Imaging Reporting and Data System (BI-RADS) from an annotated corpus of 540 Chinese ultrasound reports [97]. While these studies demonstrated the effectiveness of extracting finding entities from radiology reports, they did not attempt to identify the association or relation between the entities.

5.2.2 Clinical finding relation extraction from radiology reports

Identifying the relationships between entities provides more contextual information associated with clinical finding. For example, an anatomical location can be associated with one or more tumors, or a negative assertion can indicate the absence of a clinical finding. Early studies used rule-based methods with lexicons and grammars to extract clinical finding relations from radiology reports. Sevenster et al. identified the relations between finding observations and body locations using MedLEE [54]. Savova et al. used Mayo's clinical Text Analysis and Knowledge Extraction System (cTAKES) to extract evidence entities and assertion relations of peripheral arterial disease cases from 455 reports [98]. Other studies employed statistical approaches to extract disease specific entities and relations from radiology reports, including metastatic lung disease [99], and hepatocellular carcinoma [57]. Recent relation extraction work used the recurrent neural network (RNN) model. Steinkamp et al. extracted clinical finding observations and their relations to modifier entities, such as location, size and change over time using GRU [100]. Most recent state-of-the-art relation extraction work used the BERT model. Zhang et al. fine-tuned a BERT model to extract both breast cancer entities and relations from a corpus of 600 Chinese clinical notes (100 radiology reports) [62].

5.3 Methods

This section describes (1) the event-based annotation schema, (2) the event evaluation scoring method, (3) a new corpus annotated based on the schema, and (4) the new extraction framework that can fine-tune a BERT model to extract both entities and relations.

5.3.1 Data

We collected 706,908 CT reports between 2008 and 2018 from our UW clinical data warehouse. We randomly sampled 500 reports from the collection, and annotated as our gold standard corpus.

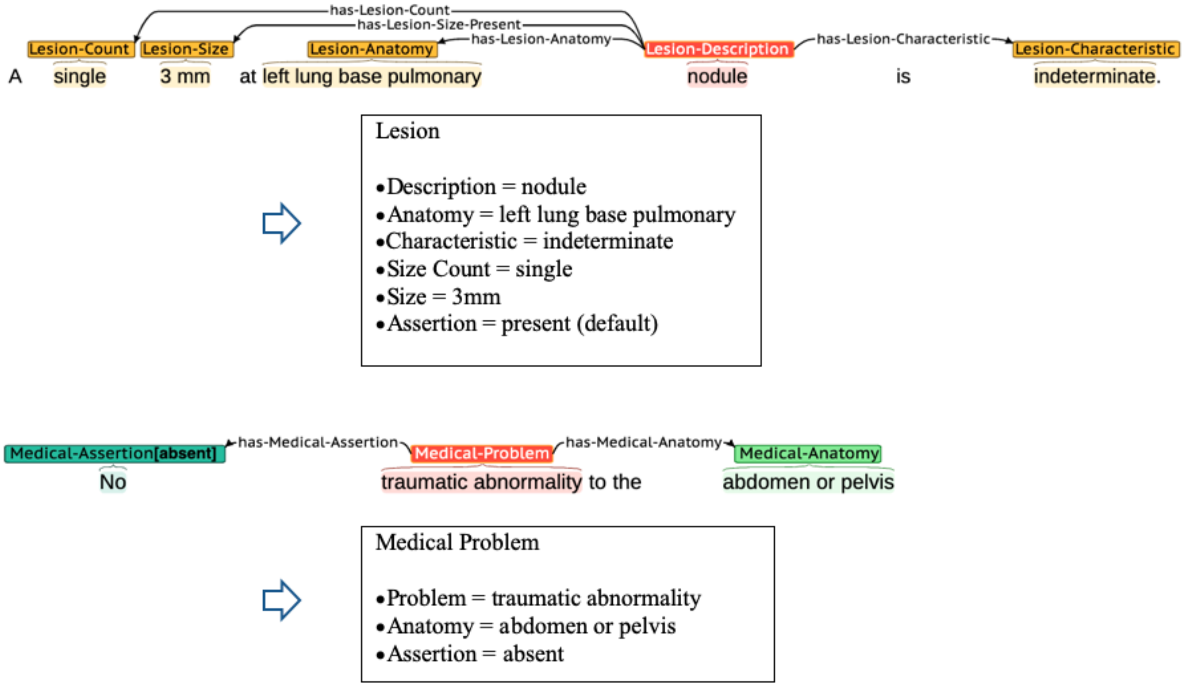
5.3.2 Annotation schema

An event-based representation was used to capture the details of the two clinical findings. Each event was characterized with a trigger and a set of connected arguments. The trigger was a required key phrase identifying the finding event, while the arguments provided fine-grained details about the event. The arguments were linked to the corresponding triggers through argument roles, forming a detailed and nuanced semantic representation of the clinical findings. A finding event comprised two types of arguments: *span-only* and *span-with-value*. The annotation of *span-only* arguments included the selection of the relevant phrase and connection to the trigger. The annotation of *span-with-value* arguments included the selection of the relevant phrase and connection to the trigger, as well as the assignment of a categorical label that captures the clinical meaning of the selected phrase (e.g., assertion). The categorical labels normalized the contents of the annotated phrase, allowing the extracted information to more easily be incorporated into secondary use applications. For example, annotating the phrase ”concern for” as Assertion would include the assignment of the categorical label *possible* (Please refer to Appendix A for example sentences). Because the presence of a lesion or medical problem could be implied rather than explicit, the present label of the argument Assertion was the default value for finding events, unless a *possible* or *absent* label was explicitly annotated. The annotation schema is summarized in Table 5.1.

	Argument	**Type**	**Value**	**Example**
Lesion Finding	Lesion Description (Trigger)	span-only	-	"mass", "lesion", "nodule"
	Anatomy	span-only	-	"left lower lobe"
	Assertion	span-with-value	present (default), absent, possible	"no", "possible"
	Characteristics	span-only	-	"hypodense", "septal"
	Count	span-only	-	"2", "numerous", "multiple"
	Size	span-only	-	"4.1 x 3.1 cm", "small"
	Size Trend	span-with-value	new, increasing, decreasing, no-change	"stable", "unchanged"
Medical Problem Finding	Medical Problem (Trigger)	span-only	-	"atherosclerotic calcifications"
	Anatomy	span-only	-	"abdominal aorta", "right kidney"
	Assertion	span-with-value	present (default), absent, possible	"no", "possible"

Table 5.1: Annotation schema of lesion finding and medical problem finding.

Extraction of these findings was treated as a slot filling task by the text spans that corresponded to the arguments (argument entities with roles) of the clinical finding events. Figure 5.1 presents example annotations for a Lesion event and a Medical Problem event. For *span-only* arguments, the slot values would be the identified text spans. For *span-with-value* arguments, the slot values would be the identified categorical labels, which captured the meaning of the annotated phrase. A finding event might include multiple arguments of the same type. For example, a medical problem could be linked to multiple anatomical locations, or a lesion could be described by multiple characteristics.



Lesion

- Description = nodule
- Anatomy = left lung base pulmonary
- Characteristic = indeterminate
- Size Count = single
- Size = 3mm
- Assertion = present (default)

Medical Problem

- Problem = traumatic abnormality
- Anatomy = abdomen or pelvis
- Assertion = absent

Figure 5.1: Example annotations for Lesion and Medical Problem events.

5.3.3 Evaluation

Inter-annotator agreement and model extraction performance was evaluated using the same scoring criteria. The annotated and extracted events include trigger and argument entities that are connected through argument roles. The pairing of triggers and arguments (entities with identified roles) assembles events from the individual entities. The scoring criteria for trigger and argument entities and argument roles are presented below.

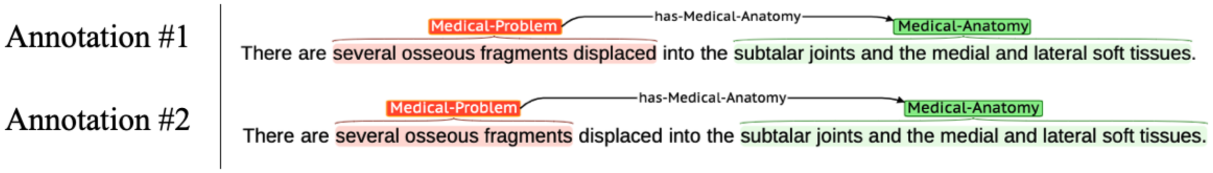
Trigger and argument entities

Trigger and argument entities scoring considered the span identification and labeling, without considering the roles linking trigger and argument entities. All trigger and argument entities were compared at the token-level (rather than span-level) to allow partial matches, since partially matched text spans could still contain clinically relevant information, e.g. "mass lesions" vs "lesions".

Argument roles

Argument role scoring considered three annotated/extracted phenomena: (1) the trigger entity, (2) the argument entity, and (3) the argument role (linking the trigger-argument entity pair). Argument role equivalence required the trigger entity, argument entity, and role label to be equivalent. In argument role scoring, the entity equivalence criteria for triggers, span-only arguments, and span-with-value arguments were based on their semantics in the event representation, and the most salient information being captured by the entities.

Trigger: Events were aligned based on trigger equivalence, and the arguments associated with aligned events (events with equivalent triggers) were compared based on the argument types. Triggers were considered equivalent if the spans overlapped by at least one token. Figure 5.2 shows an example of two Medical Problem annotations. Although the word "displaced" is not part of the trigger in Annotation 2, their overlapping text spans and connections to the Medical-Anatomy argument entities indicates that both argument entities belong to the same event and can be scored accordingly.

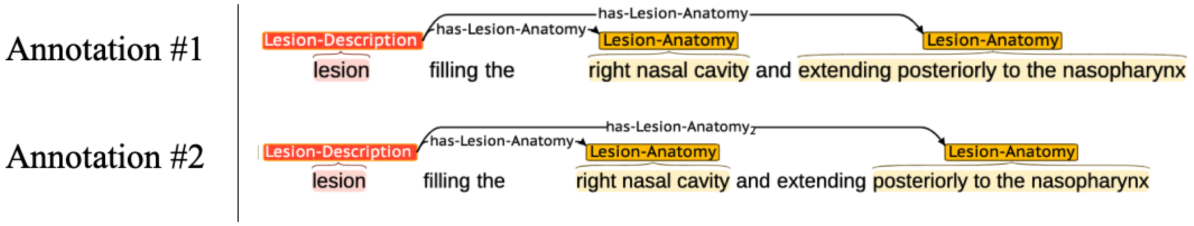


There are several osseous fragments displaced into the subtalar joints and the medial and lateral soft tissues.

There are several osseous fragments displaced into the subtalar joints and the medial and lateral soft tissues.

Figure 5.2: Two Medical Problem Finding event annotations with equivalent triggers.

Span-only: When evaluating argument role performance, span-only argument entity equivalence was assessed at the token-level rather than span-level, because partial matches can capture clinically relevant information. The example in 5.3, includes the same sentence with two sets of annotations for a Lesion event with multiple Lesion-Anatomy arguments. The second Lesion-Anatomy entities in the annotation do not match exactly. The discrepancy between the Lesion-Anatomy annotations (”extending” in Annotation 1) includes some clinical information; however, a majority of the clinically relevant information is captured by both spans (”posteriorly to the nasopharynx”). The token-level equivalence criteria for span-only argument entities was intended to reward such partial matches.



lesion filling the right nasal cavity and extending posteriorly to the nasopharynx

lesion filling the right nasal cavity and extending posteriorly to the nasopharynx

Figure 5.3: Two Lesion Finding events with partially matched span-only arguments.

Span-with-value: The categorical labels of span-with-value argument normalized the contents of the annotated phrase, allowing the extracted information to more easily be incorporated into secondary use applications. When evaluating argument role performance, the span-with-value argument entity equivalence was assessed based on the categorical labels only, without considering the spans. In Figure 5.4, although the Lesion-Size-Trend argument entity in Annotation 2 does not include the words “and number”, both Lesion-Size-Trend annotations have the same categorical label and slot value (increasing). Hence both annotations are considered equivalent.

