Summer 8-15-2018

Ontology-Based Clinical Information Extraction Using SNOMED CT

Jun Li

Ontology-Based Clinical Information Extraction Using SNOMED CT

A
Dissertation

Presented to the Faculty of
The University of Texas
Health Science Center at Houston
School of Biomedical Informatics
in Partial Fulfilment of the Requirements for the Degree of

Doctor of Philosophy

By

Jun Li, M.S.

University of Texas Health Science Center at Houston

2018

Dissertation Committee:

Hua Xu, PhD[1], Advisor
Cui Tao, PhD[1]
Jiajie Zhang, PhD[1]
Yang Gong, MD, PhD[1]

[1]The School of Biomedical Informatics

Dedication

To my parents Pingwen Li and Junfang Xiong

Acknowledgements

First, I must thank my advisor Professor Hua Xu for his mentorship. I would not be able to complete this dissertation without his guidance and help.

My deepest thanks to my other committee members, Dr. Cui Tao, Dr. Jiajie Zhang, and Dr. Yang Gong, for their advice on my dissertation research.

A special thank you to Dr. Yaoyun Zhang. Your suggestions and thoughtfulness made this dissertation better. Additionally, I would like to thank Dr. Ergin Soysal, Dr. Zongcheng Ji, Dr. Jun Xu, Harish Siddhanamatha, and Jingqi Wang for many help and support they have provided.

Finally, to my wife Jing, my son Jeffrey and my daughter Annie, thank you for giving me the time and support. I never thought I would ever attempt something like this, but you helped me believe I could.

Abstract

Extracting and encoding clinical information captured in unstructured clinical documents with standard medical terminologies is vital to enable secondary use of clinical data from practice. SNOMED CT is the most comprehensive medical ontology with broad types of concepts and detailed relationships and it has been widely used for many clinical applications. However, few studies have investigated the use of SNOMED CT in clinical information extraction.

In this dissertation research, we developed a fine-grained information model based on the SNOMED CT and built novel information extraction systems to recognize clinical entities and identify their relations, as well as to encode them to SNOMED CT concepts. Our evaluation shows that such ontology-based information extraction systems using SNOMED CT could achieve state-of-the-art performance, indicating its potential in clinical natural language processing.

Vita

1990.....Bachelor of Science, Information Science, Wuhan University

1998.....Master of Science, Computer Science, University of Houston

2006.....Master of Science, Health Informatics, University of Texas Health Science Center at Houston

Field of Study

Health Informatics

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

Rapid growth in the adoption of electronic health records (EHRs) has led to an unprecedented expansion in the availability of large practice-based clinical datasets. Tremendous efforts have been devoted to the secondary use of EHRs, which greatly promotes genomic, clinical, and translational research. One critical challenge of the secondary use of EHRs is that much of the clinically important information in EHRs is provided in unstructured clinical narratives only. Therefore, Natural Language Processing (NLP) technologies, which can extract structured information from narrative documents, have received great attention in the medical domain and many successful stories of applying NLP to the clinical text have been reported widely [1–3].

1.1 NLP in the Medical Domain

Clinical NLP has been an active research area of the Biomedical Informatics field for over 20 years. It is likely to become more important in the future because of the growth of healthcare and more advanced information technologies for electric data capture. NLP provides an efficient way to extract clinical information and encode them to concepts in standard terminologies, comparing to costly manual data extraction processes. Coded clinical concepts by NLP systems can be then used for downstream computational

applications, e.g., to improve the accuracy of information retrieval from a massive amount of EHR data [4].

1.1.1 NLP Tasks in the Medical Domain

Current clinical NLP activities range from lower to higher level tasks in term of the use of different linguistics information [5,6]. Typical low-level NLP tasks include:

- *Sentence Boundary Detection* (SBD) is the process of deciding where sentences begin and end. Most NLP tools require their input to be divided into sentences. It is challenging because punctuation marks are often ambiguous. For example, the periods in "m.g." denote abbreviation and in "Dr." denote title.
- *Tokenization* is the process of identifying individual words and punctuation marks as tokens within a sentence. The resulting tokens are then passed on to some other processes.
- *Part-of-speech Tagging* (POS Tagging) is the process of marking up a word in a text as corresponding to a particular part of speech. It is based on both the definition and context of the word. POS tagging is now done using algorithms in the context of computational linguistics.
- *Morphological Decomposition* is the process of decomposing a compound word into its constituent morphemes. Stemming and lemmatization are used to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. For example, words "am", "are", and "is" all have the common base form "be".

- *Shallow Parsing (chunking)* is the process of identifying phrases (noun groups, verb groups, etc.) from constituent part-of-speech tagged tokens. However, it does not specify their internal structure or their role in the sentence.
- *Problem-specific Segmentation* is the process of segmenting text into meaningful groups. For example, the clinical text could include sections as Chief Complaint, Past Medical History, etc.

Higher-level NLP tasks are usually built on low-level tasks and are often problem specific. They include:

- *Named Entity Recognition* (NER) [7,8] is to locate and classify specific words or phrases in text into pre-defined categories such as persons, locations, diseases, genes, or medications.
- *Word Sense Disambiguation* (WSD) [9,10] is to identify which sense or meaning of a word is used in a sentence, when the word has multiple meanings.
- *Relationship Extraction* is to detect and classify relationships between entities or events. For example, to extract relations between temporal expressions and clinical events [11,12]. This information can be used to infer that something has occurred in the past or may occur in the future.
- *Modifier Identification* [13–15] is to recognize the information modifying or completing the semantic indication of named entities or relations. For example, one important task is to infer whether a named entity is present or absent (negation) and to quantify the uncertainty of the inference.

- *Encoding or Normalization* [16–18] is to map named entities/relations to standard concepts/relations in a domain ontology. Assigning a code within a standardized coding system for a specific diagnosis or procedure provides a way of standardizing the recording of clinical information that can be subsequently used for a wide range of automated applications. Clinical coding is used for hospital billing, clinical audit, epidemiological studies, measuring treatment effectiveness, assessing health trends, cost analysis, health-care planning, and resource allocation [19].

1.1.2 NLP Applications in the Medical Domain

NLP has a wide range of potential applications in the medical domain [20]. Some important applications of NLP are as follows:

Information Extraction is the most common NLP application in biomedicine. It locates and structures specific information in the text. The structured information can be used for a number of different tasks. In biosurveillance, symptoms are extracted from the chief complaint field in the notes written for patients admitted to the emergency department of a hospital [21] or from ambulatory electronic health records [22] to help understand the prevalence and progression of a particular epidemic. In biology, biomolecular interactions extracted from different articles are used to construct biomolecular pathways [23]. In the clinical domain, pharmacovigilance systems use structured data obtained by NLP to discover adverse drug events [24].

Text Summarization produces a single text that synthesizes the main points from several input documents. It identifies and presents the salient points in texts automatically. There are several steps in the text summarization process. Content selection is to identify salient pieces of information in the input documents, content organization is to identify redundancy and contradictions among the selected pieces of information and to order them so the resulting summary is coherent, and content re-generation is to produce natural language from the organized pieces of information. Text summarization has focused on the literature [25,26].

Question Answering (QA) is a process of recognizing natural language questions, extracting the meaning, and providing the answer. This type of application becomes increasingly important as health care consumers, health care professionals, and biomedical researchers frequently search the Web to obtain information about diseases, medications, or medical procedures. A QA system can be very useful for obtaining the answers to factual questions, like "In children with an acute febrile illness, what is the efficacy of single-medication therapy with acetaminophen or ibuprofen in reducing fever?" [27]

1.1.3 Existing Clinical NLP Systems

Many NLP systems have been developed for analyzing clinical text. Linguistic String Project – Medical Language Processor (LSP-MLP) by Sager [28,29] at the New York University in 1965 was a pioneering NLP system and has greatly influenced subsequent

systems. Medical Language Extraction and Encoding (MedLEE) by Friedman [30,31] at the Columbia University in 1994 was designed for processing radiology reports and later extended to other domains. SymText and MPLUS by Haug [32,33] at the University of Utah in 1994 were created for processing chest radiograph reports. MetaMap by Aronson [7,34] at the National Library of Medicine in 1994 was developed for mapping biomedical text to concepts in the Unified Medical Language System (UMLS) Metathesaurus. Health Information Text Extraction (HITEx) by researchers at the Brigham and Women's Hospital and Harvard Medical School is an open-source clinical NLP system. The clinical Text Analysis and Knowledge Extraction System (cTAKES) [35] originated from the Mayo Clinic is an NLP system for extraction of information from electronic medical record clinical free-text. Clinical Language Annotation, Modeling, and Processing (CLAMP) by Xu [36] at the University of Texas School of Biomedical Informatics (SBMI) is a newly developed clinical NLP toolkit that provides not only state-of-the-art NLP components, but also a user-friendly graphic user interface that can help users quickly build customized NLP pipelines for their individual applications.

1.2 Ontology

A body of formally represented knowledge is based on a conceptualization: the objects, concepts, and other entities that are assumed to exist in some area of interest and the relationships that hold among them. Every knowledge base or knowledge-based system is committed to some conceptualization, explicitly or implicitly. An ontology is an explicit

specification of a conceptualization [37]. It is a declarative model of a domain that defines and represents the concepts existing in that domain, their attributes and the relationships between them. Ontology gives the description of concepts and the relations that can exist between them. The concept is very important for data sharing and knowledge representation [38].

Ontology can be classified according to the level of detailed knowledge they provide:

- *Upper Ontologies* provides very generic knowledge with low domain-specific knowledge.
- *General Ontologies* represent knowledge detail at an intermediate level. They are independent of a specific task.
- *Domain Ontologies* represent knowledge about a particular domain, such as medicine.
- *Application Ontologies* are designed for specific tasks.

1.2.1 Ontology in the Medical Domain

Numerous ontologies have been developed in the medical domain to represent biomedical terminology in common vocabularies so that they can be shared and reused across various fields. The billing terminologies such as International Classifications of Diseases (ICD), Diagnosis-related groups (DRGs), and Current Procedural Terminology (CPT) are used by all healthcare organizations to support aspects of medical billing. ICD is a diagnosis code set. ICD-10 is the version currently being used for billing in the U.S. and is also used for morbidity and mortality reporting. DRGs are commonly used in the

inpatient setting for billing a patient's hospital stay. CPT is used to code procedures for billing. Logical Observation Identifiers Names and Codes (LOINC) is used to encode lab observations and to represent clinical observations. The pharmacy terminologies are well-represented with many commercially available solutions like First Databank, Multum, Micromedex, and Medi-Span. The open-source RxNorm is the recommended pharmacy terminology for interoperability. Health Level 7 (HL7) is a messaging standard but also a terminology standard. It contains the code sets that aren't found in other standard terminologies, for example, the code sets for admission type and administrative gender. Generalised Architecture for Languages, Encyclopedia and Nomenclature in Medicine (GALEN) is a European project developed for reuse of terminology in clinical systems. It has been used to study nursing terminologies, decision support knowledge, surgical procedure, and anatomy. Foundational Model of Anatomy (FMA) structural represents knowledge about human anatomy.

Among them, the Unified Medical Language System (UMLS) [39] and the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) [40] have probably the greatest impact on biomedical ontology work because of their long history, their early focus on knowledge representation and its free availability.

1.2.2 The Unified Medical Language System (UMLS)

The Unified Medical Language System (UMLS) was created in 1986 and is maintained by the National Library of Medicine. It is a compendium of more than 100 controlled

vocabularies in the biomedical sciences. The UMLS provides a mapping structure among many health and biomedical vocabularies and standards to enable interoperability between computer systems. It may also be considered as a comprehensive thesaurus and an ontology of biomedical concepts and their relations.

The UMLS contains three knowledge sources:

The *Metathesaurus* includes over one million biomedical concepts and five million concept names from over 100 source vocabularies and code sets. Terms from each source vocabulary are organized by meaning and assigned a concept unique identifier (CUI). There are many categories in the Metathesaurus and vocabularies may fall into more than one category. Major vocabularies and categories include: Logical Observation Identifier Names and Codes (LOINC) in category Diagnosis, Current Procedural Terminology (CPT) in category Procedures & Supplies, International Classification of Diseases (ICD) in category Diseases, and Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) in category Comprehensive Vocabularies.