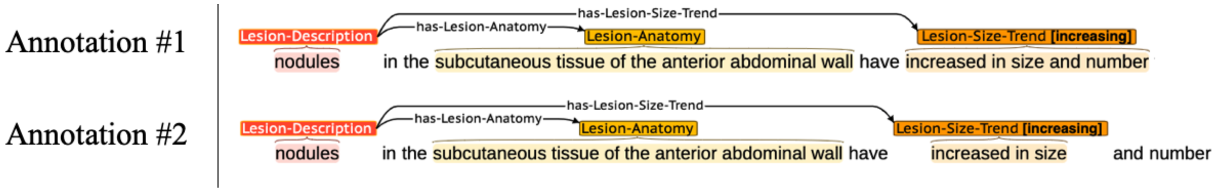

Annotation #1
has-Lesion-Size-Trend
has-Lesion-Anatomy
Lesion-Description
Lesion-Anatomy
Lesion-Size-Trend [increasing]
nodules in the subcutaneous tissue of the anterior abdominal wall have increased in size and number
Annotation #2
has-Lesion-Size-Trend
has-Lesion-Anatomy
Lesion-Description
Lesion-Anatomy
Lesion-Size-Trend [increasing]
nodules in the subcutaneous tissue of the anterior abdominal wall have increased in size and number

Figure 5.4: Two Lesion Finding event annotations with the same value for Lesion-Size-Trend.

5.3.4 Annotation agreement

The annotation was performed by one medical student and one graduate student using the BRAT rapid annotation tool [101]. An annotation guideline was provided to describe the details of each clinical finding event. Our annotators were trained in the initial iterations. They were given the same samples to annotate independently. After each iteration, they met to discuss disagreements and consulted a radiologist for advice. The annotation guideline was updated accordingly.

Inter-annotator agreement was evaluated using the same event scoring criteria described in the previous section. The agreement was calculated using pair-wise F1 score [102]. After four iterations, the final inter-annotator agreement over 30 CT reports was 93.0% F1 for triggers, 83.7% F1 for span-only arguments, and 86.9% F1 for span-with-value arguments. The medical student then annotated the remaining 470 CT reports.

5.3.5 Gold standard corpus

The final corpus contained 2,344 Lesion events (6,337 arguments and 6,617 roles), and 8,065 Medical Problem events (5,783 arguments and 7,406 roles). The argument counts represented the number of annotated spans (entities), and the role counts indicated the number of trigger-argument pairings. Since an argument could be linked to multiple triggers, the argument role counts could be greater than the argument counts. The distributions of the annotated arguments and roles are shown in Table 5.2. The number of annotated Medical Problem events was more than 3 times higher than the number of Lesion events. In general, each argument corresponded to a single role in the event, with the exception of Lesion-Size, which could be either identified as the size at the present time or in the past.

Trigger/Argument	**Frequency**	**Argument role**	**Frequency**
Lesion-Description	2,344	-	
Lesion-Anatomy	2,039	Lesion-Anatomy	2,187
Lesion-Assertion	945	Lesion-Assertion	1,008
Lesion-Characteristic	1,931	Lesion-Characteristic	1,968
Lesion-Count	235	Lesion-Count	237
Lesion-Size	816	Lesion-Size-Past	94
		Lesion-Size-Present	736
Lesion-Size-Trend	371	Lesion-Size-Trend	387
Medical-Problem	8,065	-	
Medical-Anatomy	2,990	Medical-Anatomy	3,952
Medical-Assertion	2,793	Medical-Assertion	3,454

Table 5.2: Event annotation statistics.

Overall gold standard corpus statistics are presented in Table 5.3. On average, there were 16 Medical Problem events and 5 Lesion events in a radiology report. Some radiology reports in the gold standard were very dense and contained over 100 Medical Problem events.

	Minimum	**Mean**	**Median**	**Maximum**
Number of words per report	50	327	288	1383
Number of events per report	2	21	18	130
Number of Medical Problem events per report	0	16	13	129
Number of Lesion events per report	0	5	3	36
Number of arguments per Medical Problem event	0	1	1	5
Number of arguments per Lesion event	0	3	3	16

Table 5.3: Gold standard corpus statistics.

5.3.6 Event extraction

The finding events were extracted in two separate steps: (1) the trigger and argument entities were extracted and (2) the argument roles were identified by connecting extracted trigger and argument entities through relations. The pairing of the trigger and argument entities through the argument roles assembles events from the individual entity extractions. Our event extraction pipeline operated on sentences, which were treated as independent samples.

Trigger and argument entity extraction

The extraction of trigger and argument entities was defined as a NER task. We evaluated two state-of-the-art neural network architectures: BiLSTM-CRF [103] and BERT [21].

BiLSTM-CRF:

BiLSTM-CRF was considered a strong NER baseline by multiple studies [61, 62, 104]. We used the open source NeuroNER [71] for the BiLSTM-CRF implementation, which was also used in our first study [90] (Chapter 3). In the BiLSTM-CRF architecture, each token in the input sentence was represented by the concatenation of a pretrained word embedding and a character-aware word embedding. The character-aware word embedding was generated by a BiLSTM operating on the individual characters associated with each token. The character-aware word embedding enabled the model to learn the morphological structure in each word and to encode out-of-vocabulary tokens. The word sequence was then encoded using a second BiLSTM layer to create a contextualized representation of the sentence. The label of each word was predicted by a CRF output layer which took into account the conditional dependencies across the neighboring labels. To create input data for the NER model from our annotated corpus, a series of preprocessing steps was taken. First, each annotated report was segmented into sentences. We used the Begin, Inside, Outside (BIO) tagging schema, based on whether the token was at the beginning, inside or outside of a label. For instance, consider the sentence “Probable malignant pancreatic mass with no evidence of vascular encasement”. The labels would be classified by the model, as illustrated in Figure 5.5.

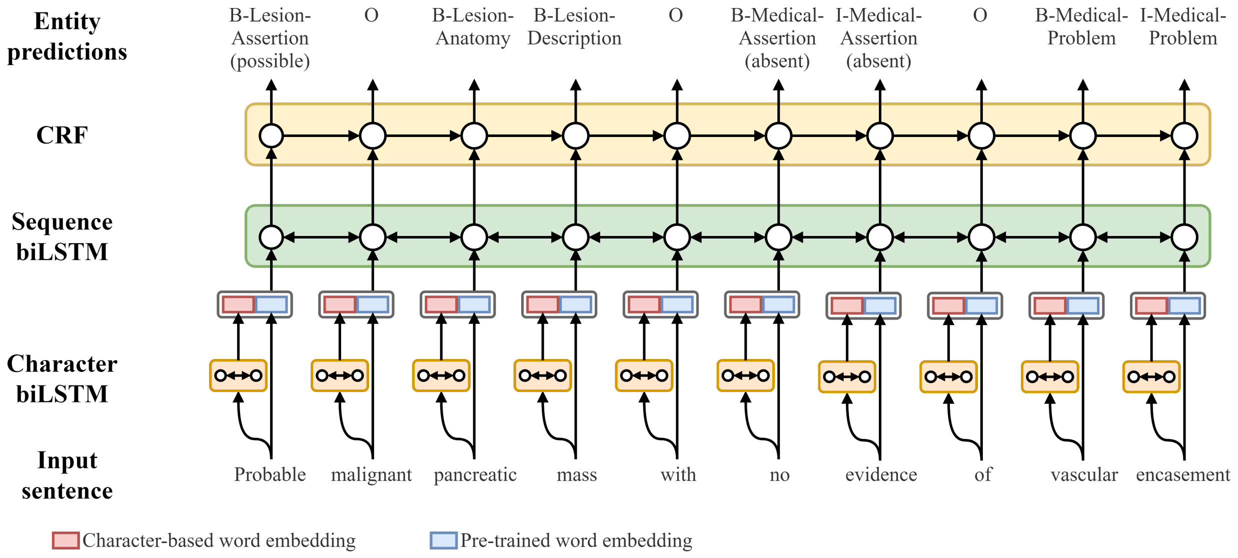


Figure 5.5: Architecture of the NeuroNER BiLSTM-CRF model.

BERT NER:

The BERT NER model was implemented by adding a single linear layer to the BERT output hidden states and fine-tuning a pre-trained BERT model, as described by Devlin et al. [21]. To prepare input data for the BERT NER model, the reports were also segmented into sentences. Because BERT utilized WordPiece tokenization [105], each word that was not in the BERT vocabulary would be segmented into multiple sub-tokens. These sub-tokens, prefixed by "##" if not the first sub-token, allowed the segments of the words to be represented in a deterministic fashion. Rather than representing all the out-of-vocabulary words with a universal token like [UNK], the sub-token representation provided richer contextual embeddings for the model to generalize. During the BIO labeling, the sub-tokens starting with "##" were assigned a special label #. In addition, the BERT input included the special tokens [CLS] and [SEP] at the beginning and end of a sentence respectively, to signify the sentence boundaries. Figure 5.6 illustrates how the labels of an input sentence were classified by BERT NER.

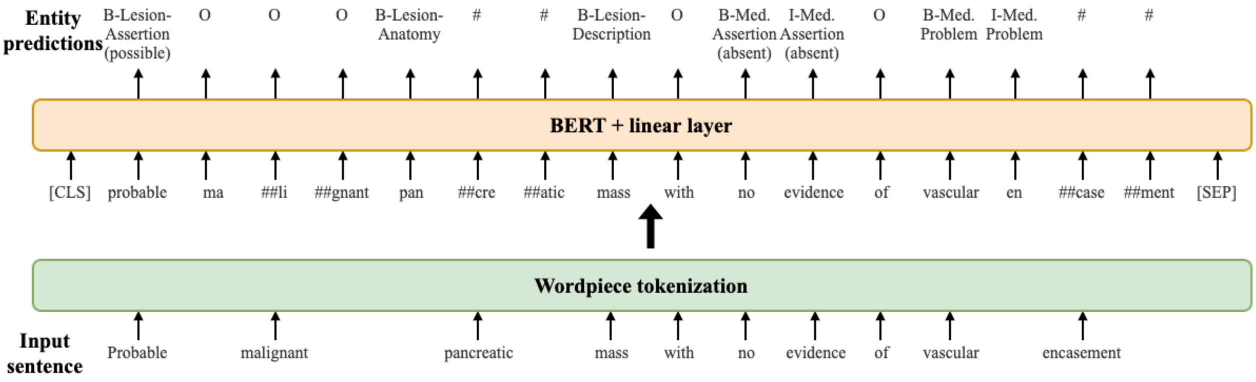


Figure 5.6: Architecture of the BERT NER model.

Argument Role Extraction

Once the triggers and argument entities were extracted, the argument roles were identified by predicting the relations between triggers and argument entities. All events comprised a single trigger that anchored the event, with zero or more argument connections. Each relation was unidirectional where the head was the trigger and the tail was an argument.

Relations were predicted using BERT by adding a linear layer to the pooled output state (encoded in the [CLS] token) and fine-tuning the model. Figure 5.7 presents the BERT RE model with an example input sentence. A unique input sentence was created for each candidate trigger-argument relation. The trigger and argument locations were marked with two pairs of special tokens, namely ([unused0], [unused1]) and ([unused2], [unused3]), which provided positional information about the entities and the direction of the relation. These special tokens were part of the BERT vocabulary designed for introducing new domain specific samples for fine tuning purposes. Consider the aforementioned example where the word "Probable" is the Lesion-Assertion of the Lesion trigger "mass". The trigger would be marked as "[unused0] mass [unused1]" and the Lesion-Assertion would be marked as "[unused2] probable [unused3]". Furthermore, we introduced a new relation called "No_relation" for negative training instances indicating the absence of relations between some arguments and triggers.

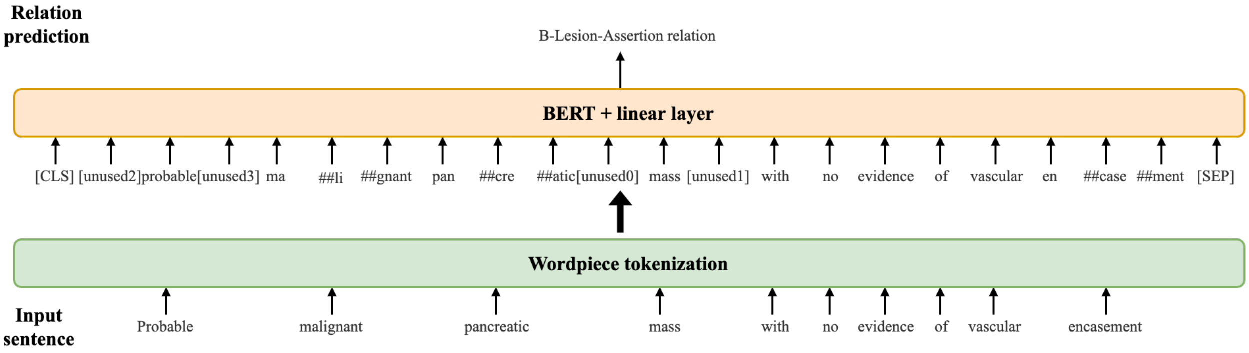

Relation prediction
B-Lesion-Assertion relation
BERT + linear layer
[CLS] [unused2]probable[unused3] ma ##li ##gnant pan ##cre ##atic[unused0] mass [unused1] with no evidence of vascular en ##case ##ment [SEP]
Wordpiece tokenization
Input sentence
Probable malignant pancreatic mass with no evidence of vascular encasement

Figure 5.7: Architecture of the BERT RE model.

We fine-tuned a single BERT model for both of the NER and RE tasks. While the input sequence encoding is task specific, the Wordpiece tokenization as well as the BERT model was shared. During training, the NER and RE batches were randomly alternated and each minimizing the cross-entropy loss of its target labels. The gradient of the loss is then applied to the same BERT model, effectively allowing the model to learn from both tasks.

Algorithm 1 Training NER and RE one a single BERT model

```
Preprocess the training samples for each task, i.e. NER, RE
Group the training samples into batches
Combine all sample batches
for each epoch do
    Shuffle the training samples
    for each batch in the samples do
        1. Train the batch samples
        2. Calculate the task-specific cross-entropy loss for the batch
        3. Calculate gradient
        4. Update model
    end for
end for
```

5.3.7 Experiment settings

All experiments were performed by 5-fold cross validation (CV) with the same data splits ratio (80% for training, 10% for validation, 10% for testing). For the entity extraction baseline (BiLSTM-CRF$_{rad}$), we used the word2vec embedding pre-trained on a radiology report dataset from our previous work [90]. This dataset contained over 3 million reports covering a wide range of imaging modalities, and were collected from two institutions including University of Washington Medical Center and Harborview Medical Center. In terms of the model hyperparameters, the embedding dimension and the hidden state dimension of the character and sequence LSTM layers were 25 and 100. We used the Adam Optimizer with a learning rate of 0.005, as recommended by NeuroNER. We applied early stopping with the validation set in order to avoid overfitting the training data [106]. The training was stopped when the validation results no longer showed improvement.

Three different pre-trained BERT models (BERT$_{base}$, BERT$_{clinical}$ and BERT$_{rad}$) were used experimentally. BERT$_{base}$ was pre-trained on Wikipedia and BookCorpus, and made available by Google [21]. BERT$_{clinical}$ was pre-trained on 2 million clinical documentation, including over 500,000 radiology reports, from the MIMIC-III database [45, 43]. BERT$_{rad}$ was pre-trained on over 3 million UW radiology reports and was initialized from the BERT$_{clinical}$. We pre-trained BERT$_{rad}$ for 150,000 steps with a batch size of 32, maximum sequence length of 128, and a learning rate of 2e-5. In our experiments, both entities and relations were extracted by fine-tuning the same BERT model. We used the same set of hyperparameters in all the extraction experiments, using the recommended values for fine-tuning suggested by Devlin et al., with a learning rate of 3e-5, a drop-out rate of 0.1. Early stopping was also employed using the validation set.

To better assess the general performance of the models with different subsamples, we repeated the cross validation 10 times. For each run, the cross validation data splits were created with a different random seed [106]. We reported the average precision, recall and F1 scores across these 50 different runs and included the 95% confidence intervals.

5.4 *Results*

5.4.1 Trigger and argument entity extraction results

All of the trigger and argument entities were extracted first before their relations were identified. Trigger and argument entity extraction performance was evaluated at the token-level. Table 5.4 presents the results.

Entity	**BiLSTM-CRF$_{rad}$**			**BERT$_{base}$**			**BERT$_{clinical}$**			**BERT$_{rad}$**		
	P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Medical-Problem	88.8	84.9	86.7 (±0.45)	89.1	83.9	86.4 (±0.37)	90.5	83.6	86.8 (±0.37)	91.3	85.0	**88.0** (±0.34)
Medical-Anatomy	79.1	79.9	79.3 (±0.92)	82.3	77.9	79.9 (±0.87)	83.8	77.3	80.3 (±0.84)	85.7	78.5	**81.8** (±0.75)
Medical-Assertion	85.6	84.5	84.9 (±0.79)	86.9	85.7	86.3 (±0.70)	87.8	84.7	86.1 (±0.63)	88.5	86.3	**87.3** (±0.78)
Lesion-Description	87.2	87.9	87.5 (±0.71)	89.1	86.8	87.9 (±0.66)	89.0	87.6	88.2 (±0.62)	90.0	88.4	**89.1** (±0.63)
Lesion-Anatomy	80.2	78.6	79.0 (±0.92)	85.5	76.5	80.6 (±0.94)	85.8	76.8	80.8 (±0.89)	86.8	80.7	**83.5** (±0.86)
Lesion-Assertion	81.3	72.1	76.2 (±1.55)	86.0	70.0	76.8 (±1.60)	85.6	70.5	77.1 (±1.48)	86.5	73.6	**79.3** (±1.26)
Lesion-Characteristic	76.6	72.6	74.1 (±1.36)	81.8	70.5	75.4 (±1.22)	82.8	71.3	76.3 (±1.11)	84.2	73.6	**78.3** (±1.14)
Lesion-Size	84.1	85.8	84.4 (±1.88)	91.1	84.2	87.3 (±1.37)	89.1	84.4	86.4 (±1.56)	90.7	88.2	**89.3** (±1.43)
Lesion-Count	89.1	85.6	86.7 (±2.20)	90.9	86.6	88.0 (±2.15)	92.0	88.0	**89.3** (±2.07)	91.0	87.5	88.7 (±2.16)
Lesion-Size-Trend	69.0	63.2	65.5 (±3.20)	78.0	60.7	67.6 (±3.14)	75.2	59.5	65.5 (±2.98)	77.3	63.6	**68.9** (±3.06)
Overall	84.2	82.1	83.1 (±0.37)	86.7	80.9	83.7 (±0.36)	87.7	80.6	84.0 (±0.28)	88.8	82.4	**85.5** (±0.28)

Table 5.4: Entity extraction results (average precision, recall and F1 in %), based on 10 runs of 5-fold cross validation. The numbers in brackets are 95% confidence intervals of the averages. The best F1 values are in bold.

The BERT models outperformed the BiLSTM-CRF$_{rad}$ baseline. With in-domain pretraining, BERT$_{rad}$ achieved higher overall performance other other BERT variants. In Lesion-Count prediction, BERT$_{clinical}$ was slightly higher than BERT$_{rad}$. In Lesion-Size-Trend prediction, the decreasing label had relatively low extraction performance due to the small sample size. For the Assertion extraction, the absent label was easier to predict since most of the annotated text spans comprised a single word “no”, which constituted 70% of the Medical-Assertion and 84% of the Lesion-Assertion. We conducted statistical significance tests using the overall F1 to access whether the difference in model results were due to randomness or sampling variability. In cross validation, the training sets overlap between different folds. As a result, the classification performance from each fold is not completely independent, and can lead to misleading statistical results when applying standard paired t-tests [107]. Hence, we applied the corrected resampled t-test, as suggested by Nadeau and Bengio [108], to better estimate the sample variance. The test results in Table 5.5 show that the overall performance of BERT$_{rad}$ was better than the other architectures with significance (p-value $<$ 5e-6).

	BERT$_{base}$	**BERT**$_{clinical}$	**BERT**$_{rad}$
BiLSTM-CRF$_{rad}$	0.000665	0.00143	**5.70E-08**
BERT$_{base}$	-	0.001506	**0.000028**
BERT$_{clinical}$	-	-	**0.000005**

Table 5.5: Statistical test results on trigger and argument entity extraction.

5.4.2 Argument role extraction results

In this section, we present the argument role extraction results, shown in Table 5.6. Specifically, we predicted the argument roles using the extracted triggers and argument entities rather than the gold standard entities. The evaluation of the event arguments was based on the scoring described in section 5.3.3, which considered the most salient information, i.e. slot value of each argument type in the finding events. The in-domain contextualized representations helped the BERT$_{rad}$ model achieved higher performance in general, with the exception of Lesion-Count.

Argument type	**Argument role**	**BERT**$_{base}$			**BERT**$_{clinical}$			**BERT**$_{rad}$		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Span-only	Medical-Anatomy	78.4	67.1	72.1 (±1.12)	80.0	66.6	72.5 (±1.02)	81.4	68.3	**74.2** (±1.00)
Span-with-value	Medical-Assertion	86.8	82.3	84.5 (±0.54)	87.5	81.7	84.4 (±0.43)	88.6	83.0	**85.6** (±0.45)
Span-only	Lesion-Anatomy	83.6	67.7	74.7 (±1.15)	84.2	68.1	75.1 (±0.98)	84.7	71.3	**77.3** (±1.06)
	Lesion-Characteristic	80.4	65.2	71.6 (±1.32)	81.5	66.0	72.6 (±1.21)	82.6	67.9	**74.2** (±1.28)
	Lesion-Count	87.0	81.6	83.4 (±2.11)	89.8	83.6	**86.0** (±2.18)	88.1	83.3	85.1 (±2.09)
	Lesion-Size	85.1	59.9	69.9 (±2.56)	85.5	60.6	70.5 (±2.10)	86.4	62.5	**72.0** (±2.25)
Span-with-value	Lesion-Assertion	85.4	79.7	82.4 (±0.69)	84.9	80.0	82.3 (±0.76)	86.1	81.2	**83.5** (±0.61)
	Lesion-Size-Trend	82.1	71.4	76.0 (±1.94)	80.3	70.4	74.4 (±2.21)	81.9	74.1	**77.4** (±2.28)

Table 5.6: End-to-end argument roles extraction results (average precision, recall and F1 in %), based on 10 runs of 5-fold cross validation. The numbers in brackets are 95% confidence intervals of the averages. The best F1 values are in bold.

5.4.3 Overall trigger and argument role extraction results

Table 5.7 presents the overall extraction performance for the triggers and the arguments (entities with roles). The BERT$_{rad}$ model achieved the highest average F1 of 92.9% for triggers, 75.0% for span-only arguments and 84.8% for span-with-value arguments. The performance of BERT$_{base}$ was comparable to BERT$_{clinical}$. While BERT$_{clinical}$ performed slightly better than BERT$_{base}$ on triggers and span-only arguments, BERT$_{base}$ performed slightly better on span-with-value arguments.

Argument	**BERT**$_{base}$			**BERT**$_{clinical}$			**BERT**$_{rad}$		
type	**P**	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Trigger	90.9	92.1	91.5 (±0.24)	91.5	92.2	91.8 (±0.26)	92.6	93.2	**92.9** (±0.25)
Span-only	79.8	67.1	72.8 (±0.71)	81.1	67.0	73.3 (±0.66)	82.3	69.0	**75.0** (±0.66)
Span-with-value	86.3	81.2	83.6 (±0.46)	86.3	76.3	83.5 (±0.41)	87.6	82.1	**84.8** (±0.39)

Table 5.7: Overall extraction performance for triggers and arguments (average Precision, Recall and F1 in %).

Similar statistical tests were conducted on the overall extraction results. As can be observed in Table 5.8, BERT$_{rad}$ achieved the best overall performance with significance (p-values $<$ 1.6e-4).

	Trigger		Span-only		Span-with-value	
	BERT$_{clinical}$	**BERT**$_{rad}$	**BERT**$_{clinical}$	**BERT**$_{rad}$	**BERT**$_{clinical}$	**BERT**$_{rad}$
BERT$_{base}$	0.002617	**0.000001**	0.005315	**0.001638**	0.001639	**0.000459**
BERT$_{clinical}$		**0.000037**		**0.000924**		**0.000224**

Table 5.8: Statistical test results on trigger and argument role extraction.

5.5 Extracting findings from MIMIC-CXR radiology reports

Section 5.4.2 shows the end-to-end event extraction results of our repeated cross-validation using the gold standard CT radiology reports. To explore how the extraction model generalizes on radiology reports from another institution as well as from another imagine modality, we created a validation dataset from the MIMIC-CXR chest X-ray reports [1]. 50 reports were randomly selected from the 227,835 reports in the database and annotated using the same finding event schema. This validation set included 257 Medical Problem finding events (141 argument entities and 313 roles) and 7 Lesion finding events (9 argument entities, 15 roles). We extracted the findings from this validation dataset and evaluated the extraction performance using the same argument role scoring described in section 5.3.3. The overall F1 scores on this validation set were 95.6% for triggers, 79.1% for span-only arguments and 89.7% for span-with-value argument. The performance was comparable to our repeated 5-fold cross validation performance, despite the fact that the MIMC-CXR reports were from a different institution and based on a different imaging modality.