The *Semantic Network* provides the categorization of all concepts in the Metathesaurus by grouping concepts according to semantic types. Currently there are 133 semantic types and major semantic types include organism, anatomical structure, biologic function, chemical, physical object, and idea or concept. The Semantic Network also defines semantic relationships between semantic types. For example, the semantic type “Disease”

has a relationship “associated_with” with the semantic type “Finding”. There are 54 semantic relationships. Semantic types and semantic relationships create an information model that represents the biomedical domain.

The *SPECIALIST Lexicon* contains syntactic (syntax), morphological (inflection, derivation, and composition), and orthographic (spelling) information for biomedical terms as well as commonly occurring English words [41]. Currently it has over 200,000 terms and is used by the lexical tools for NLP tasks.

1.2.3 SNOMED CT

The Systematized Nomenclature of Medicine-Clinical Terms (SNOMED CT) was the 2002 merge result of the Systematized Nomenclature of Medicine (SNOMED) International originally developed by Dr. Roger Cote and the Clinical Terms Version 3 (CTV3) originally developed by Dr. James Read. SNOMED CT is maintained by the International Health Terminology Standards Development Organisation (IHTSDO). It is the most comprehensive, multilingual clinical healthcare terminology in the world [42].

SNOMED CT content is represented using three types of components:

Concepts representing clinical meanings are organized into hierarchies. Every concept has a unique numeric identifier called Concept ID. Within a hierarchy concepts range from the more general to the more detailed. This allows detailed clinical data to be

recorded and later accessed or aggregated at a more general level. For example, “Finding by site”, “Musculoskeletal finding”, “Joint finding”, “Arthropathy”, “Arthropathy of knee joint”, and “Arthritis of knee” are all concepts in “Clinical finding” hierarchy. But their granularities range from low to high. SNOMED CT currently contains more than 400,000 medical concepts, divided into 37 hierarchies.

Descriptions link appropriate human-readable terms to concepts. Every description has a unique numeric identifier called Description ID. A concept can have several associated descriptions, each description representing a synonym for the same concept. For example, “Weak heart”, “Cardiac failure”, and “Myocardial failure” are all descriptions of the concept “Heart failure (disorder)”. There are approximately 1,290,000 descriptions in SNOMED CT.

Relationships link each concept to other related concepts. Every relationship has a unique numeric identifier called Relationship ID. The relationships provide formal definitions and other properties of the concepts. One type of relationship is the “is a” relationship which is used to relate a concept to more general concepts. Related concepts in the concept hierarchy are linked using the “is a” relationship. For example, the concept “Arthropathy” has an “is a” relationship to the concept “Joint finding”. Attribute relationships are used to connect concepts in different hierarchies. For example, the concept “Appendicitis” in “disorder” hierarchy has an “associated morphology” attribute relationship to the concept “Inflammation” in “morphologic abnormality” hierarchy.

There are other types of relationships for representing aspects of the meaning of a concept. For example, the concept “Viral pneumonia” has a “causative agent” relationship to the concept “Virus” and a “finding site” relationship to the concept “Lung”. There are approximately 1,580,000 relationships, 65 unique relationship types and 836 different relationships between concepts in SNOMED CT.

1.3 Ontology-Based Information Extraction

Ontology-Based Information Extraction (OBIE) is a subfield of information extraction. In OBIE, ontologies are used as the backbone in the information extraction process and the output is generally presented through an ontology.

1.3.1 OBIE Definition

An OBIE system is a system that processes unstructured or semi-structured natural language text through a mechanism guided by ontologies to extract certain types of information and presents the output using ontologies [43]. There are key characteristics of OBIE systems:

- *Process unstructured or semi-structured natural language text*: OBIE system inputs can be either unstructured text files or semi-structured files using a particular template.
- *Present the output using ontologies*: The use of a formal ontology as the target output is an important characteristic that distinguishes OBIE systems from other IE systems.

- *Use an information extraction process guided by an ontology*: In OBIE systems, the information extraction process is guided by the ontology to extract classes, properties, and instances. No new information extraction method is invented but an existing method is oriented to identify the components of an ontology.

Figure 1-1 shows the general architecture of an OBIE system by Wimalasuriya and Dou [43].

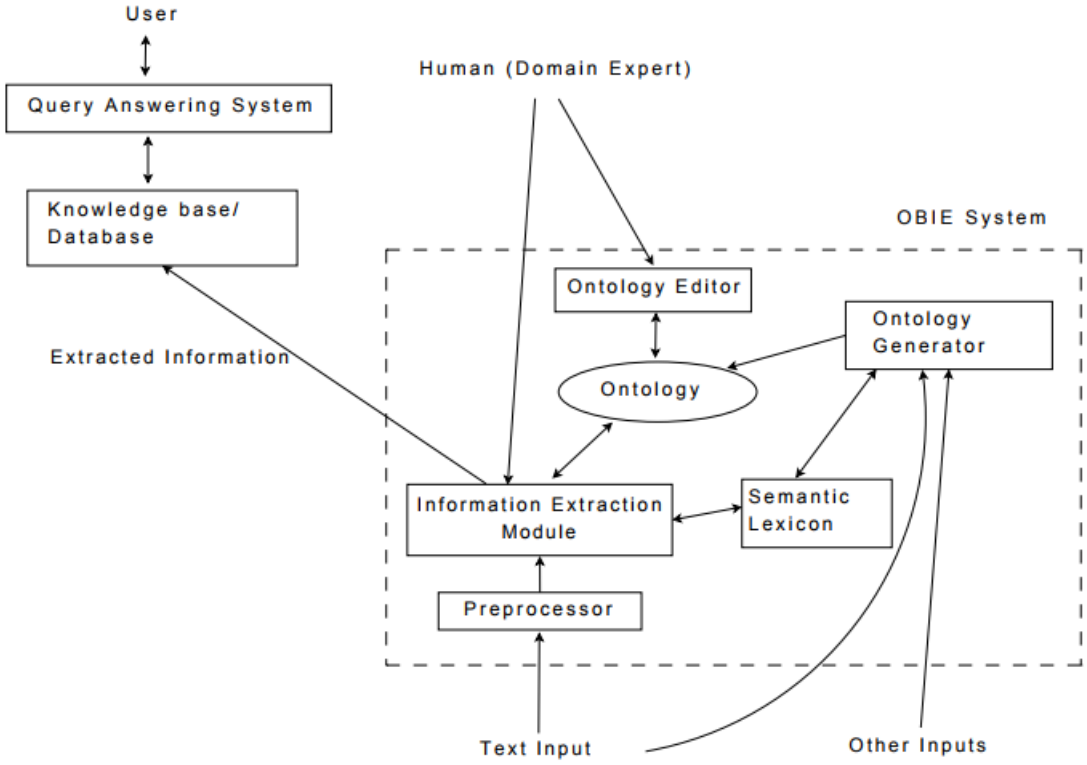

Figure 1-1. OBIE system, by Wimalasuriya and Dou

1.3.2 OBIE Methods in the Medical Domain

Many OBIE systems use linguistic rules to capture certain types of information. These rules are represented by regular expressions. For example, the expression `(diagnosed with <NP>)`, where `<NP>` denotes a noun phrase, might capture the names of diseases in a set of documents. By specifying a set of rules like this, it is possible to extract a significant amount of information. In practice, the rules are combined with NLP tools such as part-of-speech (POS) taggers and noun phrase chunkers. The General Architecture for Text Engineering (GATE) [44], which is a widely used NLP framework, provides an easy-to-use platform to employ this technique. Textpresso [45] and NLP-SNOMED [46] are examples of using this technique.

It is a common practice to convert an information extraction task into a classification task. When using classification for OBIE, classifiers are trained to identify different components of an ontology such as concepts and attribute values. Different classification techniques such as support vector machines (SVM), Hidden Markov Models (HMM), Conditional Random Fields (CRF), maximum entropy models, and decision trees have been used. Linguistic features such as POS tags, capitalization information and individual words are typically used as input for classification.

1.3.3 OBIE Systems in the Medical Domain

Most clinical NLP systems have encoding component which uses clinical ontologies to code clinical information. These systems can be seen as OBIE systems as well. Table 1-1

shows the existing NLP systems and the clinical ontologies used for encoding. The table was originally by Doan et al. [2] and we extended it with more NLP systems.

Table 1-1

Existing NLP Systems

System	**Creator**	**Ontology**	**Encoding**
MedLEE	Columbia University	Developed its own medical lexicons and terminologies	UMLS's CUI
SPRUS/SymText/MPLUS	University of Utah	UMLS	ICD-9
MetaMap	National Library of Medicine	UMLS	UMLS's CUI
HITEx	Harvard University	UMLS	UMLS's CUI
cTAKES	Mayo Clinic and IBM	UMLS + Trained models	UMLS's CUI and RxNorm
CLAMP	University of Texas	UMLS	UMLS's CUI

Currently, the UMLS are used as the clinical ontology for most of the NLP systems. However, the UMLS is not a classification system by design. It is a translation tool primarily designed for information retrieval. It is not sufficiently complete nor organized in such a way to serve as a controlled terminology. The UMLS is much more dichotomous (a clean hit or a clean miss) than SNOMED with substantially less completeness due to its precoordinated paradigm. It publishes terms from both

compositional and precoordinated schemes that may overlap without a definition of a canonical or preferred concept. It remains focused on the content of the source vocabularies that it connects and that material is not chosen primarily for clinical descriptive purposes [47].

1.4 Motivation and Specific Aims

NLP systems that can extract and encode clinical information captured in unstructured clinical narratives with concepts and relations in standard medical terminologies are vital to enable secondary use of clinical data. SNOMED CT is the most comprehensive medical terminology, covering broad types of concepts and well-defined semantic relationships. However, few studies have leveraged SNOMED CT for clinical NLP tasks. In this dissertation research, we propose to develop novel ontology-based information extraction approaches that leverage SNOMED CT for extracting important clinical concepts and relations in clinical text. Our hypothesis is that NLP systems guided by SNOMED CT can be built to effectively extract important clinical concepts and their relations with good performance. To achieve this goal, we propose the following specific aims:

Specific Aim 1 – Develop a fine granular information model based on SNOMED CT and clinical corpora. The information model will cover core clinical concepts and relations in the SNOMED CT. Additional concepts and relations of clinical importance that are only presented in clinical corpora will also be incorporated. An annotation guideline will be

developed with the guidance of the information model. Then a corpus of clinical notes will be manually annotated, which will be used as the gold standard for clinical concept recognition and relation extraction.

Specific Aim 2 – Recognize clinical entities defined in the information model using different NER approaches. This problem will be considered as a typical NER task. We will investigate three types of commonly used methods, the dictionary lookup based method, the conditional random field algorithm based on feature engineering, and deep learning based method using unsupervised features learned from the large-scale clinical dataset.

Specific Aim 3 – Extract relations between clinical entities and their modifiers following the information model using different algorithms. Relation extraction is essentially a classification problem. We will systematically compare a feature-based approach, a dependency graph kernel-based approach, and a joint learning based approach for this task.

Specific Aim 4 – Encode extracted clinical entities and modifiers into SNOMED concepts using different entity-linking algorithms. We will first manually assign SNOMED CT codes to extracted clinical entities organized in different granularities. The annotation will be used as the gold standard for training and evaluating our encoding methods. Next, we will propose novel encoding approaches using the Learning to Rank

framework with multiple features. In particular, a translation-based language model will be generated from synonym pairs in SNOMED CT, to capture the semantic correspondence of terms and alleviate the severe problem of string mismatch. We will compare the performance of our approaches with the encoding performances of existing clinical NLP systems such as MetaMap and cTAKES.

Chapter 2: SNOMED-based Information Model for Clinical NLP

2.1 Introduction

An information model is a representation of concepts and their relationships, properties and operations that can be performed on them, often created for a specific domain or a specific task. It provides the framework for organizing the information so that it can be delivered and reused. In many NLP tasks such as information extraction, an information model is often created based on semantic patterns in clinical documents and used to guide the annotation of clinical corpora [48]. Most of these information models are relatively simple, as they are often developed for a specific information extraction task, e.g., temporal information [49]. Few studies have investigated information models that cover broad types of clinical entities and relations. One important work is the information model used in the MedLEE system [50], which covers critical clinical concepts (e.g., problems, medications, and labs) and their allowable modifiers (e.g., negation and certainty). It is time-consuming to develop such comprehensive information models for clinical NLP as it often relies on the manual review of the targeted clinical documents.

Medical ontologies are often developed through iterative review and discussion by domain experts, and can naturally serve as information models for specific medical domains. However, many existing medical terminologies contain relatively simple

semantic types and relations (e.g., ICD is focused on disease and provides parent-child relation only), which do not cover comprehensive patterns occurred in the clinical text; therefore not very useful for clinical NLP tasks. One exception is the SNOMED CT, which contains broad types of clinical concepts and comprehensive relations among concepts. For example, the current version of SNOMED CT (September 2016 US Edition) contains 37 types of concepts and 65 types of relations. Nevertheless, few studies have investigated the use of the SNOMED CT as an information model for clinical NLP systems, probably due to its complexity.

In this chapter, we describe the first study of leveraging the SNOMED CT as an information model for developing clinical NLP systems. We assessed the actual occurrence of SNOMED CT concept types and relations in clinical text and refined them to build a practical information model for NLP, and then followed this information model to annotate a clinical corpus, which is used for following named entity recognition and relation extraction tasks.

2.2 SNOMED-based Information Model Development

SNOMED CT provides comprehensive types of clinical concepts and their relations. As the initial step, we focus on the several core clinical concepts such as clinical findings, procedure and medications. Besides, not all the concept and relation types are observed in clinical text. Therefore, one task here is to remove concept and relation types that are rarely seen in clinical documents. On the contrary, clinical documents may contain

additional important types of information that need to be captured, but are not represented in the SNOMED CT. Therefore, we need to add such additional concept and relation types into the information model for NLP.

2.2.1 Details of SNOMED CT

SNOMED CT is a core medical terminology that contains concepts with unique meanings and formal logic-based definitions, which are organized into hierarchies. Concepts are linked together into a semantic network in which different link types are used to express formal relationships. SNOMED CT content is represented using three types of component:

- Concepts representing clinical meanings are organized into hierarchies.
- Descriptions which link appropriate human-readable terms to concepts.
- Relationships which link each concept to other related concepts.

Figure 2-1 shows SNOMED CT components and how the concepts are organized in hierarchies [42].

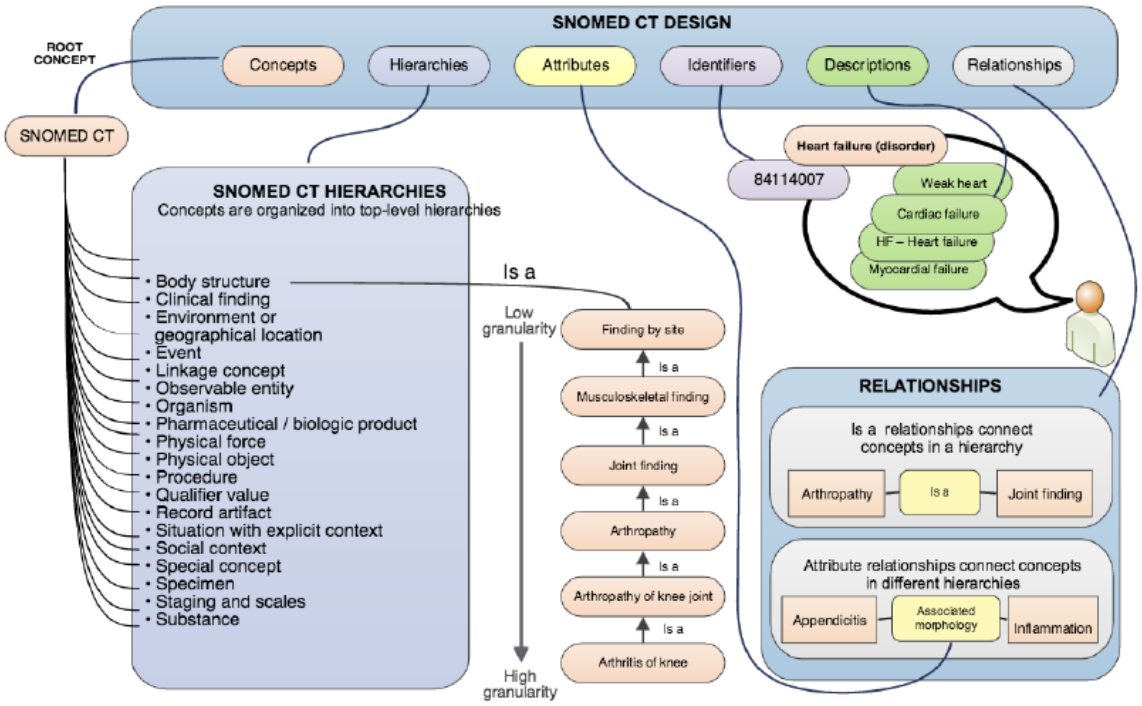

Figure 2-1. SNOMED CT Design, from SNOMED CT Starter Guide

In this dissertation, we used the September 2016 US Edition of SNOMED CT. Table 2-1 lists the SNOMED CT hierarchies with their semantic tags and total concept counts.

Table 2-1

SNOMED CT Hierarchies

Hierarchy	**Semantic Tag**	**Total Concepts**
Body structure	body structure	27,700
Body structure, altered from its original anatomical structure	morphologic abnormality	5,572
Cell structure	cell structure	519
Entire cell	cell	646

Clinical finding	finding	48,240
Disease	disorder	103,171
Environment or geographical location	environment / location	
Environment	environment	1,385
Geographical and/or political region of the world	geographic location	620
Event	event	9,016
Linkage concept	linkage concept	
Attribute	attribute	1,173
Link assertion	link assertion	8
Observable entity	observable entity	9,549
Organism	organism	37,946
Pharmaceutical / biologic product	product	25,285
Physical force	physical force	178
Physical object	physical object	15,890
Procedure	procedure	78,811
Regimes and therapies	regime/therapy	4,008
Qualifier value	qualifier value	10,886
Record artifact	record artifact	357
Situation with explicit context	situation	10,221
Social context	social concept	32
Ethnic group	ethnic group	374
Life style	life style	30
Occupation	occupation	6,497

Person	person	692
Racial group	racial group	21
Religion / philosophy	religion/philosophy	228
Special concept	special concept	31
Inactive concept	inactive concept	8
Namespace concept	namespace concept	201
Navigational concept	navigational concept	733
Specimen	specimen	1,798
Staging and scales	staging scale	41
Assessment scales	assessment scale	1,270
Tumor staging	tumor staging	262
Substance	substance	28,604

Note. Concepts with semantic tags "administrative concept", "biological function", "context-dependent category", and "foundation metadata concept" are inactive concepts. They are not included in this table.

As shown in Table 2-1, there are 37 hierarchies defined in SNOMED CT. Between these hierarchies, there are 65 unique relationship types and 836 different relationships.

2.2.2 Information Model Construction

2.2.2.1 Semantic Types for Clinical Concepts

After careful review of SNOMED CT by domain experts and discussion with NLP experts, we have selected the most clinically relevant semantic types for the information

model for NLP, most of which are top-level domains in SNOMED CT. In order to better represent the qualifier values related to their semantic meanings, we separated qualifier value concepts based on attribute, course, degree, episodicity, intent, laterality, priority, severity, and site.

In addition, following feedback from NLP experts, we added 3 new semantic types:

- Certainty: It is used to define if a clinical concept or fact is true or not.
- Demographics: It is used to define concepts related to a person's age, gender, marital status, name, race, etc. This type is similar to the SNOMED CT "Social context" type. The SNOMED CT "Social context" type has 6 subtypes. We combine "Social context" and all its subtypes into one "Demographics" type.
- Medication: It is used to define concepts related to the medications. SNOMED CT contains concepts for pharmaceutical products but it does not have medication brand names. For example, medication names such as Amoxicillin, Lipitor, etc. are not SNOMED CT concepts or descriptions.

Table 2-2 lists all the semantic types in the proposed information model. Column 2 in the table shows the corresponding SNOMED CT semantic type. Column 3 shows the semantic tag used for annotation in our corpus. Column 4 shows the abbreviation for the semantic type.

Table 2-2

Semantic Types in the Proposed Information Model

Semantic Type	**SNOMED CT Semantic Type**	**Semantic Tag**	**Abbreviation**
Body structure	Body structure	body_structure	BS
Certainty		certainty	CER
Clinical finding	Clinical finding, Disease	clinical_finding	CF
Demographics	Social context, Ethnic group, Life style, Occupation, Racial group, Religion / philosophy	demographics	DEM
Device	Physical object	device	DEV
Laboratory	Substance, Procedure	laboratory	LAB
Medication	Pharmaceutical / biologic product	medication	MED
Observable entity	Observable entity	observable_entity	OE
Organism	Organism	organism	ORG
Person	Social context -> Person	person	PER
Procedure	Procedure	procedure	PRO
Qualifier value - attribute	Qualifier value	qualifier_value::attribute	QV_AT
Qualifier value - course	Qualifier value	qualifier_value::course	QV_CO
Qualifier value - degree	Qualifier value	qualifier_value::degree	QV_DE

Qualifier value - episodicity	Qualifier value	qualifier_value::episodicity	QV_EP
Qualifier value - intent	Qualifier value	qualifier_value::intent	QV_IN
Qualifier value - laterality	Qualifier value	qualifier_value::laterality	QV_LA
Qualifier value - priority	Qualifier value	qualifier_value::priority	QV_PR
Qualifier value - severity	Qualifier value	qualifier_value::severity	QV_SE
Qualifier value - site	Qualifier value	qualifier_value::site	QV_SI
Substance	Substance	substance	SUB

2.2.2.2 Relationships for Clinical Concepts

The main relationships between clinical concepts included in the information model are:

Clinical finding

- Has_location (Body structure): This relationship shows the location of a clinical finding. The location refers to a body structure.
- Belons_to (Person): This relationship specifies the person from which the clinical finding information is obtained.
- Associated_with (Clinical finding | Procedure | Substance): This relationship represents a clinically relevant association between concepts.
- Has_causative_agent (Organism | Medication | Substance): This relationship identifies the direct causative agent of a disease. The agent refers to an organism, medication, or substance.

- After (Procedure): This relationship represents a sequence of events where a clinical finding occurs after a procedure.
- Has_finding_method (Procedure): This relationship specifies the means by which a clinical finding was determined.
- Due_to (Clinical finding): This relationship relates a clinical finding directly to a cause such as another clinical finding.
- Has_interpretation (Clinical finding): This relationship designates the judgment aspect being evaluated or interpreted for a concept when grouped with the attribute interprets. It may point to a finding value as a quantitative value; a qualitative value showing absence, degree increased; or a string value for normality, presence, etc.
- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Procedure

- Has_procedure_site (Body structure): This relationship describes the body site acted on or affected by a procedure.
- Has_focus (Clinical finding | Procedure): This relationship specifies the clinical finding or procedure which is the focus of a procedure.
- Has_interpretation (Clinical finding): This relationship designates the judgment aspect being evaluated or interpreted for a concept when grouped with the attribute interprets. It may point to a finding value as a quantitative value; a qualitative value showing absence, degree increased; or a string value for normality, presence, etc.

- Procedure_device (Device): This relationship describes the devices associated with a procedure.
- Using_substance (Substance): This relationship describes the substance used to execute the action of a procedure. It is not the substance on which the procedure's method directly acts.
- Has_location (Body structure): This relationship shows the location of a procedure. The location refers to a body structure.
- Direct_substance (Medication): This relationship describes the substance or pharmaceutical / biologic product on which the procedure's method directly acts.
- Has_method (Body structure): This relationship represents the action being performed to accomplish the procedure. It does not include the surgical approach, equipment or physical forces.
- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Laboratory

- Has_interpretation (Clinical finding | Organism): This relationship designates the judgment aspect being evaluated or interpreted for a concept when grouped with the attribute interprets. It may point to a finding value as a quantitative value; a qualitative value showing absence, degree increased; or a string value for normality, presence, etc.

- Has_intent (Clinical finding | Organism): This relationship specifies the intent of a laboratory test.
- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Observable entity

- Has_location (Body structure): This relationship shows the location of an observable entity. The location refers to a body structure.
- Has_interpretation (Clinical finding): This relationship designates the judgment aspect being evaluated or interpreted for a concept when grouped with the attribute interprets. It may point to a finding value as a quantitative value; a qualitative value showing absence, degree increased; or a string value for normality, presence, etc.
- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Medication

- Has_indication (Clinical finding): This relationship shows the reason for the treatment.
- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Body structure

- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Device

- Has_modifier (Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Organism

- Has_modifier (Certainty | Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

Substance

- Has_modifier (Qualifier value): This relationship specifies the values that further explain the concept behavior or properties.