Interestingly, the extracted MIMIC-CXR findings contained clinical concepts that were absent in the UW CT corpus. For instance, the words "plasmacytoma" and "fibroadenomas" were correctly identified as lesions and "acute respiratory distress syndrome" was correctly identified as medical problem, even though these lesion and medical problem mentions did not appear in any radiology reports in the training corpus. This could be attributed to the pre-training of BERT$_{rad}$ with 3 million UW radiology reports covering a wide range of modalities. To contribute to the core aim of the MIMIC-CXR project and facilitate future research studies in medical imaging, we extracted all clinical findings from 227,835 radiology reports in MIMIC-CXR using the fine-tuned BERT$_{rad}$ model. A total of 1,420,604 medical problem findings and 31,706 lesion findings were extracted. We are releasing the finding extraction results to the research community [2].

[1] https://physionet.org/content/mimic-cxr/2.0.0/

[2] https://github.com/uw-bionlp/MIMIC-CXR$_{c}linical_{f}indings$

5.6 Discussion and error analysis

In section 5.4.1, we show the superior performance of BERT models in entity extraction compared to the BiLSTM-CRF baseline. Furthermore, the in-domain pre-training allowed BERT$_{rad}$ to develop better contextual representations and generally achieved higher performance in both entities and argument roles extraction. Knowledge of clinical concepts that are unseen in the training corpus can be learned and transferred to other tasks.

In this section, we analysed the extraction results and discussed some limitations. Among the finding entities, Medical-Problem and Medical-Anatomy had relatively long text spans. Over 25% of Medical-Problem spans and 35% of Medical-Anatomy spans contained at least 5 words. We found that some entities with lengthy spans were extracted into multiple separate entities, particularly before and after a conjunctive word. About 4% of all Medical-Problem entities and 7% of all Medical-Anatomy entities were split into multiple entities by the entity extraction models. Figure 5.8 presents an example of each case.

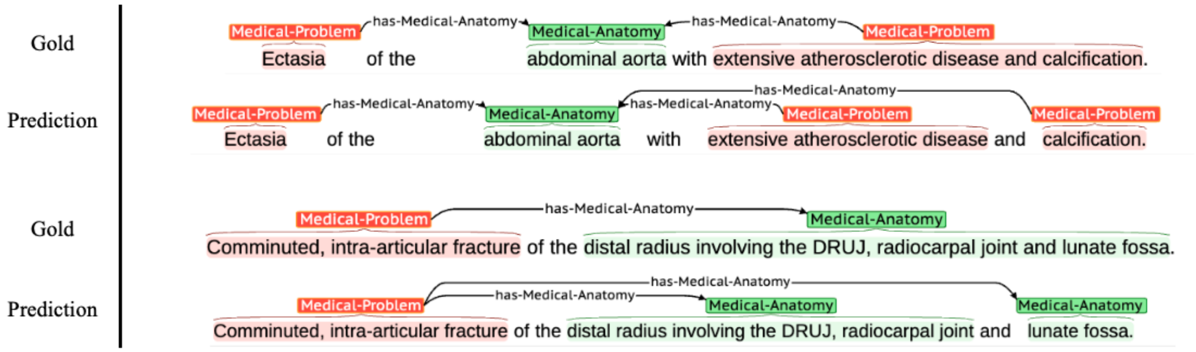

Gold
Medical-Problem
has-Medical-Anatomy
Medical-Anatomy
has-Medical-Anatomy
Medical-Problem
Ectasia of the abdominal aorta with extensive atherosclerotic disease and calcification.
Prediction
has-Medical-Anatomy
Medical-Problem
has-Medical-Anatomy
Medical-Anatomy
has-Medical-Anatomy
Medical-Problem
Medical-Problem
Ectasia of the abdominal aorta with extensive atherosclerotic disease and calcification.
Gold
Medical-Problem
has-Medical-Anatomy
Medical-Anatomy
Comminuted, intra-articular fracture of the distal radius involving the DRUJ, radiocarpal joint and lunate fossa.
Prediction
has-Medical-Anatomy
Medical-Problem
has-Medical-Anatomy
Medical-Anatomy
Medical-Anatomy
Comminuted, intra-articular fracture of the distal radius involving the DRUJ, radiocarpal joint and lunate fossa.

Figure 5.8: Examples of long text spans being extracted into multiple entities.

Our annotation schema allowed a text span assigned with multiple labels. For example, a body location can be both the anatomical region of a lesion and a medical problem. However, our NER model could only predict a single label for each token, and therefore cannot assign the same text spans for both Medical-Anatomy and Lesion-Anatomy. Approximately 1% of all entities in the corpus had multiple labels, so this limitation does not impact the overall extraction performance. One way to circumvent this single-label limitation is by having a single entity for both findings. Although a single anatomy entity no longer carries any clinical finding connotation, its association with the finding events can still be identified by the RE model.

Our extraction framework employed multi-task learning to optimize the parameters of a single BERT model. Other fine-tuning approaches applied additional components to the architecture to boost performance. One example is using graph structures to jointly model the span relations in the different tasks [109]. Another example is using entity aware markers to encode input sentences in a relation extraction model, which was shown to outperform joint modeling architectures [110]. Our BERT$_{rad}$ model was pre-trained using the common transfer learning paradigm by initializing its weight from another BERT model in relevant domain. This approach is particularly advantageous when the target data are scarce. However, a recent study showed that pre-training the language model from scratch in a domain with abundant unlabeled text could derive better in-domain vocabulary and result in substantial performance improvement [111]. Since our UW dataset contained more than 3 million radiology reports, this pre-training approach could potentially improve the contextual representation of the BERT$_{rad}$ model and possibly lead to better event extraction performance.

5.7 Summary

In this work, we present a new schema for extracting lesion and medical problem findings from radiology reports. Based on the schema, we annotated a new corpus of 500 CT reports. We used the corpus to train the same BiLSTM-CRF model in the first study [90] (Chapter 3) as the baseline for entity extraction. We then employed the same pre-trained language model, BERT, in the second study [112] (Chapter 4) to extract both entities and relations. We demonstrated that the BERT model not only outperformed the baseline in finding argument entities extraction, but also achieved superior performance in relation extraction. In particular, the one that was pre-trained with 3 million radiology reports achieved the highest performance in both entity and relation extraction. Because our multi-institutional corpus covered a wide range of imaging modalities, the acquired deep contextual knowledge from these reports allowed the model to perform comparatively well in another imaging modality. We demonstrated that by extracting clinical findings from the MIMIC-CXR chest X-ray reports.

Chapter 6

CONCLUSIONS AND FUTURE WORK

This thesis details how unstructured text in radiology reports can be applied to secondary use. In the first study, we extracted follow-up recommendations and the related entities from 3 million UW radiology reports. We analysed patient follow-up adherence rate by imagine modalities and discovered that mammogram exams had the highest follow-up rate of 77%. However, over 38% of patients with CT exams did not have any follow-up tests as recommended by radiologists. This finding may be concerning as missing follow-up diagnosis can compromise patient health, especially if an unexpected lesion was discovered in the image. In the second study, we developed a deep learning model using BERT to classify CT exam protocols. To train our models, we used structured EHR data (age, gender) as well as unstructured text of patient history and diagnosis. The most common exams, specifically the CT Chest, Abdomen, Pelvis with IV and Oral, could be classified at 93% F1. Some exams could be classified at up to 96% F1 (CT Pelvis Cystogram). Potentially, the machine learning model can be integrated into clinical workflow to make suggestions for radiologists. As the advancement of NLP and AI research continue to progress, the model performance is approaching human performance. In the third study, we extracted clinical findings using an in-domain pre-trained BERT model, the extraction performance for triggers and arguments were (92.9%,75.0%,84.8%), while the human annotator performance were (93.0%, 83.7%, 86.9%). Although our span-only argument extraction performance was below annotator agreement, both triggers and span-with-value arguments extraction were close to human performance. Furthermore, the BERT$_{rad}$ model trained with CT reports performed equally well in the extraction of chest X-ray reports.

6.1 Contribution

Collectively, the information extracted in our studies provided vast opportunities for secondary use application. The clinical findings of lesions and medical problems offered additional evidence to support radiologists' follow-up recommendations. For those patients who did not have follow-up imaging tests, we can investigate the possible correlation in the presence of lesions/medical problems, and certain anatomical regions. Are patients with certain lesion/medical problem findings at specific anatomical locations more likely to miss recommended follow-up tests? How many of these patients had an unexpected lesion finding?

Although our extraction work focused on radiology reports, the deep learning NER and RE approaches can also be applied to other clinical text. For that reason, the multi-task event extraction framework used in the third study (Chapter 5) was developed to be able to train/validate/predict any annotated corpora. The file path of the annotations, as well as model parameters can be set easily in a configuration file. Any extraction experiments can therefore be performed with minimal effort, and the results are reproducible.

6.2 Future work

One of the limitations in supervised machine learning is that the model performance hinges on the quality of the annotation. Achieving high quality of annotation requires substantial time and effort from clinicians especially if the annotation schemas are complex. Unlike other clinical domains, one unique aspect of radiology is the abundance of imaging data. The raw pixels, even without annotations, depict the body locations, effusion, and lesions. Some examinations also include multiple views. Theses images offer additional training data to each radiology report, and can be trained in a multi-modal network to boost the performance of the extraction work.

In recent years, multi-modal learning has been applied to natural language generation (NLG). Using pre-trained image encoders and RNN decoders, researchers were able to generate captions describing images [113, 114]. These models were later adapted to the radiology domain, where an image encoder was trained to read the X-ray image with a RNN or transformer decoder to generate the report of clinical findings. Table 6.1 presents an example of predicted reports using the *Show, Attend, and Tell* model [114]. The gold standard images and reports were taken from the MIMIC-CXR database.

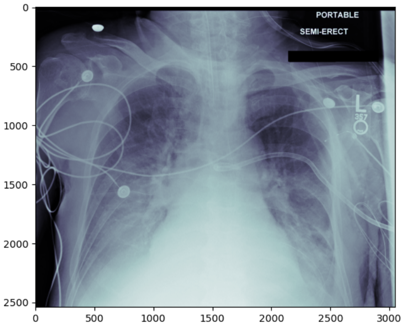
	Gold	Generated report
 PORTABLE SEMI-ERECT 0 500 1000 1500 2000 2500 0 500 1000 1500 2000 2500 3000	As compared to the previous radiograph, there is no relevant change. **No change in position of the left pleural pigtail catheter**. No evidence of recurrent left pneumothorax. Unchanged opacities at the left lung base and in the retrocardiac lung regions. Unchanged appearance of the right lung with **mild cardiomegaly** and a right **basal combined atelectasis** and **pleural effusion**. **Unchanged position of the endotracheal tube and the nasogastric tube**.	In comparison with the study of _ _ _ , **the monitoring and support devices remain in place** . There is again **enlargement of the cardiac silhouette** with pulmonary vascular congestion and bilateral **pleural effusions** with compressive **atelectasis at the bases** . **Monitoring and support devices remain in place** .

Table 6.1: Generated report of clinical findings using MIMIC-CXR image.

Radiology report generation is an emerging research area. One potential approach to improve the report generation is by leveraging our extracted MIMIC-CXR clinical findings as a secondary training objective to fine tune the sequence decoder. The decoded sequence can be used to predict the one-hot representation of the extracted clinical findings, using an additional cross-entropy loss function. This secondary training objective can reinforce the NLG model's understanding of common clinical findings in chest X-ray images, such as pleural effusion, pneumonia, and pneumothorax, thereby allowing the model to generate more accurate reports.

6.3 Final remarks

AI technology is already part of our everyday lives, from online shopping recommendations, digital voice assistant, to autonomous driving. The opportunity of integrating AI into hospital EHR systems is immense. The vast amount of clinical data offer tremendous opportunities to derive new and important insights, which can assist health care providers and improve patient care. Recognizing the importance of clinical AI, FDA proposed a framework, namely "Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning-Based Software as a Medical Device" [115], to ensure changes made to the AI algorithms are transparent so that real-world performance can be monitored for safety assurance. In addition, FDA organized a public workshop in February 2020, with the title of "Evolving Role of Artificial Intelligence in Radiological Imaging.", to discuss applications of AI in radiological imaging, focusing on computer aided-detection and diagnosis software (CADe and CADx). CADe and CADx software make suggestion on clinical relevant findings after analyzing radiological images. Unlike the software from early generation which only augmented the tasks performed by radiologists, FDA acknowledged that the latest software powered by advanced AI can perform some tasks autonomously. This perspective is concurred by Dr. Geoff Hinton, a renowned professor and researcher in deep learning, who popularized backpropagation in neural networks, and consequently contributed to the recent deep learning movement. Hinton asserted that within the next 5-10 years, deep learning can do better than radiologists in interpreting radiological images [116].