2.2.3 Annotation Guideline Development

Based on the proposed information model, an annotation guideline is developed. It describes specific types of information that should be annotated, with examples found in the clinical texts. Some general considerations have been defined in the annotation guideline. It primarily covers the main concepts constructing a clinical encounter. These primary concepts are clinical findings, procedures, laboratory tests, and their values.

Supporting concepts like body structure, person, device, organism, are also required to be annotated to refer certain clinical information properly.

The meaningful concept with the finest granularity is required to be annotated with individual labels to the main concept and each of its modifier. For example:

She has acute chest pain this morning.

In this sentence, "acute chest pain" should be annotated as three separated concepts "acute", "chest", and "pain", each of which belongs to different semantic categories "modifier", "body structure", and "clinical finding" respectively.

We limit the scope of relation annotation to the same sentence. If two related concepts are in different sentences, their relationships should be ignored and not annotated.

2.3 Clinical Corpus Annotation Using the Information Model

Medical Transcription Examples and Sample Reports (MTSamples) website [51] contains sample transcribed medical reports for many specialties and different work types. For this study, we have randomly selected 103 discharge summary notes from MTSamples and used them to create an annotated clinical corpus.

Discharge summaries were given to two annotators for annotation based on the proposed information model and the annotation guideline. We used the annotation tool provided by the Clinical Language Annotation, Modelling and Processing Toolkit (CLAMP) in this

project. CLAMP leverages the BRAT annotation interface [52], as shown in Figure 2-2 about a screenshot of the annotation interface [36].

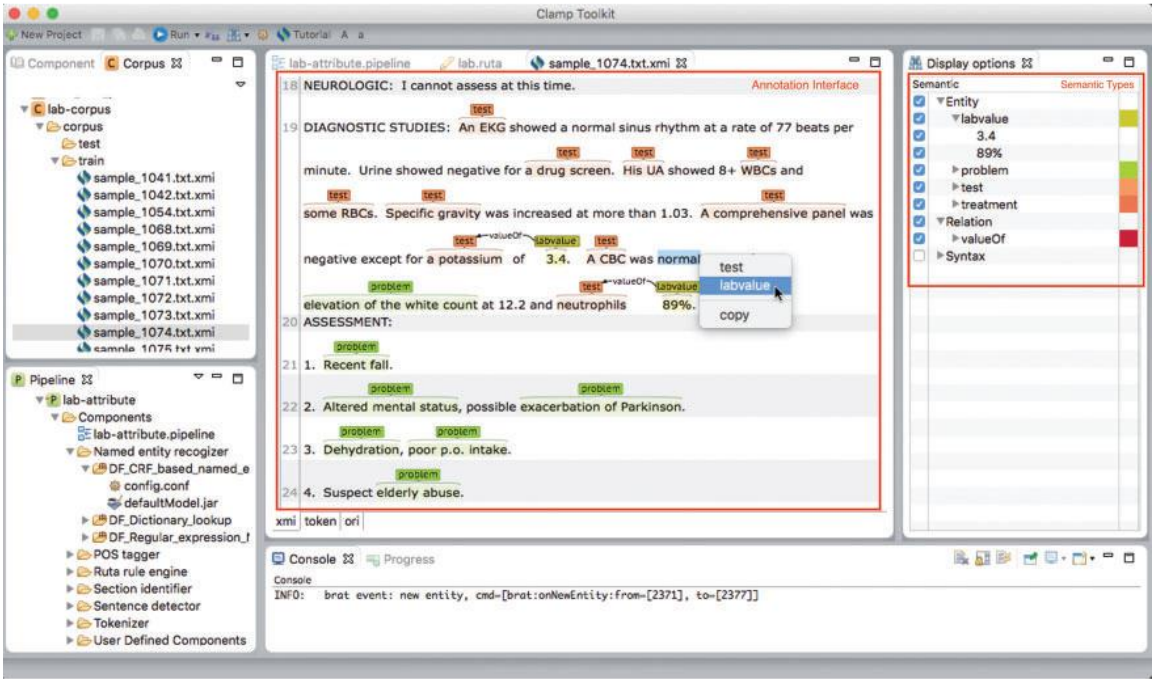

Figure 2-2. Annotation Interface in CLAMP

2.3.1 Inter-Annotator Agreement

Fleiss' kappa is a statistical measure for assessing the reliability of agreement between a fixed number of raters when assigning categorical ratings to a number of items or classifying items [53]. The calculated kappa value *k* could be interpreted using table 2-3.

Table 2-3

Kappa Value Interpretation

k	Interpretation
< 0	Poor agreement
0.01 – 0.20	Slight agreement
0.21 – 0.40	Fair agreement
0.41 – 0.60	Moderate agreement
0.61 – 0.80	Substantial agreement
0.81 – 1.00	Almost perfect agreement

To calculate inter-annotator agreement for our corpus annotation, each annotator was given the same 33 discharge summary notes for annotation. A total of 5,244 clinical concepts was annotated for 44 semantic types and 2,783 relations for 25 relationship types. R package 'irr' was used for calculating inter-annotator agreement. As shown in table 2-4, both clinical concept and concept relation annotations reach the substantial agreement between annotators. But the clinical concept has a much higher agreement value than that of concept relation, indicating relation annotation is a more challenging task.

Table 2-4

Inter-Annotator Agreement Results

	Concept	Relation
Annotated by both annotator, agree on the semantic type	4,303 (82.06%)	1,841 (66.15%)
Annotated by both annotator, not agree on the	101	27

semantic type	(1.93%)	(0.97%)
Annotated by annotator 1 only	412 (7.86%)	420 (15.09%)
Annotated by annotator 2 only	428 (8.16%)	495 (17.79%)
Total annotation	5,244	2,783
Total semantic types	44	25
Kappa value	0.803	0.612

2.3.2 Annotation Guideline Refinement

The annotation guideline was tuned and refined in several rounds of testing. Inter-annotator agreement rate was assessed in each round and the annotators met to discuss any disagreements. The annotation guideline was then updated based on the resolution and used in the next round of testing. The final version of the guideline was used to annotate the corpus.

2.3.3 Statistics of Annotated Corpus

After evaluating 103 discharge summary notes annotated by the two annotators, we have removed a few notes which only contain a few short sentences and selected 100 notes as our final corpus. The corpus has a total of 5,133 sentences, 10,932 concept annotations with 22 different semantic types, and 4,289 relation annotations with 61 different relation types between concepts. These annotations are used as the gold standard for the concept recognition and relation extraction work in the next steps. Table 2-5 and table 2-6 show the detailed statistics of the annotated corpus with some examples.

Table 2-5

Statistics of Annotated Corpus – By Entity Semantic Type

Entity Semantic Type	**Total**	**Examples**
clinical_finding	2,976	hypertension, obesity
body_structure	1,262	heart, abdomen
person	1,053	patient, sister
medication	1,046	Aspirin, Zyvox
procedure	1,028	biopsy, x-ray
laboratory	679	glucose, hemoglobin
qualifier_value::attribute	641	small, partial
observable_entity	407	
certainty	371	
qualifier_value::laterality	362	
demographics	260	
qualifier_value::site	202	
qualifier_value::severity	190	
device	188	
qualifier_value::course	168	
organism	47	
substance	23	
qualifier_value::episodicity	14	
qualifier_value::degree	10	
qualifier_value::priority	2	
qualifier_value::intent	2	

physical_object	1	
All	**10,932**	

Table 2-6

Statistics of Annotated Corpus – By Relation Type

Relation Type	**Entity From**	**Entity To**	**Total**	**Examples**
has_location	clinical_finding	body_structure	871	(pain, chest)
has_modifier	clinical_finding	certainty	471	(cancer, without)
has_modifier	clinical_finding	qualifier_value::attribute	434	(effusion, small)
belongs_to	clinical_finding	person	425	(nausea, patient)
has_procedure_site	procedure	body_structure	298	(CT, brain)
has_modifier	body_structure	qualifier_value::laterality	260	(kidney, left)
has_modifier	clinical_finding	qualifier_value::severity	194	(nausea, less)
has_modifier	clinical_finding	qualifier_value::course	165	(pain, chronic)
has_modifier	procedure	qualifier_value::attribute	133	(surgeries, multiple)
has_indication	medication	clinical_finding	119	
associated_with	clinical_finding	clinical_finding	108	
has_modifier	body_structure	qualifier_value::site	100	
has_focus	procedure	clinical_finding	96	
has_interpretation	procedure	clinical_finding	94	
has_modifier	clinical_finding	qualifier_value::laterality	62	
has_modifier	clinical_finding	qualifier_value::site	56	
procedure_device	procedure	device	52	
has_modifier	procedure	qualifier_value::site	42	

has_modifier	procedure	qualifier_value::laterality	40	
has_modifier	body_structure	qualifier_value::attribute	36	
has_causative_agent	clinical_finding	organism	20	
has_modifier	medication	qualifier_value::attribute	19	
after	clinical_finding	procedure	17	
has_causative_agent	clinical_finding	medication	17	
has_location	observable_entity	body_structure	14	
has_modifier	clinical_finding	qualifier_value::episodicity	14	
has_modifier	clinical_finding	qualifier_value::degree	9	
has_modifier	observable_entity	qualifier_value::attribute	9	
has_interpretation	laboratory	clinical_finding	8	
has_finding_method	clinical_finding	procedure	7	
has_modifier	device	qualifier_value::attribute	7	
due_to	clinical_finding	clinical_finding	6	
has_intent	laboratory	clinical_finding	6	
has_modifier	laboratory	qualifier_value::attribute	6	
using_substance	procedure	substance	6	
has_modifier	medication	certainty	5	
has_causative_agent	clinical_finding	substance	4	
has_focus	procedure	procedure	4	
has_intent	laboratory	organism	4	
has_interpretation	clinical_finding	clinical_finding	4	
has_interpretation	observable_entity	clinical_finding	4	
has_modifier	observable_entity	qualifier_value::laterality	4	
has_modifier	procedure	certainty	4	

direct_substance	procedure	medication	3	
has_interpretation	laboratory	organism	3	
has_location	procedure	body_structure	3	
has_modifier	laboratory	certainty	3	
has_modifier	medication	qualifier_value::course	3	
has_modifier	organism	certainty	3	
has_method	procedure	body_structure	2	
has_modifier	body_structure	certainty	2	
has_modifier	body_structure	qualifier_value::severity	2	
has_modifier	laboratory	qualifier_value::priority	2	
has_modifier	procedure	qualifier_value::intent	2	
associated_with	clinical_finding	procedure	1	
associated_with	clinical_finding	substance	1	
has_modifier	laboratory	qualifier_value::site	1	
has_modifier	medication	qualifier_value::site	1	
has_modifier	observable_entity	certainty	1	
has_modifier	organism	qualifier_value::attribute	1	
has_modifier	substance	qualifier_value::attribute	1	
All			**4,289**	

2.4 Discussion

Table 2-7 shows the comparison between SNOMED CT ontology and our proposed information model. We reduced the number of entity semantic types from 37 to 22 by merging and removing some SNOMED CT semantic types. However, fewer semantic types does not lose the coverage of our information model for clinical text. We only

removed the less clinically relevant semantic types such as “Linkage concept”, “Special concept”, etc. We greatly decreased the number of unique relation types and the number of relations between entity types to reduce the complexity of our information model. One of the important SNOMED CT relation type is “116680003 | Is a (attribute)” and it defines 37.4% of total relationships in SNOMED CT. It is used to link the related concepts in the concept hierarchy. We decided not to include it in our information model since our focus is on the modifier type relations between the concepts with different semantic types, not on the linking type relations between the concepts with the same semantic type.

Table 2-7

Comparison between SNOMED CT and Proposed Information Model

	SNOMED CT	**Proposed Information Model**
No. of Entity Semantic Types	37	22
No. of Unique Relation Types	65	17
No. of Relations Between Entity Types	836	61

After analyzing our annotated corpus, we discovered that “clinical finding” is a core semantic type in the clinical summary notes. Not only it has the most entity annotations (2,976 out of 10,932), it is also the semantic type which has the most relation types (26 out of 61) with other semantic types (16 out of 21).

2.5 Conclusion

In this study, we developed a comprehensive information model to represent broad types of clinical concepts and their relationships, by leverage the SNOMED CT oncology. Using the information model, we created an annotation guideline and annotated a corpus of 100 discharge summary notes. Our evaluation shows that annotators can follow the information model and the guideline to annotate discharge summaries with a good inter-annotator agreement. The annotated corpus is served for the concept recognition and relation extraction work in the next steps.

Chapter 3: Clinical Named Entity Recognition

3.1 Introduction

Recognition of clinically relevant entities such as diseases, drugs, and labs from the narrative text is the first step of the semantic interpretation of the clinical text. It is a typical Named Entity Recognition (NER) task, which is to locate and classify words and phrases into predefined semantic categories such as clinical findings and test results. Both rule-based methods and machine learning-based methods have been extensively studied for NER tasks.

Early clinical NLP systems often implement rule-based methods that use existing biomedical ontologies and knowledge engineering approaches to generate dictionaries for each semantic type and then perform dictionary lookup to identify clinical entities in the text [7,30,54]. For example, MedLEE [30] maintains large lexical files for different semantic types by leveraging existing medical terminologies and manually collecting terms from clinical corpora. One limitation of leveraging existing ontologies for semantic lexicons is that they may not cover all the terms occurred in the clinical text (i.e., lexical variants). Therefore, approaches have been developed to improve recognition of lexical variants, i.e., MetaMap [7] uses a variant generation tool from the UMLS's SPECIALIST lexicon [41].

Recently, machine learning-based NER approaches have shown superior performance in various clinical NER tasks. Machine learning-based approaches treat NER as a sequence labeling task and develop machine learning models to predict word labels using annotated corpora. As promoted by shared NLP tasks in the medical domain (i.e., i2b2 challenges [55]), extensively studies have been conducted to assess different aspects for improving machine learning-based NER, including different machine learning algorithms and diverse types of features. Conditional Random Fields (CRF) [56] and Structured Support Vector Machines (SVM) [57] are two widely used machine learning algorithms in NER. Features used in clinical NER also range broadly, including bag-of-word, part-of-speech tags, dictionaries etc., each of which more or less contributes to the performance improvements for different tasks [58].

More recently, deep learning-based methods are growing in popularity as approaches to NER. Deep learning-based methods do not need time-consuming and labor-intensive feature engineering [59,60]. Instead, word embeddings pre-trained from large-scale unlabelled corpora are usually used as features [61]. As the currently most widely-used distributional semantic representation (i.e., vector representation) of words, neural word embeddings (such as those produced by the word2vec software package [62]) are assumed to capture the latent syntactic/semantic information of a word, because the resulting vector representations for words will be similar if these words occur in similar local contexts [62]. A recent study by Habibi et al. demonstrated that using deep learning-

based methods outperformed state-of-the-art entity-specific NER tools and an entity-agnostic CRF implementation by a large margin [59].

The NER task here is to recognize entities defined in our information model derived from SNOMED CT, which contains broad types of entities (22 in total), thus making it different from previous tasks (i.e., i2b2 challenges) that are often limited to several types of entities [55]. We systematically assess all three types of approaches that are widely used in clinical NER for the proposed task: rule-based approaches leveraging existing ontologies, traditional machine learning-based NER using CRF, and deep learning-based approaches using LSTM.

3.2 Dataset

The annotated 100 discharge summaries were divided into two parts: a training set of 50 notes and a test set of 50 notes. The training set was used to generate the baseline semantic lexicon list for dictionary lookup and to train the machine learning-based NER models. The NER system was then evaluated using the test set. Table 3-1 lists the counts of each semantic type of clinical entities in the training and test datasets based on the gold standard annotation. The semantic types for numerical values are removed from the gold standard since they are relatively easy to recognize.

Table 3-1

Statistics of the Training and Test Datasets

Semantic Type	Training set	Test set	Total
clinical_finding	1,511	1,465	2,976
body_structure	670	592	1,262
person	570	483	1,053
medication	515	531	1,046
procedure	567	461	1,028
laboratory	241	438	679
qualifier_value::attribute	336	305	641
observable_entity	215	192	407
certainty	190	181	371
qualifier_value::laterality	243	119	362
demographics	118	142	260
qualifier_value::site	114	88	202
qualifier_value::severity	100	90	190
device	105	83	188
qualifier_value::course	81	87	168
organism	17	30	47
substance	16	7	23
qualifier_value::episodicity	9	5	14
qualifier_value::degree	7	3	10
qualifier_value::intent	2	0	2
qualifier_value::priority	0	2	2
physical_object	1	0	1

All	**5,628**	**5,304**	**10,932**

3.3 Rule-based Approach for Clinical Entity Recognition

Our rule-based method follows four steps: (a) generating a semantic lexicon list; (b) pre-processing discharge summary notes (i.e., sentence detection and tokenization); (c) locating clinical entities in the sentences by looking the lexicons; and (d) post-processing the matching results using heuristic rules.

3.3.1 Semantic Lexicon Generation

First, we created corpus-specific lexicons by using the gold standard annotation from the training set. The corpus-specific list contains 2,024 terms. Then we created another lexicon file by using SNOMED CT concepts and descriptions. The SNOMED lexicon file contains 707,772 terms.

As mentioned earlier, lexical variants are common in natural language. The variations may be morphological or simply orthographic [41]. Morphological variations generate different forms of the same lexical item through inflection or derivation. Orthographic variations generate different spellings of the same lexical item. Some words have several inflected forms which could be considered instances of the same word. For example, the verb “treat” has three inflectional variants: “treats” is the third person singular present tense form, “treated” is the past and past participle form, and “treating” is the present participle form.

The UMLS SPECIALIST Lexicon has been developed to provide the lexical information needed by NLP systems [63]. It includes both commonly occurring English words and the biomedical vocabulary. The syntactic, morphological, and orthographic information is recorded for each word or term. Therefore, we further extended our corpus-specific lexicons and the SNOMED CT lexicons by including the lexical variations specified in the UMLS SPECIALIST Lexicon. After that, the extended corpus-specific list contains 3,916 terms and the extended SNOMED CT list contains 760,218 terms.

3.3.2 Pre-Processing Discharge Summary Notes

We use the CLAMP toolkit [36] for pre-processing the discharge summary notes. CLAMP provides the components for common NLP tasks such as sentence boundary detection, tokenization, part-of-speech tagging, and section header identification. Using these components, we divide a discharge summary note into sections, sentences, and tokens with POS tags.

3.3.3 Dictionary Lookup Methods

Pattern-based regular expression [64] match and dictionary lookup were implemented to locate clinical entities of interest. Based on our observation, certain patterns were defined using the regular expression. For example, pattern "(\b)(\d+year-old)(\b)" is used to locate lexicons which describe the age with the semantic type "demographics" such as "37-year-old"; pattern "(\b)(Dr\. [A-Z][a-zA-Z]*)(\b)" is used to locate lexicons which describe the doctor names with the semantic type "person" such as "Dr. XYZ".

For dictionary lookup, each term in the generated semantic lexicon file was used to search the sentence. SNOMED CT has recommended a list of stop words and excluded words [65], which were removed from the lexicon list to increase the success of finding lexical matches. Our matching algorithm returns the exact matches.

3.3.4 Post-Processing the Matching Results

There are instances whereby multiple lexicon matches are found for the same word/phrase. For example, in "chest x-ray", there are three matching lexicons: "Chest (body structure)", "X-ray (procedure)", and "Chest X-ray (procedure)". Our rule is to select individual lexicons to the main concept and each of its modifier, which is the most granular description of the clinical concept. In the example above, "Chest (body structure)" and "X-ray (procedure)" will be the final results.

3.3.5 Experiments and Evaluation

To evaluate the effect of different lexicon lists, we started with the SNOMED CT lexicons as the baseline, and then combined corpus specific lexicons with it. We further compared the performance of extended lexicons using the UMLS SPECIALIST for both SNOMED CT lexicons, SNOMED CT + corpus-specific lexicons.

To report the performance of NER, we counted True Positives, True Negatives, False Positives, and False Negatives by comparing systems' results with the gold standard. We

then calculated standard metrics including Precision, Recall, and F1-score to report the performance of the NER systems:

$$Recall(R) = \frac{TruePositives(TP)}{TruePositives(TP) + FalseNegatives(FN)}$$

$$Precision(P) = \frac{TruePositives(TP)}{TruePositives(TP) + FalsePositives(FP)}$$

$$F\ Score(F) = \frac{2 * (Precision * Recall)}{(Precision + Recall)}$$

3.3.6 Results

Table 3-2 showed the results of our NER system when different lexicon files were used. The combined list of corpus-specific lexicons and SNOMED CT lexicons achieved the best F1-score of 0.506. It also achieved the best precision value of 0.381. Compared to SNOMED CT lexicons only, the combined lexicon list increased the precision/recall/F-score by 0.048/0.182/0.086 respectively. Extending the lexicon lists with UMLS SPECIALIST did not improve the performance. Although extended combination list achieved the best recall value of 0.759, its F-score decreased by 0.036 due to the 0.04 decrease of precision.

Table 3-2

Results of clinical entity recognition when different lexicon files were used

Lexicon List	**No. of Lexicons**	**No. of Entities**			**Performance**		
		Correct (TP)	**Predict (TP+FP)**	**Gold (TP+FN)**	**Precision**	**Recall**	**F**

SNOMED CT	707,772	3,384	10,165	5,933	0.333	0.570	0.420
Corpus Specific + SNOMED CT	709,792	4,460	11,698	5,933	**0.381**	0.752	**0.506**
Extended SNOMED CT (SPECIALIST)	760,218	3,611	11,893	5,933	0.304	0.609	0.405
Extended Corpus Specific + SNOMED CT (SPECIALIST)	764,130	4,504	13,219	5,933	0.341	**0.759**	0.470

Table 3-3 shows the detailed results of different semantic types for the best-performing system (Corpus-specific lexicons + SNOMED CT lexicons). The dictionary lookup-based approach achieved varied performance for different types of entities. Some semantic types achieved high performance even for this simple approach, e.g., precision/recall/F-score were 0.906/0.961/0.933 respectively for the semantic type of "person". Some types of entities had a very low frequency in the dataset, thus producing an extremely low performance.

Table 3-3

Results of clinical entity recognition by semantic type

Semantic Type	**Precision**	**Recall**	**F**
clinical_finding	0.603	0.762	0.673
body_structure	0.567	0.730	0.638
procedure	0.552	0.753	0.637
medication	0.698	0.687	0.692

person	**0.906**	**0.961**	**0.933**
laboratory	0.872	0.742	0.802
qualifier_value::attribute	0.240	0.598	0.343
observable_entity	0.313	0.608	0.413
certainty	0.185	0.735	0.296
qualifier_value::laterality	0.865	0.919	0.891
demographics	0.739	0.944	0.829
qualifier_value::site	0.597	0.527	0.560
device	0.455	0.670	0.542
qualifier_value::course	0.905	0.731	0.809
qualifier_value::severity	0.485	0.810	0.607
organism	0.682	0.649	0.665
substance	0.211	0.778	0.332
qualifier_value::episodicity	0.148	1.000	0.258
qualifier_value::degree	0.730	1.000	0.844
qualifier_value::intent	0.000	0.000	0.000
qualifier_value::priority	0.036	1.000	0.069
physical_object	0.000	0.000	0.000
Overall	**0.381**	**0.752**	**0.506**

3.4 Machine Learning-based Approach for Clinical Entity Recognition

Here we present our work on developing machine learning-based NER system for the 22 types of clinical entities, using the CRF algorithm, as well as a set of comprehensive features.

3.4.1 Conditional Random Fields

Conditional Random Fields [66] are undirected graphical models, used to calculate the conditional probability of values on designated output nodes, given values to other designated input nodes. A CRF is a type of discriminative probabilistic model used for labeling sequential data such as natural language text. When applying CRF to the NER problem, the observation sequence is the tokens of a sentence and the state sequence is its corresponding label sequence.