Whether Hinton's prediction is true or not still yet to be seen. However, undoubtedly, deep learning allows us to extract valuable information from radiology reports with state-of-the-art performance. The extracted information provides supporting evidence for clinicians to determine their course of action. At the minimum, they can serve as reminders to prevent clinicians from overlooking clinically important findings and recommendations. We are optimistic that deep learning and AI will continue to mature and, with FDA's governance and careful evaluation in prospective clinical trials, can eventually be integrated into clinical decision support systems to improve quality of patient care. We hope that this dissertation work contributes to this noble cause and brings a step closer to achieving more "meaningful use" of EHRs.

BIBLIOGRAPHY

[1] The american recovery and reinvestment act of 2009. https://www.healthit.gov/sites/default/files/hitech_act_excerpt_from_arra_with_index.pdf.

[2] Improve quality safety, efficiency and reducing health disparities. https://www.healthit.gov/sites/default/files/mu_wg_stage_3_planning_16_jul_12.docx.

[3] Ellen Kim, Samuel M Rubinstein, Kevin T Nead, Andrzej P Wojcieszynski, Peter E Gabriel, and Jeremy L Warner. The evolving use of electronic health records (ehr) for research. In *Seminars in radiation oncology*, volume 29, pages 354–361. Elsevier, 2019.

[4] Charles Safran, Meryl Bloomrosen, W Edward Hammond, Steven Labkoff, Suzanne Markel-Fox, Paul C Tang, and Don E Detmer. Toward a national framework for the secondary use of health data: an american medical informatics association white paper. *Journal of the American Medical Informatics Association*, 14(1):1–9, 2007.

[5] The acr practice parameter for the communicationof diagnostic imaging findings. https://www.acr.org/-/media/acr/files/practice-parameters/communicationdiag.pdf.

[6] The joint commission. improve the effectiveness of communication among caregivers. https://www.jointcommission.org/-/media/tjc/documents/standards/national-patient-safety-goals/2020/npsg_chapter_cah_jul2020.pdf.

[7] Joanne Callen, Andrew Georgiou, Julie Li, and Johanna I Westbrook. The safety implications of missed test results for hospitalised patients: a systematic review. *BMJ quality & safety*, 20(2):194–199, 2011.

[8] William E Holden, David M Lewinsohn, Molly L Osborne, Chris Griffin, Ann Spencer, Carol Duncan, and Mark E Deffebach. Use of a clinical pathway to manage unsuspected radiographic findings. *Chest*, 125(5):1753–1760, 2004.

[9] Eric G Poon, Jennifer S Haas, Ann Louise Puopolo, Tejal K Gandhi, Elisabeth Burdick, David W Bates, and Troyen A Brennan. Communication factors in the follow-up of abnormal mammograms. *Journal of General Internal Medicine*, 19(4):316–323, 2004.

[10] R James Brenner, Leonard L Lucey, John J Smith, and Roger Saunders. Radiology and medical malpractice claims: a report on the practice standards claims survey of the physician insurers association of america and the american college of radiology. *AJR. American journal of roentgenology*, 171(1):19–22, 1998.

[11] Jeremy S Whang, Stephen R Baker, Ronak Patel, Lyndon Luk, and Alejandro Castro III. The causes of medical malpractice suits against radiologists in the united states. *Radiology*, 266(2):548–554, 2013.

[12] Meliha Yetisgen-Yildiz, Martin L Gunn, Fei Xia, and Thomas H Payne. A text processing pipeline to extract recommendations from radiology reports. *Journal of biomedical informatics*, 46(2):354–362, 2013.

[13] Meliha Yetisgen-Yildiz, Martin L Gunn, Fei Xia, and Thomas H Payne. Automatic identification of critical follow-up recommendation sentences in radiology reports. In *AMIA Annual Symposium Proceedings*, volume 2011, page 1593. American Medical Informatics Association, 2011.

[14] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1480–1489, 2016.

[15] Jay A Harolds, Jay R Parikh, Edward I Bluth, Sharon C Dutton, and Michael P Recht. Burnout of radiologists: frequency, risk factors, and remedies: a report of the acr commission on human resources. *Journal of the American College of Radiology*, 13(4):411–416, 2016.

[16] Brad J Balint, Scott D Steenburg, Hongbu Lin, Changyu Shen, Jennifer L Steele, and Richard B Gunderman. Do telephone call interruptions have an impact on radiology resident diagnostic accuracy? *Academic radiology*, 21(12):1623–1628, 2014.

[17] McKinley Glover IV, Renata R Almeida, Pamela W Schaefer, Michael H Lev, and William A Mehan Jr. Quantifying the impact of noninterpretive tasks on radiology report turn-around times. *Journal of the American College of Radiology*, 14(11):1498–1503, 2017.

[18] Andrew Schemmel, Matthew Lee, Taylor Hanley, B Dustin Pooler, Tabassum Kennedy, Aaron Field, Douglas Wiegmann, and J Yu John-Paul. Radiology workflow disruptors: a detailed analysis. *Journal of the American College of Radiology*, 13(10):1210–1214, 2016.

[19] Giles W Boland, Richard Duszak, and Mannudeep Kalra. Protocol design and optimization. *J Am Coll Radiol*, 11(5):440–1, 2014.

[20] Computed tomography (ct) - body. `https://www.radiologyinfo.org/en/info/bodyct`.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

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

[23] Peter Spyns. Natural language processing in medicine: an overview. *Methods of information in medicine*, 35(04/05):285–301, 1996.

[24] Naomi Sager. A procedure for left to right analysis of sentence structure. `https://cs.nyu.edu/cs/projects/lsp/pubs/TDAP_27_1960.pdf`.

[25] Naomi Sager. *Report on the String Analysis Programs: Introductory Volume.* University of Pennsylvania, Department of linguistics, 1966.

[26] Naomi Sager, Carol Friedman, and Margaret S Lyman. *Medical language processing: computer management of narrative data.* Addison-Wesley Longman Publishing Co., Inc., 1987.

[27] M Lyman, N Sager, C Friedman, and E Chi. Computer-structured narrative in ambulatory care: its use in longitudinal review of clinical data. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 82. American Medical Informatics Association, 1985.

[28] Donald AB Lindberg, Betsy L Humphreys, and Alexa T McCray. The unified medical language system. *Yearbook of Medical Informatics*, 2(01):41–51, 1993.

[29] Alan R Aronson. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association, 2001.

[30] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[31] Sergey Goryachev, Margarita Sordo, and Qing T Zeng. A suite of natural language processing tools developed for the i2b2 project. In *AMIA Annual Symposium Proceedings*, volume 2006, page 931. American Medical Informatics Association, 2006.

[32] Hamish Cunningham. Gate, a general architecture for text engineering. *Computers and the Humanities*, 36(2):223–254, 2002.

[33] Amber Stubbs and Özlem Uzuner. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/uthealth corpus. *Journal of biomedical informatics*, 58:S20–S29, 2015.

[34] Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[35] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. Annotating temporal information in clinical narratives. *Journal of biomedical informatics*, 46:S5–S12, 2013.

[36] Özlem Uzuner, Imre Solti, and Eithon Cadag. Extracting medication information from clinical text. *Journal of the American Medical Informatics Association*, 17(5):514–518, 2010.

[37] Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. Semeval-2017 task 12: Clinical tempeval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 565–572, Vancouver, Canada, 8 2017. Association for Computational Linguistics.

[38] Kirk Roberts, Dina Demner-Fushman, and Joseph M Tonning. Overview of the tac 2017 adverse reaction extraction from drug labels track. In *TAC*, 2017.

[39] Hanna Suominen, Sanna Salanterä, Sumithra Velupillai, Wendy W Chapman, Guergana Savova, Noemie Elhadad, Sameer Pradhan, Brett R South, Danielle L Mowery, Gareth JF Jones, et al. Overview of the share/clef ehealth evaluation lab 2013. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 212–231. Springer, 2013.

[40] Sameer Pradhan, Wendy Chapman, Suresh Man, and Guergana Savova. Semeval-2014 task 7: Analysis of clinical text. In *Proc. of the 8th International Workshop on Semantic Evaluation (SemEval 2014*, page 5462, 2014.

[41] Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. Semeval-2015 task 14: Analysis of clinical text. In *proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, 2015.

[42] Dina Demner-Fushman, K Bretonnel Cohen, Sophia Ananiadou, and Jun'ichi Tsujii. Proceedings of the 19th sigbiomed workshop on biomedical language processing. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, 2020.

[43] Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

[44] Hye Jin Kam and Ha Young Kim. Learning representations for the early detection of sepsis with deep neural networks. *Computers in biology and medicine*, 89:248–255, 2017.

[45] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[46] Geeticka Chauhan, Ruizhi Liao, William Wells, Jacob Andreas, Xin Wang, Seth Berkowitz, Steven Horng, Peter Szolovits, and Polina Golland. Joint modeling of chest radiographs and radiology reports for pulmonary edema assessment. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 529–539. Springer, 2020.

[47] David Zingmond and Leslie A Lenert. Monitoring free-text data using medical language processing. *Computers and Biomedical Research*, 26(5):467–481, 1993.

[48] Kim N Danforth, Megan I Early, Sharon Ngan, Anne E Kosco, Chengyi Zheng, and Michael K Gould. Automated identification of patients with pulmonary nodules in an integrated health system using administrative health plan data, radiology reports, and natural language processing. *Journal of Thoracic Oncology*, 7(8):1257–1262, 2012.

[49] Sheng Yu, Kanako K Kumamaru, Elizabeth George, Ruth M Dunne, Arash Bedayat, Matey Neykov, Andetta R Hunsaker, Karin E Dill, Tianxi Cai, and Frank J Rybicki. Classification of ct pulmonary angiography reports by presence, chronicity, and location of pulmonary embolism with natural language processing. *Journal of biomedical informatics*, 52:386–393, 2014.

[50] Carol Friedman, Stephen B Johnson, Bruce Forman, and Justin Starren. Architectural requirements for a multipurpose natural language processor in the clinical environment. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 347. American Medical Informatics Association, 1995.

[51] George Hripcsak, Carol Friedman, Philip O Alderson, William DuMouchel, Stephen B Johnson, and Paul D Clayton. Unlocking clinical data from narrative reports: a study of natural language processing. *Annals of internal medicine*, 122(9):681–688, 1995.

[52] Carol Friedman, James J Cimino, and Stephen B Johnson. A conceptual model for clinical radiology reports. In *Proceedings of the Annual Symposium on Computer Application in Medical Care*, page 829. American Medical Informatics Association, 1993.

[53] Carol Friedman, Philip O Alderson, John HM Austin, James J Cimino, and Stephen B Johnson. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association*, 1(2):161–174, 1994.

[54] Merlijn Sevenster, Rob Van Ommering, and Yuechen Qian. Automatically correlating clinical findings and body locations in radiology reports using medlee. *Journal of digital imaging*, 25(2):240–249, 2012.

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

[56] Radiological Society of North America. Radlex radiology lexicon. `https://www.rsna.org/practice-tools/data-tools-and-standards/radlex-radiology-lexicon`.

[57] Wen-wai Yim, Tyler Denman, Sharon W Kwan, and Meliha Yetisgen. Tumor information extraction in radiology reports for hepatocellular carcinoma patients. *AMIA Summits on Translational Science Proceedings*, 2016:455, 2016.

[58] Savelie Cornegruta, Robert Bakewell, Samuel Withey, and Giovanni Montana. Modelling radiological language with bidirectional long short-term memory networks. *arXiv preprint arXiv:1609.08409*, 2016.

[59] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[60] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

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

[62] Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International journal of medical informatics*, 132:103985, 2019.

[63] Eric G Poon, Tejal K Gandhi, Thomas D Sequist, Harvey J Murff, Andrew S Karson, and David W Bates. "i wish i had seen this test result earlier!": dissatisfaction with test result management systems in primary care. *Archives of internal medicine*, 164(20):2223–2228, 2004.

[64] Sayon Dutta, William J Long, David FM Brown, and Andrew T Reisner. Automated detection using natural language processing of radiologists recommendations for additional imaging of incidental findings. *Annals of emergency medicine*, 62(2):162–169, 2013.

[65] Pragya A Dang, Mannudeep K Kalra, Michael A Blake, Thomas J Schultz, Markus Stout, Paul R Lemay, David J Freshman, Elkan F Halpern, and Keith J Dreyer. Natural language processing using online analytic processing for assessing recommendations in radiology reports. *Journal of the American College of Radiology*, 5(3):197–204, 2008.

[66] Thusitha Mabotuwana, Christopher S Hall, Sandeep Dalal, Joel Tieder, and Martin L Gunn. Extracting follow-up recommendations and associated anatomy from radiology reports. In *MEDINFO 2017: Precision Healthcare through Informatics*, pages 1090–1094. IOS Press, 2017.