CRFs make first-order Markov assumption. They can be viewed as conditionally trained probabilistic finite automata (FSMs). The conditional probability $P(S/O)$ of a state sequence s=<s_1,s_2….s_T> given an observation sequence o=<o_1,o_2…..o_T> is

$$P(S/O) = \frac{1}{Z_0} exp \sum_{t=1}^{T} \sum_{k} \gamma_k f_k(S_{l-1}, S_1, O, t)$$

where $f_k(S_{l-1}, S_1, O, t)$ is a feature function. Its weight γ_k is to be learned via learning. CRFs define the conditional probability $P(l/O)$ of a label sequence l based on total probability over the state sequences,

$$P(l/O) = \sum_{s:l(s)=1} P(S/O)$$

where l(s) is the sequence of labels corresponding to the labels of the states in sequences. Z_o is a normalization factor over all state sequences. To make all conditional probabilities sum up to 1, we must calculate the normalization factor

$$Z_0 = \sum_S exp \sum_{t=1}^{T} \sum_k \gamma_k f_k(S_{l-1}, S_1, O, t)$$

The feature functions could ask arbitrary questions about two consecutive states, any part of the observation sequence and the current position. For example a feature function may be defined to have a value 0 in most cases and have value 1 when S_{t-1}, S_t are certain states and the observation has certain properties.

The annotated notes are transformed into the BIO (begin-in-out) annotation format, in which each word is assigned into a label: B represents the beginning of an entity, I represents inside of an entity, and O represents outside of an entity. For example, the sentence "His midline incision is clean" will be labeled as "His/O midline/B incision/I is/O clean/O", if "midline incision" is annotated as an entity. The NER task then becomes a classification task. It is to assign one of the three labels (B, I, or O) to each word based on the characteristics and its context. For each type of entity, we define different B classes and I classes. For example, for "clinical finding" type, the B class is defined as "B-ClinicalFinding" and I class is defined as "I-ClinicalFinding". There is only one O class for all the entity types.

3.4.2 Feature Sets

CRFs can easily include a large number of arbitrary independent features. The expressive power of models increases when adding new features that are conjunctions to the original features.

The feature sets used in our CRF approach are:

- **N-Gram**: These are sequences of words of length N.
- **Prefix and Suffix**: Many diseases and treatments share same prefix or suffix, like Adrenalectomy, Sclerotomy, and Osteotomy all shares a common suffix "-tomy". Word suffix and prefix are used as features.
- **Word Shape**: There can be many variants of the same medical entity in the clinical text, like hypertension and hypertensive, tachycardia and tachycardic.
- **Words Regular Expression**: These are regular expression patterns used for matching.
- **Dictionary Lookup**: A binary unigram feature was used to check whether the word is present in a dictionary of specific types of entities (e.g., diseases, drugs, and labs) or not.
- **Sentence Pattern**: These are information of the sentence, like sentence length, the start pattern, etc.
- **Section Headers**: A clinical note is often divided into relevant segments called Section Headers, like History of Present Illness, Current Medicines, and Lab Data. These section headers provide very useful information at the discourse level.

- **Random Indexing**: Very high dimensional Vector Space Model (VSM) implementations are impractical. Random indexing is an incremental method for constructing a vector space model with reduced dimensionality.
- **Word Embedding**: Words and phrases from the notes are mapped to vectors of real numbers. It involves a mathematical embedding from a space with one dimension per word to a continuous vector space with much lower dimension.
- **Brown Clustering**: It groups words into clusters that are semantically related by virtue of their having been embedded in similar contexts.

3.4.3 Experiments and Evaluation

CLAMP Toolkit has a machine learning-based NER component that uses the CRF algorithm. We used CLAMP with a unique set of features for recognizing clinical entities. We started the experiment with the basic word features plus the unigram feature. We then incrementally added other features such as bigram, sentence pattern, word embedding, etc. Same standard metrics described in section 3.3.5 were used for evaluation. After comparing the best performance achieved in each feature combination, we decided which features to keep and which features to remove. Figure 3-1 shows the feature sets used in CLAMP.

NER feature extractors	Descriptions
DF_Brown_clustring_feauture	Run brown clustering first, then add to NER features;
DF_Dictionary_lookup_feature	Run dictionary matching first, then add to NER features;
DF_Ngram_feature	Ngram features of words and part-of-speeches;
DF_Prefix-suffix_feature	Prefix & suffix of the words
DF_Random_indexing_feature	Run random indexing first, then add to NER features;
DF_Section_feature	Add section headers to NER features;
DF_Sentence_pattern_feature	Add sentence info(length, start pattern...) to NER features;
DF_Word_embedding_feature	Run word embedding first, then add to NER features;
DF_Word_shape_feature	Add word shapes(number, stem...) to NER features;
DF_Words_regular_expression_feature	Regular expressions matching, then add to NER features;

Figure 3-1. Feature Sets used for CRF-based NER in CLAMP

3.4.4 Results

The results in Table 3-4 were evaluated using both exact matching, which requires that the starting and ending offsets of a concept have to be exactly same as those in the gold standard, and inexact matching, which refers to cases where their offsets are not exactly same as those in gold standard, but they overlap with each other. The overall precision value is 0.813, recall value is 0.769, and F score is 0.790 for exact matching. The overall precision value is 0.876, recall value is 0.821, and F score is 0.848 for inexact matching.

Among 22 semantic types, two had F-scores higher than 0.90 and five had F-scores higher than 0.80 for exact matching. When inexact matching was used, five semantic types had F-scores higher than 0.90 and three had F-scores higher than 0.80. For semantic types which had low performance (F score < 0.50), all of them had very small sample sizes (size < 50). In general, the performance increases when the sample sizes increases.

Table 3-4

Results of clinical entity recognition (CRF)

Semantic Type	**Exact matching**			**Inexact matching**		
	Precision	**Recall**	**F**	**Precision**	**Recall**	**F**
clinical_finding	0.831	0.828	0.829	0.901	0.894	0.898
body_structure	0.756	0.802	0.778	0.837	0.857	0.847
person	0.933	0.914	0.923	0.950	0.929	0.939
medication	0.850	0.833	0.841	0.932	0.905	0.919
procedure	0.751	0.664	0.705	0.851	0.748	0.796
laboratory	0.841	0.797	0.818	0.911	0.854	0.881
qualifier_value::attribute	0.656	0.580	0.616	0.673	0.590	0.628
observable_entity	0.757	0.619	0.681	0.832	0.681	0.749
certainty	0.761	0.617	0.682	0.804	0.652	0.720
qualifier_value::laterality	0.928	0.928	0.928	0.931	0.931	0.931
demographics	0.912	0.873	0.892	0.980	0.935	0.957
qualifier_value::site	0.619	0.490	0.547	0.662	0.515	0.579
qualifier_value::severity	0.697	0.568	0.626	0.727	0.589	0.651
device	0.774	0.473	0.587	0.870	0.532	0.660
qualifier_value::course	0.924	0.863	0.892	0.942	0.875	0.907
organism	0.714	0.213	0.328	1.000	0.277	0.433
substance	0.500	0.087	0.148	0.500	0.087	0.148
qualifier_value::episodicity	0.333	0.071	0.118	0.333	0.071	0.118
qualifier_value::degree	0.000	0.000	0.000	0.000	0.000	0.000
qualifier_value::priority	0.000	0.000	0.000	0.000	0.000	0.000

qualifier_value::intent	0.000	0.000	0.000	0.000	0.000	0.000
physical_object	0.000	0.000	0.000	0.000	0.000	0.000
Overall	**0.813**	**0.769**	**0.790**	**0.876**	**0.821**	**0.848**

3.5 Deep Learning-based Approach for Clinical Entity Recognition

Here we present our work on developing deep learning-based NER system for the clinical entities, using the LSTM-CRF model.

3.5.1 LSTM-CRF Model

Recurrent neural networks (RNNs) are a family of neural networks that operate on the sequential data. They take input as a sequence of vectors (x_1, x_2, …, x_n) and they return another sequence (h_1, h_2, …, h_n) that represents some information about the sequence at every step in the input. Though, RNNs can learn long-distance dependencies in theory, they fail to do so in practice due to the gradient vanishing and tend to be biased towards their most recent inputs in the sequence [67]. Long Short-term Memory Networks (LSTMs) have been designed to solve this gradient vanishing issue. They incorporate a memory-cell and have been shown to capture long-distance dependencies [68]. The following implementation is used:

$$i_t = \sigma(W_{xi}X_t + W_{hi}h_{t-1} + W_{ci}C_{t-1} + b_i)$$

$$C_t = (1 - i_t)\odot C_{t-1} + i_t \odot \tanh(W_{xc}X_t + W_{hc}h_{t-1} + b_c)$$

$$O_t = \sigma(W_{xo}X_t + W_{ho}h_{t-1} + W_{co}C_t + b_o)$$

$$h_t = O_t \odot \tanh(C_t)$$

where σ is the element-wise sigmoid function, and $\odot$ is the element-wise product.

For a given sentence (x_1, x_2, …, x_n) containing n words, each represented as a *d*-dimensional vector, an LSTM computes a representation of the left context of the sentence at every word. A second LSTM reads the same sequence in reverse. The former is referred to as the forward LSTM and the latter as the backward LSTM. These are two distinct networks with different parameters. This forward and backward LSTM pair is referred to as a bidirectional LSTM [69].

The representation of a word in this model is obtained by concatenating the left and right context representations of the word. These representations effectively include the representation of a word in context. Despite this model's success in simple problems like POS tagging, its independent classification decisions are limiting when there are strong dependencies across output labels in NER task.

Therefore, instead of modeling tagging decisions independently, we model them jointly using a conditional random field [56]. Figure 3-2 shows the neural network architecture of the Bi-LSTM algorithm.

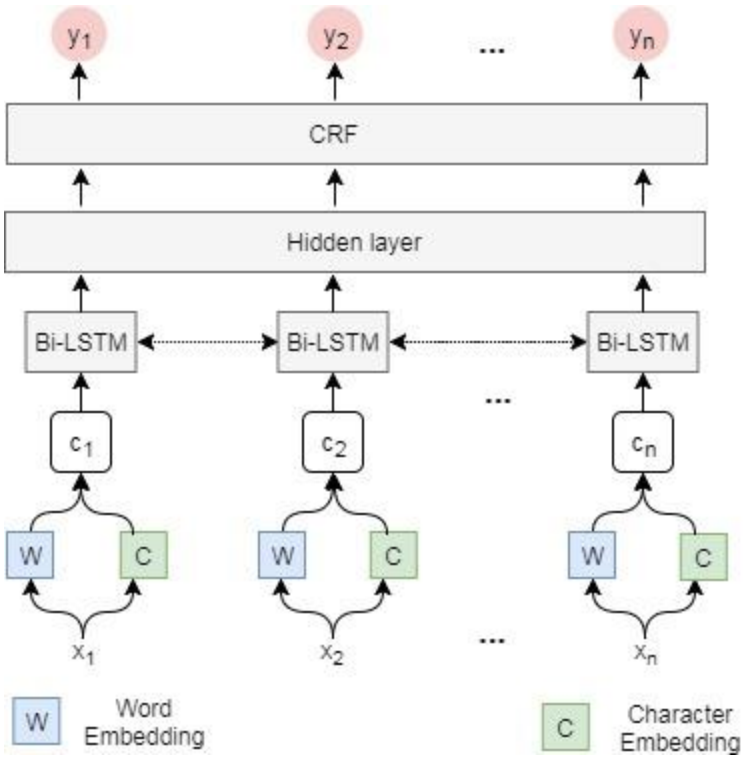

Figure 3-2. Neural Network Architecture of the Bi-LSTM Algorithm

3.5.2 Experiments and Evaluation

Our architecture is similar to the ones presented by Lample et al. [61]. Same standard metrics described in section 3.3.5 were used for evaluation.

3.5.3 Results

Table 3-5 shows the detailed results of different semantic types. Among 22 semantic types, three had F-scores higher than 0.90 and four had F-scores higher than 0.80. Similar to the previous machine learning-based CRF approach, for semantic types which had low performance (F score < 0.50), all of them had very small sample sizes (size < 50). In general, the performance increases when the sample sizes increases.

Table 3-5

Results of clinical entity recognition (LSTM-CRF)

Semantic Type	**Precision**	**Recall**	**F**
clinical_finding	0.811	0.799	0.805
body_structure	0.714	0.758	0.736
procedure	0.703	0.647	0.674
medication	0.811	0.831	0.821
person	0.948	0.908	0.927
laboratory	0.823	0.864	0.843
qualifier_value::attribute	0.604	0.516	0.557
observable_entity	0.705	0.680	0.692
certainty	0.741	0.671	0.704
qualifier_value::laterality	0.899	0.924	0.911
demographics	0.924	0.925	0.924
qualifier_value::site	0.611	0.478	0.536
device	0.764	0.553	0.641
qualifier_value::course	0.794	0.869	0.830
qualifier_value::severity	0.737	0.660	0.697
organism	0.475	0.385	0.425
substance	0.333	0.167	0.222
qualifier_value::episodicity	0.000	0.000	0.000
qualifier_value::degree	0.000	0.000	0.000

qualifier_value::priority	0.000	0.000	0.000
qualifier_value::intent	0.000	0.000	0.000
physical_object	0.000	0.000	0.000

3.6 Discussion

In this study, we applied the rule-based method, CRF-based method, and LSTM-based method to recognize broad types of clinical entities in discharge summaries.

For dictionary lookup approaches, semantic lexicon files are the key. Simply using lexicons from the SNOMED CT along did not achieve good performance. Combining corpus-specific lexicons with the SNOMED CT lexicons increased recall by 0.182 and F score by 0.086, indicating the importance of extracting terms from corpora of the target domain. Error analysis shows that medication names such as Amoxicillin, Lipitor, etc. are not included in SNOMED CT concepts or descriptions, thus often missed by the baseline method. The recall value increased from 0.357 to 0.687 for “medication” semantic type when medication terms from the training corpus were used. However, it is time-consuming and less practical to generate corpus-specific lexicons, as it requires manual annotation of a large number of clinical documents.

It is a bit surprising that expanding lexicons using the UMLS SPECIALIST Lexicon decreased the overall performance. We did observe an increased recall when more lexical variants were added and used for lookup. However, precision decreased more than the increase of recall, thus making the F-score lower. The decrease of precision is mainly due

to more false positives that were generated by the expanded lexicons. For example, we found the SNOMED CT concept "419652001 | Take - dosing instruction imperative (qualifier value)" when we searched the word "take". After more lexical variants were added, the words "takes", "taken", and "taking" were also mapped to the same SNOMED CT concept and increased false positives.

When compared to the rule-based method, both CRF and LSTM-based approaches achieved much better performance, indicating the potential of machine learning in clinical NER.

3.7 Conclusion

NER for broad types of clinical entities is still challenging. Our study shows that dictionary lookup with heuristic rules is not sufficient to achieve high performance for NER of SNOMED concepts. Machine learning and deep learning-based approaches could significantly improve the performance of the proposed NER task. However, issues such as annotation cost and overfitting should still be considered when developing statistical NER approaches.

Chapter 4: Relation Extraction

4.1 Introduction

Semantic relations between clinical entities such as the treatment relationship between drugs and diseases are critically important information embedded in the clinical text. Therefore, extracting semantic relations between entities from the clinical text is an essential task of clinical information extraction.

Early work of relation extraction (RE) focused on limited linguistic context and relied on word co-occurrences and pattern matching [70–72]. Later, machine learning-based supervised approaches were widely employed. Relation classification models were trained on annotated data. The most important information to be considered for model training is the syntactic or semantic structures of the context surrounding named entities. Generally, the frameworks of supervised learning based relation extraction techniques can be classified into several major categories [73]:

(1) Feature-based methods where a set of features is generated for each relation instance in the labeled data, and a classifier is then trained to classify new relation instances [74].

(2) Tree kernel-based methods where syntactic tree kernel functions are designed to compute similarities between representations of two relation instances. Support Vector Machine (SVM) is usually employed for relation classification with its accommodation of various kernels [73].

(3) Deep learning-based methods where distributional representations (embeddings) of words and dependency-based syntactic structures are used as input features to the algorithms of convolutional neural networks (CNNs) [75] or RNNs [76] for relation classification.

(4) Joint learning of entities and relations. Traditionally, the relation extraction task is completed using a pipeline of two separated tasks: NER and RE. Once entities and their types are identified, then RE techniques can be applied. Such a pipeline method is prone to the propagation of errors from the first phase (extracting entities) to the second phase (extracting relations). To avoid this propagation of errors, joint modeling of entity and relation has become increasingly popular because of their high performance since relations depend highly on entity information [77].

Several clinical NLP challenges have been organized for clinical relation extraction, such as the temporal relation extraction task in SemEval 2016 [78], relations between modifiers and diseases in SemEval 2015 [79], and assertions of diseases, medications and lab tests in the i2b2 2010 [55]. Besides, existing clinical NLP tools also contain relation

recognition modules. For example, CLAMP can recognize assertions, modifiers of diseases and medications [36]. cTAKES can also recognize the negation of named entities [35], as well as temporal modifiers of entities. Machine learning based methods are the current state-of-the-art methods in clinical NLP challenges, especially deep learning based methods [79]. However, most of the information models designed for relation extraction works are relatively simple, focusing on several specific clinical concepts and relation types.

Guided by the information model based on SNOMED CT designed in Chapter 2, this study takes the initiative to build relation extraction systems for a comprehensive set of core clinical concepts and relations. In total, relation extraction systems are built for 19 relations. We investigate the common frameworks of supervised relation extraction, including feature-based, tree-kernel based and joint learning of entities and relations using deep learning based methods for the task here.

4.2 Methods

We have used a feature-based supervised learning approach, a kernel-based supervised learning approach, and a deep learning approach to joint extract entities and relations for our relation extraction task. The feature-based approach was used to set the performance baseline for the state-of-the-art kernel-based approach and deep learning-based joint extraction approach.

4.2.1 Feature-based Approach

SVM is a supervised machine learning technique motivated by the statistical learning theory [80]. SVM seeks an optimal separating hyperplane based on the structural risk minimization. It divides the training examples into two classes and selects the only effective instances in the training set based on support vectors.

SVMs are used to build binary classifiers. Therefore, we must adapt SVMs for multi-class classification. We applied the one vs. others strategy, which builds K classifiers to separate one class from all others. The class that has the maximal SVM output will determine the final decision of an instance in the multiple binary classifications.

A semantic relation is determined between two entities. We define the argument order of the two entity mentions, M1 for the first mention and M2 for the second mention: Relation(M1, M2). An example of relation with ordered arguments is Has_location(“head”, “injury”).

Our feature selection follows the work by Zhou et al. [74]. According to their positions, four categories of words are used as features:

1) The words of both M1 and M2
2) The words between M1 and M2
3) The words before M1
4) The words after M2

The headword is generally much more important. For the words of both mentions, we differentiate the headword of a mention from other words. The words between the two mentions can be classified into three bins: the first word in between, the last word in between and other words in between. Both the words before M1 and after M2 can be classified into two bins: the first word next to the mention and the second word next to the mention. The entity type of both mentions and combination of mention entity types are also used as features.

4.2.2 Graph Kernel-based Approach

The overall performance of feature-based methods largely depends on the effectiveness of the designed features. The main advantage of kernel-based methods is that such explicit feature engineering is avoided. In kernel-based methods, kernel functions are designed to compute similarities between representations of two relation instances and SVM is employed for classification.

Our method follows the all-paths graph kernel proposed by Airola et al. [81–83]. A graph kernel calculates the similarity between two input graphs by comparing the relations between common vertices. The weights of the relations are calculated using all possible paths between each pair of vertices. The kernel represents the target pair using graph matrices based on two sub-graphs. The first sub-graph is built from the dependency analysis and represents the structure of a sentence. It is a directed graph which includes

two types of vertices. One type is a word vertex contains its lemma and part-of-speech tags (POS). Another type is a dependency vertex contains the dependency relation between words. Both types of vertices contain their positions. Their positions differentiate them from other vertices. Figure 4-1 illustrates the dependency graph. Since the words connecting the candidate entities in a syntactic representation are particularly likely to carry information regarding their relationship [84], the labels of the vertexes on the shortest undirected paths connecting word1 and word2 are differentiated from the labels outside the paths using a special tag "IP". Further, the edges are assigned weights; all edges on the shortest paths receive a weight of 0.9 and other edges receive a weight of 0.3 as in [81]. Thus, the shortest path is emphasized while also considering the other words outside the path as potentially relevant. Furthermore, semantic classes, representing the sentence content at a fine-grained semantic level, can be integrated into the dependency graph kernel by replacing the word vertices with semantic class vertices.

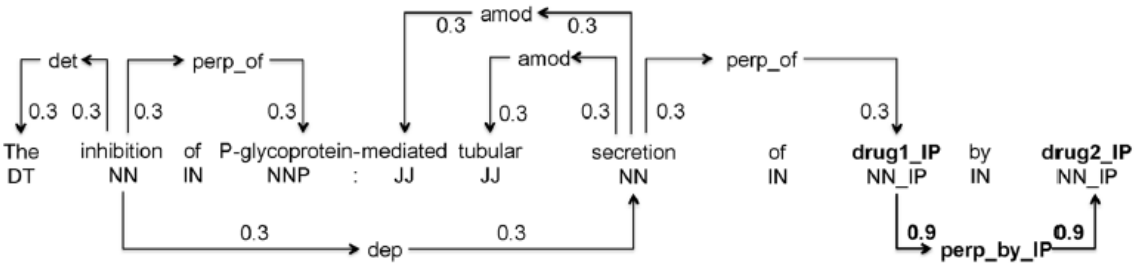

Figure 4-1. Dependency Graph

The second sub-graph is built from the linear structure of the sentence and represents the word sequence in the sentence. Each of its word vertices contains its lemma, its relative

position to the target pair and its POS. All edges are given the weight 0.9 as in [81]. Figure 4-2 illustrates the linear order graph.

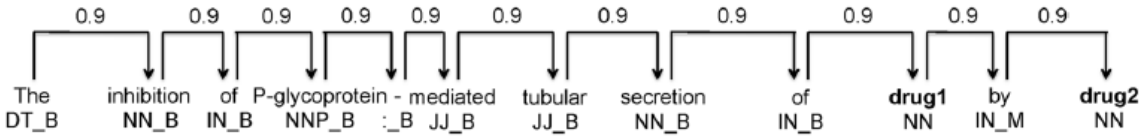

Figure 4-2. Linear Order Graph

Assuming V represents the set of vertices in the graph, the calculation of the similarity between two graphs uses two types of matrices which are edge adjacent matrix A and label matrix L. The graph is represented with the adjacent matrix $A \in R^{|V| \times |V|}$ whose rows and columns are indexed by the vertices, and $[A]_{i,j}$ contains the weight of the edge connecting $v_i \in V$ and $v_j \in V$ if such an edge exists, and 0 otherwise. In addition, the labels are presented as a label allocation matrix $L \in R^{|I| \times |V|}$, so that $L_{i,j} = 1$ if the *j-th* vertex has the *i-th* label, and $L_{i,j} = 0$ otherwise. Using the Neumann Series, a graph matrix G is calculated as:

$$G = L^T \sum_{n=0}^{\infty} A^n L = L^T((I - A)^{-1} - I)L$$

This matrix sums up the weights of all the paths between any pair of vertices. Each entry represents the strength of the relation between a pair of vertices. Given two instances of graph matrices G' and G'', the graph kernel $K(G', G'')$ is defined as follows:

$$K(G', G'') = \sum_{i=1}^{|L|} \sum_{j=1}^{|L|} G'_{ij} G''_{ij}$$

4.2.3 Deep learning-based Joint Learning Approach

Our method follows the end-to-end relation extraction method proposed by Miwa et al. [85]. The recurrent neural network based model captures both word sequence and dependency tree substructure information by stacking bidirectional tree-structured LSTM-RNNs on bidirectional sequential LSTM-RNNs. This allows the model to jointly represent both entities and relations with shared parameters in a single model.