[67] Brian E Chapman, Danielle L Mowery, Evan Narasimhan, Neel Patel, Wendy Chapman, and Marta Heilbrun. Assessing the feasibility of an automated suggestion system for communicating critical findings from chest radiology reports to referring physicians. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 181–185, 2016.

[68] Eamon Johnson, W Christopher Baughman, and Gultekin Ozsoyoglu. Modeling incidental findings in radiology records. In *Proceedings of the International Conference on Bioinformatics, Computational Biology and Biomedical Informatics*, pages 940–945, 2013.

[69] Emmanuel Carrodeguas, Ronilda Lacson, Whitney Swanson, and Ramin Khorasani. Use of machine learning to identify follow-up recommendations in radiology reports. *Journal of the American College of Radiology*, 16(3):336–343, 2019.

[70] Shang Gao, Michael T Young, John X Qiu, Hong-Jun Yoon, James B Christian, Paul A Fearn, Georgia D Tourassi, and Arvind Ramanthan. Hierarchical attention networks for information extraction from cancer pathology reports. *Journal of the American Medical Informatics Association*, 25(3):321–330, 2018.

[71] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. Neuroner: an easy-to-use program for named-entity recognition based on neural networks. *arXiv preprint arXiv:1705.05487*, 2017.

[72] Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. In *Lrec*, volume 3735, page 3740, 2012.

[73] Mahadevappa Mahesh. Variability in ct protocols. *Journal of the American College of Radiology*, 10(10):805–806, 2013.

[74] Peter B Sachs, Geralyn Gassert, Michael Cain, David Rubinstein, Melody Davey, and Danielle Decoteau. Imaging study protocol selection in the electronic medical record. *Journal of the American College of Radiology*, 10(3):220–222, 2013.

[75] Kalpana M Kanal, Monica S Vavilala, Colin Raelson, Abhishek Mohan, Wendy Cohen, Jeffrey Jarvik, Frederick P Rivara, and Brent K Stewart. Variation in pediatric head ct imaging protocols in washington state. *Journal of the American College of Radiology*, 8(4):242–250, 2011.

[76] J Yu John-Paul, Akash P Kansagra, and John Mongan. The radiologist's workflow environment: evaluation of disruptors and potential implications. *Journal of the American College of Radiology*, 11(6):589–593, 2014.

[77] Andrew D Brown and Thomas R Marotta. Using machine learning for sequence-level automated mri protocol selection in neuroradiology. *Journal of the American Medical Informatics Association*, 25(5):568–571, 2018.

[78] Hari Trivedi, Joseph Mesterhazy, Benjamin Laguna, Thienkhai Vu, and Jae Ho Sohn. Automatic determination of the need for intravenous contrast in musculoskeletal mri examinations using ibm watson's natural language processing algorithm. *Journal of digital imaging*, 31(2):245–251, 2018.

[79] Ignat Drozdov, Daniel Forbes, Benjamin Szubert, Mark Hall, Chris Carlin, and David J Lowe. Supervised and unsupervised language modelling in chest x-ray radiological reports. *Plos one*, 15(3):e0229963, 2020.

[80] Qiang Wei, Zongcheng Ji, Yuqi Si, Jingcheng Du, Jingqi Wang, Firat Tiryaki, Stephen Wu, Cui Tao, Kirk Roberts, and Hua Xu. Relation extraction from clinical narratives using pre-trained language models. In *AMIA Annual Symposium Proceedings*, volume 2019, page 1236. American Medical Informatics Association, 2019.

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

[82] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[83] Minlong Lin, Ke Tang, and Xin Yao. Dynamic sampling approach to training neural networks for multiclass imbalance classification. *IEEE Transactions on Neural Networks and Learning Systems*, 24(4):647–660, 2013.

[84] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[85] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[86] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[87] Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *International Conference on Machine Learning*, pages 1607–1616. PMLR, 2018.

[88] Kevin Clark, Minh-Thang Luong, Urvashi Khandelwal, Christopher D Manning, and Quoc Le. Bam! born-again multi-task networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5931–5937, 2019.

[89] Raphael Tang, Yao Lu, Linqing Liu, Lili Mou, Olga Vechtomova, and Jimmy Lin. Distilling task-specific knowledge from bert into simple neural networks. *arXiv preprint arXiv:1903.12136*, 2019.

[90] Wilson Lau, Thomas H Payne, Ozlem Uzuner, and Meliha Yetisgen. Extraction and analysis of clinically important follow-up recommendations in a large radiology dataset. *AMIA Summits on Translational Science Proceedings*, 2020:335, 2020.

[91] Dina Demner-Fushman, Wendy W Chapman, and Clement J McDonald. What can natural language processing do for clinical decision support? *Journal of biomedical informatics*, 42(5):760–772, 2009.

[92] Janet P Haas, Eneida A Mendonça, Barbara Ross, Carol Friedman, and Elaine Larson. Use of computerized surveillance to detect nosocomial pneumonia in neonatal intensive care unit patients. *American journal of infection control*, 33(8):439–443, 2005.

[93] Basel Qenam, Tae Youn Kim, Mark J Carroll, and Michael Hogarth. Text simplification using consumer health vocabulary to generate patient-centered radiology reporting: translation and evaluation. *Journal of medical Internet research*, 19(12):e417, 2017.

[94] Bryan Rink, Kirk Roberts, Sanda Harabagiu, Richard H Scheuermann, Seth Toomay, Travis Browning, Teresa Bosler, and Ronald Peshock. Extracting actionable findings of appendicitis from radiology reports using natural language processing. *AMIA Summits on Translational Science Proceedings*, 2013:221, 2013.

[95] Jason J Zopf, Jessica M Langer, William W Boonn, Woojin Kim, and Hanna M Zafar. Development of automated detection of radiology reports citing adrenal findings. *Journal of digital imaging*, 25(1):43–49, 2012.

[96] Yanshan Wang, Saeed Mehrabi, Sunghwan Sohn, Elizabeth J Atkinson, Shreyasee Amin, and Hongfang Liu. Natural language processing of radiology reports for identification of skeletal site-specific fractures. *BMC medical informatics and decision making*, 19(3):23–29, 2019.

[97] Shumei Miao, Tingyu Xu, Yonghui Wu, Hui Xie, Jingqi Wang, Shenqi Jing, Yaoyun Zhang, Xiaoliang Zhang, Yinshuang Yang, Xin Zhang, et al. Extraction of bi-rads findings from breast ultrasound reports in chinese using deep learning approaches. *International journal of medical informatics*, 119:17–21, 2018.

[98] Guergana K Savova, Jin Fan, Zi Ye, Sean P Murphy, Jiaping Zheng, Christopher G Chute, and Iftikhar J Kullo. Discovering peripheral arterial disease cases from radiology notes using natural language processing. In *AMIA Annual Symposium Proceedings*, volume 2010, page 722. American Medical Informatics Association, 2010.

[99] Ricky K Taira, Stephen G Soderland, and Rex M Jakobovits. Automatic structuring of radiology free-text reports. *Radiographics*, 21(1):237–245, 2001.

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

[101] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 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*, pages 102–107, 2012.

[102] George Hripcsak and Adam S Rothschild. Agreement, the f-measure, and reliability in information retrieval. *Journal of the American medical informatics association*, 12(3):296–298, 2005.

[103] Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, 2016.

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

[105] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[106] Lutz Prechelt. Automatic early stopping using cross validation: quantifying the criteria. *Neural Networks*, 11(4):761–767, 1998.

[107] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[108] Claude Nadeau and Yoshua Bengio. Inference for the generalization error. *Machine learning*, 52(3):239–281, 2003.

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

[110] Zexuan Zhong and Danqi Chen. A frustratingly easy approach for entity and relation extraction. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 50–61, 2021.

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

[112] Wilson Lau, Laura Aaltonen, Martin Gunn, and Meliha Yetisgen. Automatic assignment of radiology examination protocols using pre-trained language models with knowledge distillation. *arXiv preprint arXiv:2009.00694*, 2020.

[113] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[114] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[115] Food, Drug Administration, et al. Proposed regulatory framework for modifications to artificial intelligence/machine learning (ai/ml)-based software as a medical device (samd). 2019.

[116] Geoff Hinton. Creative destruction lab. geoff hinton: on radiology. 2016. `https://www.youtube.com/watch?v=2HMPRXstSvQ`.

[117] Leonard Berlin. Duty to directly communicate radiologic abnormalities: has the pendulum swung too far? *American Journal of Roentgenology*, 181(2):375–381, 2003.

[118] Leonard M Berlin. Failure of radiologic communication: an increasing cause of malpractice litigation and harm to patients. *Applied Radiology*, 39(1):17, 2010.

[119] 2020 physician compensation report. `https://c8y.doxcdn.com/image/upload/v1/Press%20Blog/Research%20Reports/compensation-report-2020.pdf`.

[120] Yajuan Wang, Steven R Steinhubl, Chrisopher Defilippi, Kenney Ng, Shahram Ebadollahi, Walter F Stewart, and Roy J Byrd. Prescription extraction from clinical notes: towards automating emr medication reconciliation. *AMIA Summits on Translational Science Proceedings*, 2015:188, 2015.

[121] Naveen Ashish, Lisa Dahm, and Charles Boicey. University of california, irvine–pathology extraction pipeline: The pathology extraction pipeline for information extraction from pathology reports. *Health informatics journal*, 20(4):288–305, 2014.

[122] Qing T Zeng, Sergey Goryachev, Scott Weiss, Margarita Sordo, Shawn N Murphy, and Ross Lazarus. Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC medical informatics and decision making*, 6(1):1–9, 2006.

[123] Sam Henry, Yanshan Wang, Feichen Shen, and Ozlem Uzuner. The 2019 national natural language processing (nlp) clinical challenges (n2c2)/open health nlp (ohnlp) shared task on clinical concept normalization for clinical records. *Journal of the American Medical Informatics Association*, 27(10):1529–1537, 2020.

[124] Feichen Shen, Sijia Liu, Sunyang Fu, Yanshan Wang, Sam Henry, Ozlem Uzuner, and Hongfang Liu. Family history extraction from synthetic clinical narratives using natural language processing: overview and evaluation of a challenge data set and solutions for the 2019 national nlp clinical challenges (n2c2)/open health natural language processing (ohnlp) competition. *JMIR Medical Informatics*, 9(1):e24008, 2021.

[125] Yanshan Wang, Sunyang Fu, Feichen Shen, Sam Henry, Ozlem Uzuner, and Hongfang Liu. The 2019 n2c2/ohnlp track on clinical semantic textual similarity: overview. *JMIR Medical Informatics*, 8(11):e23375, 2020.

[126] Amber Stubbs, Michele Filannino, Ergin Soysal, Samuel Henry, and Özlem Uzuner. Cohort selection for clinical trials: n2c2 2018 shared task track 1. *Journal of the American Medical Informatics Association*, 26(11):1163–1171, 2019.

[127] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Ozlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.

[128] Keith J Dreyer, Mannudeep K Kalra, Michael M Maher, Autumn M Hurier, Benjamin A Asfaw, Thomas Schultz, Elkan F Halpern, and James H Thrall. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology*, 234(2):323–329, 2005.

[129] Thusitha Mabotuwana, Vadiraj Hombal, Sandeep Dalal, Christopher S Hall, and Martin Gunn. Determining adherence to follow-up imaging recommendations. *Journal of the American College of Radiology*, 15(3):422–428, 2018.

[130] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

[131] Selen Bozkurt, Jafi A Lipson, Utku Senol, and Daniel L Rubin. Automatic abstraction of imaging observations with their characteristics from mammography reports. *Journal of the American Medical Informatics Association*, 22(e1):e81–e92, 2015.

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

[133] Ferris M Hall. Language of the radiology report: primer for residents and wayward radiologists. *American Journal of Roentgenology*, 175(5):1239–1242, 2000.

[134] Katikapalli Subramanyam Kalyan and Sivanesan Sangeetha. Secnlp: A survey of embeddings in clinical natural language processing. *Journal of biomedical informatics*, 101:103323, 2020.

[135] Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*, 2019.

[136] Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng Wang. A dependency-based neural network for relation classification. *arXiv preprint arXiv:1507.04646*, 2015.

[137] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 1785–1794, 2015.

[138] Justin Lovelace and Bobak Mortazavi. Learning to generate clinically coherent chest x-ray reports. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 1235–1243, 2020.

Appendix A

ANNOTATION GUIDELINES FOR CLINICAL FINDING EXTRACTION

The goal of this annotation is to extract two different clinical findings from radiology reports, namely Medical problem finding, and Lesion finding. Each finding is represented by an event consisting of a trigger and multiple arguments. The annotation process involves identifying text spans within the notes that directly associate with the different clinical information as well as the relationships among them. Each piece of information (entity) is related to an event trigger, which link all information together cohesively. The following diagram shows the different clinical entities with the event triggers marked in red. The entities with * are categorical with pre-defined values. The following sections describe how each piece of information will be annotated. Not all entities are present in a radiology report. However, when annotating an entity within a clinical finding, the corresponding trigger (red) should always be identified and annotated first.

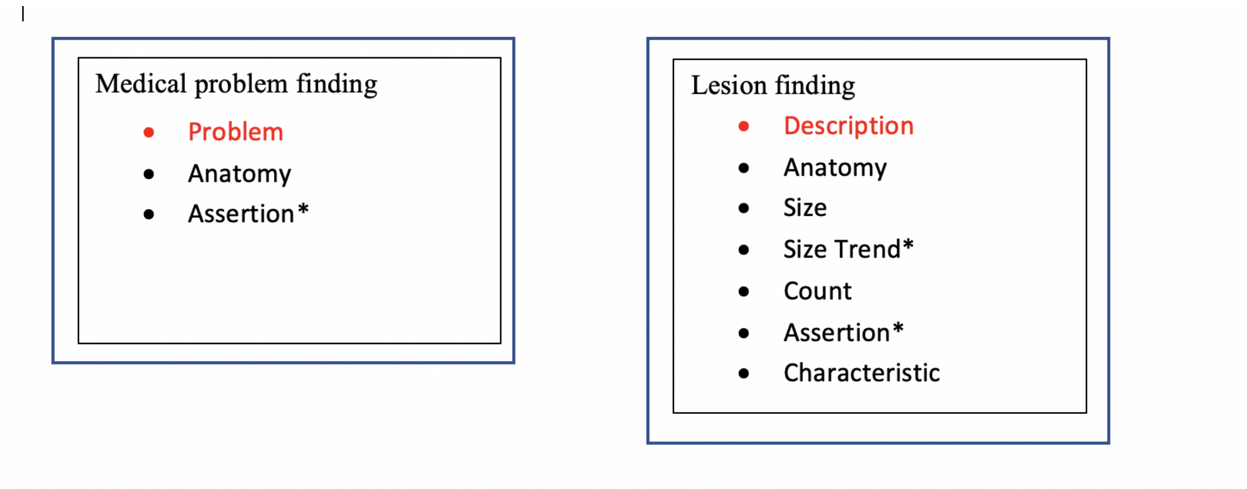
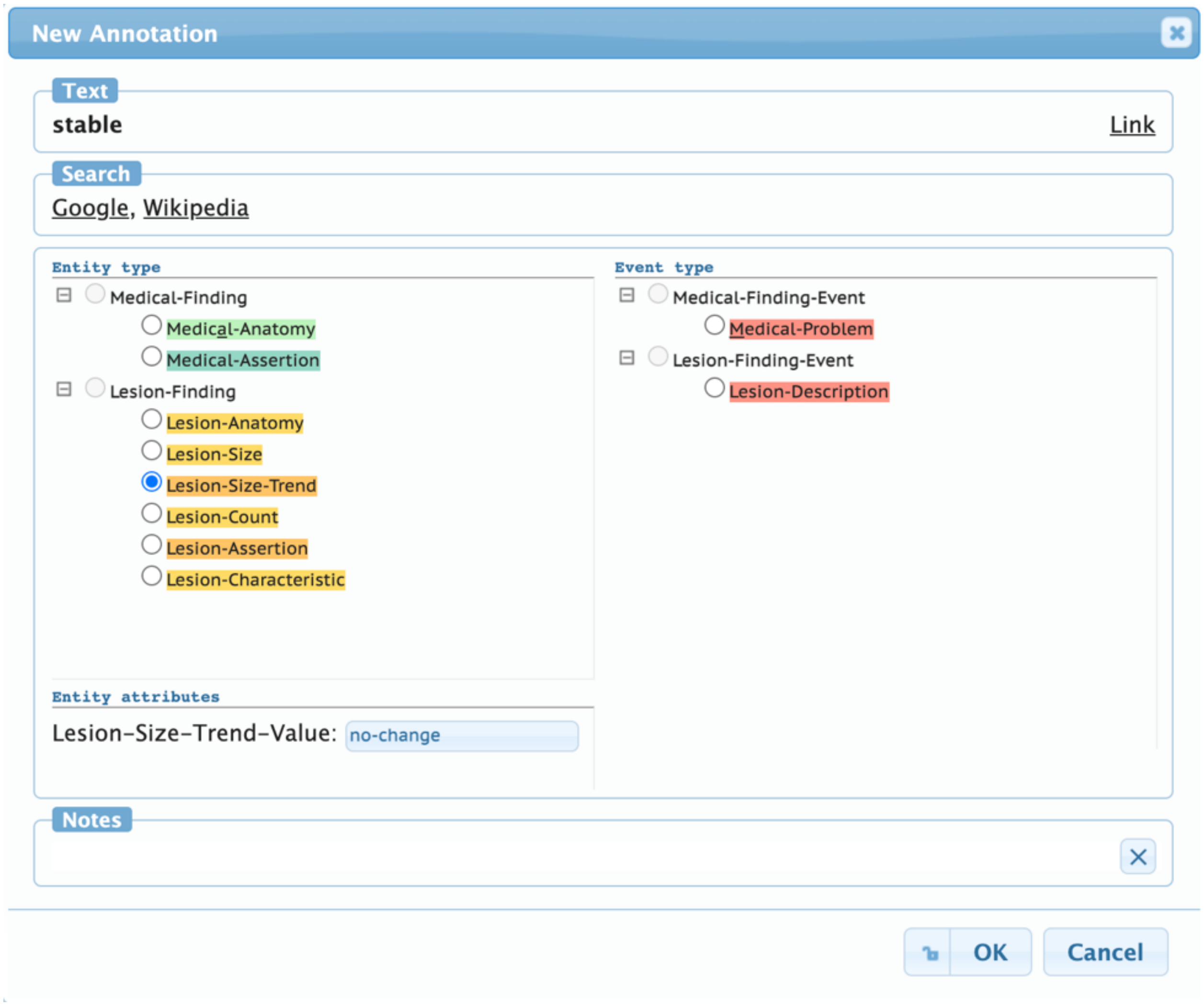


Figure A.1: Clinical finding events and their associated arguments.

Annotation is done on the BRAT tool. When highlighting a text span, a window will pop up showing the selections.



New Annotation

Text: stable — Link

Search: Google, Wikipedia

Entity type

- ⊟ ○ Medical-Finding
  - ○ Medical-Anatomy
  - ○ Medical-Assertion
- ⊟ ○ Lesion-Finding
  - ○ Lesion-Anatomy
  - ○ Lesion-Size
  - ● Lesion-Size-Trend
  - ○ Lesion-Count
  - ○ Lesion-Assertion
  - ○ Lesion-Characteristic

Event type

- ⊟ ○ Medical-Finding-Event
  - ○ Medical-Problem
- ⊟ ○ Lesion-Finding-Event
  - ○ Lesion-Description

Entity attributes

Lesion-Size-Trend-Value: no-change

Notes

OK Cancel

Figure A.2: BRAT tool entity selections for new annotation.

The right panel indicates the event triggers for the clinical findings. The left panel shows the entity types that are associated with each event trigger. The medical finding entities are highlighted in green while the lesion finding entities are in yellow. The ones in dark green and dark yellow are categorical which require selecting one of the possible values in the drop-down box down below. More details will be provided in the following sections. In general, entities highlighted in red are the event triggers and should be annotated first before others.

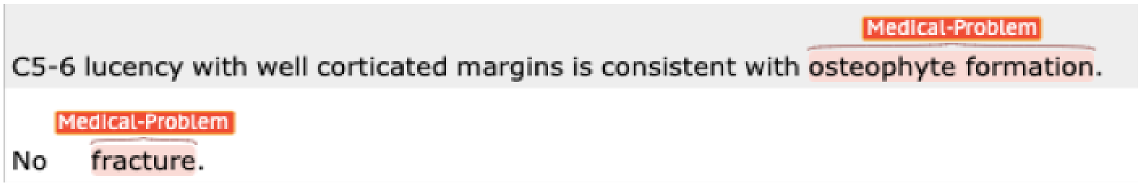
Annotating a text span involves selecting the entire text span with a mouse, and then choose one of the entities on Figure 2. If the entity is categorical, a drop-down box will be presented on the window. Choose the appropriate value from the drop-down box.

MEDICAL PROBLEM FINDING

Medical problem findings are abnormal pathological process uncovered by the radiology imaging test, such as cirrhosis, air-trapping, fracture, and effusion. A medical finding includes problem description, affected anatomy, and assertion.

A.1 Problem Description (required)

The description of medical problem serves as the event trigger. The text span can be a multi-word phase that identifies the actual medical problem, such as "osteophyte formation" and "fracture".

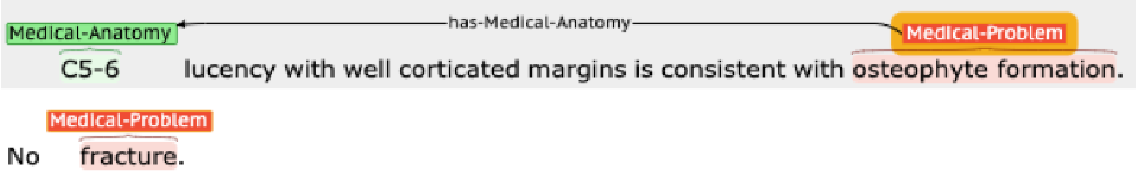


C5-6 lucency with well corticated margins is consistent with osteophyte formation.

No fracture.

A.2 Anatomy

Medical finding anatomy entity is a text span which captures one or more body parts associated with the medical problem, such as "C5-6" in the following example. Notice that a link "has-Medical-Anatomy" needs to be created from the trigger "osteophyte formation" to "C5-6" indicating their relationship. The link can be created on the UI by simply dragging an arrow from the event trigger to the anatomy entity.

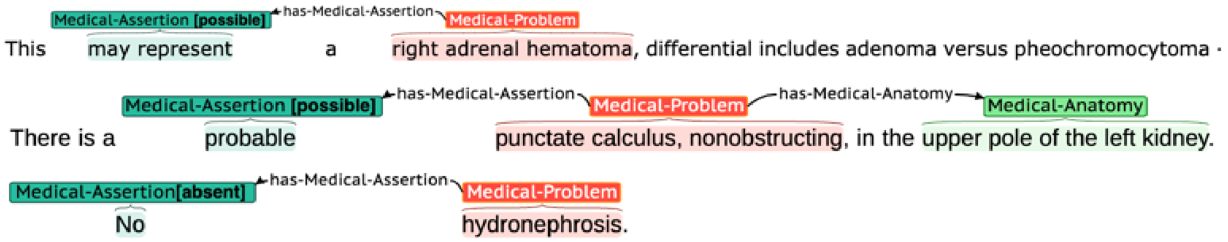


C5-6 lucency with well corticated margins is consistent with osteophyte formation.

No fracture.

A.3 Assertion

Medical finding assertion is a categorical value (possible, absent) indicating the likelihood of the medical problem. The following shows some assertions highlighted in dark green. Assertion has a default value of present. If no other explicit assertion value is annotated for the medical problem, it is implied that the medical problem has a present assertion.



This may represent a right adrenal hematoma, differential includes adenoma versus pheochromocytoma ·

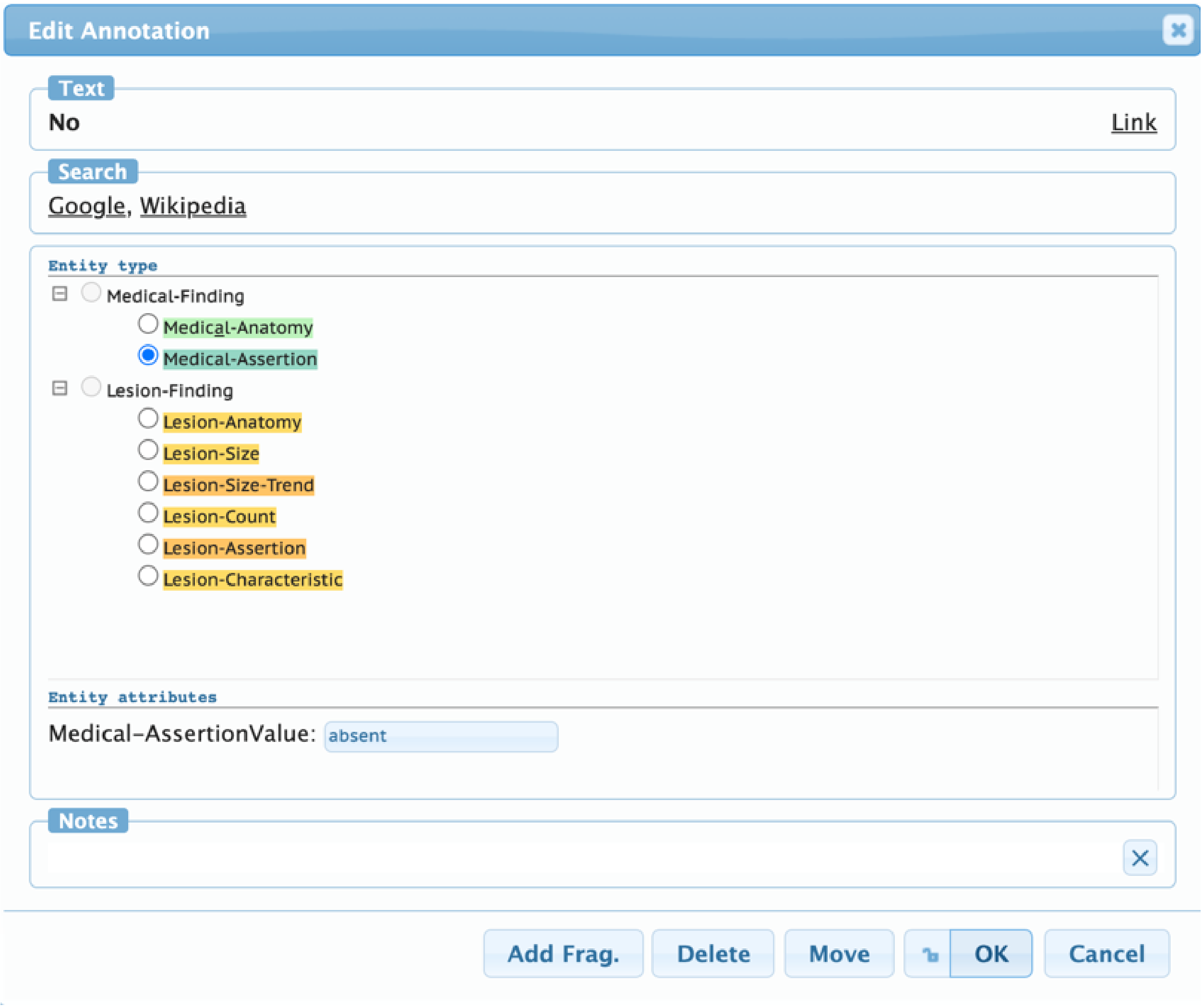
There is a probable punctate calculus, nonobstructing, in the upper pole of the left kidney.

No hydronephrosis.

The following table presents some examples of each category. The underlined text spans are the medical problem event triggers.

value	**examples**
possible	There is a **possible** nondisplaced L5 spinous process fracture. Liver: There is a mildly nodular contour of the liver as before, **possibly** representing cirrhosis.
present (default)	Calcified atherosclerosis of the LAD. C5-6 lucency with well corticated margins is consistent with osteophyte formation.
absent	**No evidence** of radiopaque nephrolith. Visualized osseous structures show **no** acute osseous abnormality.

When annotating a text span with this entity type or any categorical types, make sure a corresponding entity attribute is selected from the bottom drop-down on the BRAT tool. Select the values from the corresponding drop-down for each type



Edit Annotation

Text

No

Search

Google, Wikipedia

Entity type

- ⊟ ○ Medical-Finding
  - ○ Medical-Anatomy
  - ● Medical-Assertion
- ⊟ ○ Lesion-Finding
  - ○ Lesion-Anatomy
  - ○ Lesion-Size
  - ○ Lesion-Size-Trend
  - ○ Lesion-Count
  - ○ Lesion-Assertion
  - ○ Lesion-Characteristic

Entity attributes

Medical–AssertionValue: absent

Notes

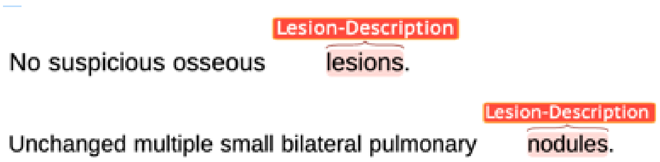
Figure A.3: Brat annotation screen showing categorical drop-down at the bottom.

LESION FINDING

Lesion finding describes the extent of lesion development that can be observed on the imaging, which includes description, anatomy, lesion size, size trend, count and assertions. Noun phrases containing anatomical location as part of a lesion description, (e.g. brain lesion or pulmonary nodules) should be annotated as two separate entities, i.e. lesion-anatomy (brain, pulmonary) and lesion-description (lesion, nodules).

A.4 Lesion description (required)

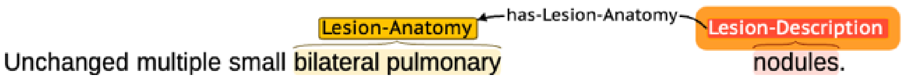
The description of lesion finding serves as the event trigger and is mandatory. Common text spans are (“mass”, “node”, “nodule”, “nodular opacity”, “lesion”). “Opacity” on its own is considered a medical problem.


No suspicious osseous lesions.

Unchanged multiple small bilateral pulmonary nodules.

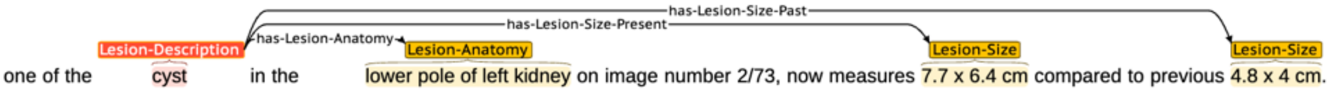
A.5 Anatomy

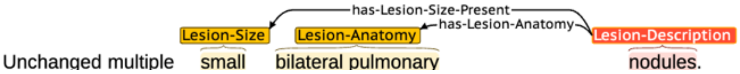
Lesion anatomy entity is a text span capturing one or more body parts where the lesion is located, such as “bilateral pulmonary” in this example. The links labelled “has-Lesion-Anatomy” indicate the relations between the Lesion-Anatomy entities and the corresponding lesion descriptions.


Unchanged multiple small bilateral pulmonary nodules.

A.6 Size

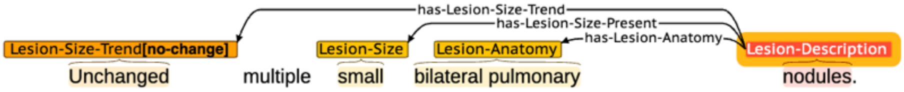
Some lesion descriptions contain size, such as “7.7 x 6.4 cm” and “4.8 x 4 cm” in the example. The links labelled “has-Lesion-Size-Present” indicate the relations between the Lesion-Size entities and the corresponding lesion descriptions. A separate relation “has-Lesion-Size-Past” should be used to link to lesion sizes in the past exams.


has-Lesion-Size-Past
has-Lesion-Size-Present
has-Lesion-Anatomy
Lesion-Description
Lesion-Anatomy
Lesion-Size
Lesion-Size
one of the cyst in the lower pole of left kidney on image number 2/73, now measures 7.7 x 6.4 cm compared to previous 4.8 x 4 cm.



A.7 Size trend

Some lesion descriptions contain size trend which is a categorical value (new, increasing, deceasing, no-change), such as the word “Unchanged” in the following example. The links labelled “has-Lesion-Size-Trend” indicate the relations between the Lesion-Size-Trend entities and the corresponding lesion descriptions.

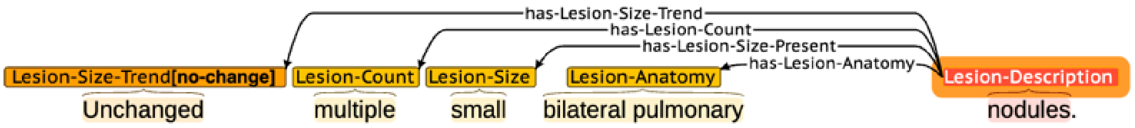


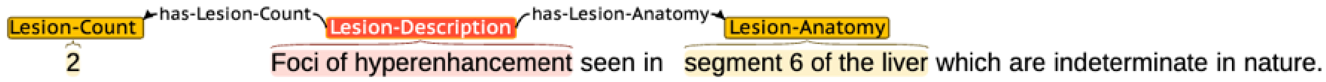
The following table presents some examples of each category. The underlined text spans are the medical problem event triggers.

value	**examples**
new	On the current exam, there is a **newly** identified hypodense well-delineated mass. Liver: There is a mildly nodular contour of the liver as before, **possibly** representing cirrhosis.
increasing	One nonhypermetabolic lymph node which is **increased in size**. More peripherally, there is 10 x 9 mm nodule adjacent to the suture line (4/63) which is gradually **increasing in size** since 2/17/2017, too small to characterize on PET.
decreasing	**Decreasing size** of the hypodense lesion within the inferior aspect of the right hepatic lobe now measuring 0.6 cm compared to 1.5 cm on 06/04/2014 The mass in the proximal ureter has **decreased significantly in size**, currently measuring 3 mm (4/104), **decreased** from 7 x 8 mm.
no-change	There is a hypoattenuating left adrenal nodule that has been increasing in size since 2009 though it is **unchanged** since May. Enlarged inferior mediastinal and right hilar lymph nodes are **unchanged** since January.

A.8 Count

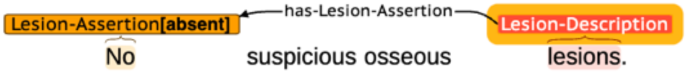
Some lesion findings include the number of nodules or lesions, such as "multiple" in the following example. The link "has-Lesion-Count" indicates the relation between the Lesion-Count entity and the corresponding lesion description.




Lesion-Count
has-Lesion-Count
Lesion-Description
has-Lesion-Anatomy
Lesion-Anatomy
2
Foci of hyperenhancement seen in segment 6 of the liver which are indeterminate in nature.

A.9 Assertion

Assertion is a categorical value (possible, absent) indicating the likelihood of the lesion finding, such as the word “no” in this example. The link “has-Lesion-Assertion” indicate the relation between the assertion entity and the corresponding lesion description. Like Medical Assertion, Lesion Assertion also has a default value of present. If no other explicit assertion value is annotated for the lesion finding, it is implied that the lesion finding has a present assertion.

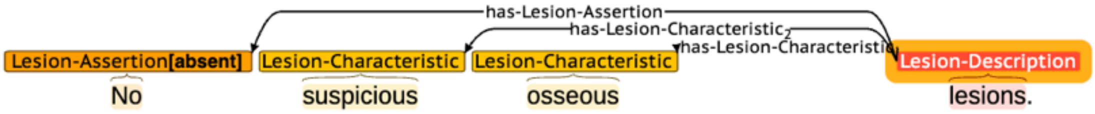


The following table presents some examples of each category. The underlined text spans are the medical problem event triggers.

value	examples
possible	**Cannot completely exclude** mass. Focal lesion seen in segment five shows delayed phase contrast washout indeterminate in nature, **possibly** dysplastic nodule vs. low grade HCC.
present (default)	Stable segment 7 metastasis status post radiation therapy. Intense FDG uptake (max SUV 17.1) is noted within 27 x 21 mm nodule in left lower lobe (4/76), consistent with biopsy-proven invasive adenocarcinoma.
absent	Findings: **No** suspicious enhancing nodule is seen. **No** obvious intracystic septations or mural nodularity are seen.

A.10 Characteristic

Characteristic attribute indicates the lesion characteristics such as the word “osseous” in this example. The link “has-Lesion-Characteristc” indicates the relation between the Characteristic entity and the corresponding lesion description.



Notes

- Avoid annotate articles (e.g. a, an, the), and unnecessary adjectives.
- Avoid annotate overlapping text spans. i.e. text spans overlapped with more than one annotation.
- Typical lesion description noun phrases ("mass", "node", "nodule", "nodular opacities", "lesion")
- "opacity" itself is considered a medical problem.
- Noun phrases containing anatomical location as part of a lesion description, (e.g. brain lesion or pulmonary nodules) should be annotated as two separate entities, i.e. lesion-anatomy (brain, pulmonary) and lesion-description (lesion, nodules).
- Avoid annotating assertion modifiers ('likely', 'possible') for non-triggers. E.g. in the span, *much less likely a metastatic lesion*, "much less likely" is not Lesion-Assertion (possible) for "lesion", as it is describing the extent of metastasis.