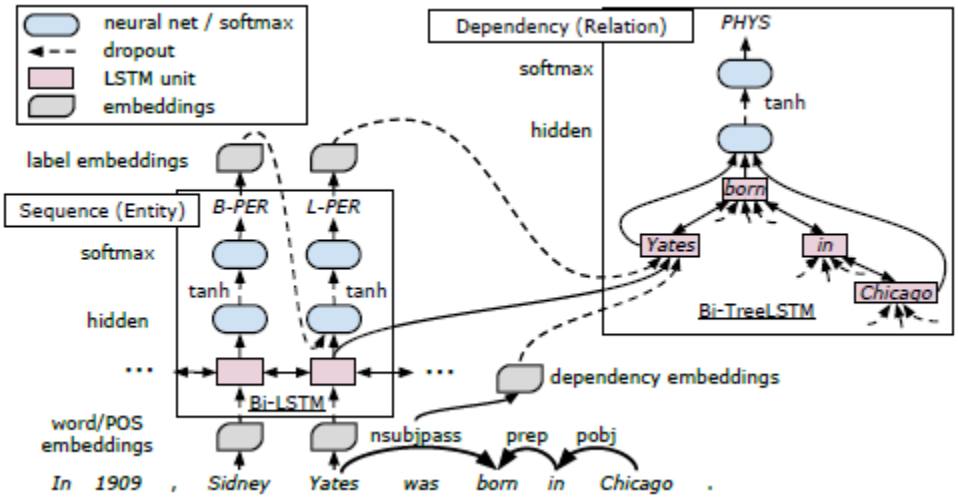

Figure 4-3. End-to-end Relation Extraction Model

Figure 4-3 shows the overview of the model. The model mainly consists of three representation layers: a word embeddings layer (embedding layer), a word sequence based LSTM-RNN layer (sequence layer), and finally a dependency subtree based LSTM-RNN layer (dependency layer). The embedding layer handles embedding representations of words, part-of-speech (POS) tags, dependency types, and entity labels. The sequence layer represents words in a linear sequence using the representations from

the embedding layer. This layer represents sentential context information and maintains entities. The entity detection is treated as a sequence labeling task. The dependency layer represents a relation between a pair of two target words in the dependency tree. It is corresponding to a relation candidate in relation classification.

The left-to-right entity detection is built on the sequence layer and relation classification is realized on the dependency layers, where each subtree based LSTM-RNN corresponds to a relation candidate between two detected entities. The parameters are simultaneously updated via backpropagation through time (BPTT) [86]. The dependency layers are stacked on the sequence layer, so the embedding and sequence layers are shared by both entity detection and relation classification, and the shared parameters are affected by both entity and relation labels.

4.3 Experiments and Evaluation

In our annotated discharge summary corpus, there are 4,289 relation instances for 61 unique relationships. Many relationships have only a few instances. We selected 19 relationships that have more than 40 instances and applied all three approaches including feature-based approach, graph kernel-based approach, and joint learning based approach. Parameters for each algorithm were optimized using the training set via a 10-fold cross-validation method. POS-tags and dependency trees of the datasets were generated using the Stanford parser [87].

For each approach, we evaluated the relation recognition performance for both using gold-standard entities and using the entities recognized from the last chapter (an end-to-end system). We used the standard measures (Precision, Recall, and F-measure) to evaluate the performance of each approach.

The feature-based SVM approach was used to set our performance baseline. The study showed that using only a set of basic features could already achieve reasonable performance and adding more complex features may not improve the performance much [88]. Therefore, we used some basic word and entity type features for the SVM classifier in our experiments.

The package of the graph kernel-based algorithm provided in [81] was employed in our experiments. This package is built on the lease squares SVM and provides configuration options for some SVM parameters, as well as graph kernel related parameters. For graph kernels, all edges on the shortest paths received a weight of 0.9, the other edges received a weight of 0.3. For the word sequence based kernel, all edges received a weight of 0.9.

The package of the deep learning-based joint learning algorithm provided in [85] was employed in our experiments. The package is implemented using the Dynamic Neural Network Toolkit (DyNet) [89]. Sentences were parsed using the Stanford neural dependency parser [90] with the original Stanford Dependencies.

4.4 Results

Table 4-1 shows the performance of relation extraction using annotated gold standard clinical entities. Table 4-2 shows the end-to-end performance by recognizing the clinical entities first and then extracting the relations among recognized entities. Relation(ConceptType1, ConceptType2) defines that Concept Type 1 has a Relation with Concept Type 2. We use the abbreviations (defined in Chapter 2 Table 2-2) for clinical concept semantic types. For example, Has_location(CF, BS) defines that the concept type Clinical Finding (CF) has a relation Has_location with the concept type Body Structure (BS).

We highlighted the best F-measure for each relation in the result tables. When using gold standard entities for relation extraction, joint learning based approach achieved best F-measures for 10 relations, feature-based approach had best F-measures for 9 relations, and graph kernel-based approach did not have any best F-measures. The best F-measure performance was 0.894 for relation Has_modifier(CF, CER) using the feature-based approach. In the end-to-end relation extraction, joint learning based approach had best F-measures for 11 relations, graph kernel-based approach had best F-measures for 7 relations, and feature-based approach only had best F-measures in 1 relation. The best F-measure performance was 0.718 for relation Has_modifier(BS, QV_LA) using joint learning based approach.

Table 4-1

Relation Extraction Performance (Gold Standard Entities)

Relation (ConceptType1, ConceptType2)	No. of Instances	SVM			Graph Kernel			Joint		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Has_location (CF, BS)	871	0.691	0.775	0.731	0.781	0.919	0.844	0.821	0.935	**0.874**
Has_modifier (CF, CER)	471	0.857	0.934	**0.894**	0.761	0.919	0.833	0.754	0.942	0.838
Has_modifier (CF, QV_AT)	434	0.734	0.853	0.789	0.845	0.898	0.871	0.866	0.910	**0.887**
Belongs_to (CF, PER)	425	0.685	0.821	**0.747**	0.519	0.739	0.610	0.540	0.752	0.629
Has_procedure_site (PRO, BS)	298	0.661	0.681	0.671	0.727	0.877	0.795	0.743	0.891	**0.810**
Has_modifier (BS, QV_LA)	260	0.684	0.826	0.748	0.724	0.822	0.770	0.750	0.853	**0.798**
Has_modifier (CF, QV_SE)	194	0.671	0.892	**0.765**	0.639	0.879	0.740	0.640	0.881	0.741
Has_modifier (CF, QV_CO)	165	0.719	0.885	0.793	0.741	0.945	0.830	0.745	0.957	**0.838**
Has_modifier (PRO, QV_AT)	133	0.587	0.835	**0.689**	0.523	0.852	0.648	0.531	0.865	0.658
Has_indication (MED, CF)	119	0.655	0.622	**0.638**	0.275	0.909	0.423	0.294	0.930	0.447
Associated_with (CF, CF)	108	0.333	0.009	0.018	0.181	0.704	0.288	0.201	0.738	**0.316**
Has_modifier (BS, QV_SI)	100	0.705	0.91	**0.795**	0.596	0.831	0.694	0.600	0.841	0.700
Has_focus (PRO, CF)	96	0.429	0.031	**0.058**	0.000	0.000	0.000	0.000	0.000	0.000

Has_interpretation (PRO, CF)	94	0.286	0.021	0.040	0.034	0.600	0.065	0.355	0.620	**0.451**
Has_modifier (CF, QV_LA)	62	0.000	0.000	0.000	0.016	1.000	0.032	0.633	0.724	**0.675**
Has_modifier (CF, QV_SI)	56	0.533	0.143	0.225	0.018	1.000	0.036	0.407	0.210	**0.277**
Procedure_device (PRO, DEV)	52	0.640	0.615	**0.627**	0.314	0.727	0.438	0.322	0.756	0.452
Has_modifier (PRO, QV_SI)	42	0.667	0.571	**0.615**	0.357	0.577	0.441	0.376	0.602	0.463
Has_modifier (PRO, QV_LA)	40	0.429	0.075	0.128	0.300	0.632	0.407	0.318	0.650	**0.427**

Table 4-2

Relation Extraction Performance (End-to-End)

Relation (ConceptType1, ConceptType2)	**No. of Instances**	**SVM**			**Graph Kernel**			**Joint**		
		P	**R**	**F**	**P**	**R**	**F**	**P**	**R**	**F**
Has_location (CF, BS)	871	0.487	0.554	0.519	0.570	0.682	0.621	0.665	0.698	**0.681**
Has_modifier (CF, CER)	471	0.642	0.535	0.584	0.559	0.680	0.613	0.720	0.600	**0.655**
Has_modifier (CF, QV_AT)	434	0.466	0.535	0.498	0.624	0.623	**0.623**	0.580	0.650	0.613
Belongs_to (CF, PER)	425	0.564	0.653	0.605	0.371	0.347	0.359	0.605	0.630	**0.617**
Has_procedure_site (PRO, BS)	298	0.513	0.523	0.518	0.536	0.571	0.553	0.650	0.660	**0.655**
Has_modifier	260	0.577	0.741	0.649	0.508	0.465	0.485	0.680	0.760	**0.718**

(BS, QV_LA)										
Has_modifier (CF, QV_SE)	194	0.433	0.469	0.450	0.479	0.578	0.524	0.610	0.570	**0.589**
Has_modifier (CF, QV_CO)	165	0.582	0.648	0.613	0.545	0.763	0.636	0.670	0.705	**0.687**
Has_modifier (PRO, QV_AT)	133	0.371	0.421	0.394	0.389	0.519	0.445	0.480	0.520	**0.499**
Has_indication (MED, CF)	119	0.536	0.496	0.515	0.189	0.652	0.293	0.680	0.590	**0.632**
Associated_with (CF, CF)	108	0.000	0.000	0.000	0.132	0.308	**0.185**	0.000	0.000	0.000
Has_modifier (BS, QV_SI)	100	0.393	0.590	0.472	0.443	0.480	0.460	0.540	0.690	**0.606**
Has_focus (PRO, CF)	96	0.375	0.031	**0.058**	0.000	0.000	0.000	0.000	0.000	0.000
Has_interpretation (PRO, CF)	94	0.167	0.011	0.020	0.024	0.220	**0.043**	0.020	0.100	0.020
Has_modifier (CF, QV_LA)	62	0.000	0.000	0.000	0.012	1.000	**0.024**	0.000	0.000	0.000
Has_modifier (CF, QV_SI)	56	0.043	0.018	0.025	0.013	1.000	0.026	0.365	0.123	**0.184**
Procedure_device (PRO, DEV)	52	0.400	0.192	0.260	0.231	0.333	**0.273**	0.192	0.400	0.259
Has_modifier (PRO, QV_SI)	42	0.048	0.024	0.032	0.268	0.204	**0.231**	0.030	0.100	0.046
Has_modifier (PRO, QV_LA)	40	0.375	0.075	0.125	0.225	0.243	**0.234**	0.045	0.345	0.080

4.5 Discussion

Even we used some basic word and entity type features only for SVM classifier in our experiments, the results showed that feature-based approach had the best performance in extracting almost half of the relation types than more sophistic graph kernel-based and joint learning based approaches when using gold standard entities. Our graph kernel-based and joint learning based approaches achieved similar performances even though joint learning based approach achieved slightly better performance when using gold standard entities for predicting relations.

It is not surprising that the corpus size plays an important role for performance. When the number of instances for a relation type is greater than 150, most of the relation types achieved high F score ($F > 0.7$). When the number of instances for a relation type is less than 100, the performance greatly decreased ($F < 0.5$). This finding suggests that we should annotate more clinical documents, in order to achieve optimal performance for machine learning-based relation extraction tasks.

In the end-to-end extraction of both entities and relations, all three approaches suffered big performance loss. Joint learning based approach reported better results than the other two approaches. The highest F score for feature-based approach decreased from 0.894 to 0.584; the highest F score for graph kernel-based approach decreased from 0.871 to 0.623; the highest F score for joint learning based approach decreased from 0.887 to 0.613.

These findings indicate it is still challenging to build NLP systems that can extract both entities and relations with high accuracy.

4.6 Conclusion

We used a feature-based SVM approach, a graph kernel-based approach, and a joint learning-based approach to extract a comprehensive set of relation types. All three approaches achieved good performance when the number of instances used for training the algorithms was large enough. Joint learning based approach achieved better performance for the end-to-end system that extracts both entities and relations.

Chapter 5: SNOMED CT Encoding

5.1 Introduction

Entities and relations extracted in previous Chapters have to be encoded into standard concepts in ontologies, in order to be used for other computerized applications [91]. An automated encoding system has to be developed to map entities (often in various surface forms) and relations in clinical documents into standardized representations in an ontology. Standardized clinical codes are then used for hospital billing, clinical audit, epidemiological studies, measuring treatment effectiveness, assessing health trends, cost analysis, health-care planning, and resource allocation [19].

An encoder that maps extracted mentions of entities to concepts in ontologies is also known as the entity linking task in NLP. The entity linking task has been extensively studied in Computer Science including shared tasks such as TAC KBP [92]. Diverse heuristic and machine learning based methods have been proposed for a framework of entity liking that includes candidate generation, candidate ranking, and un-linkable mention prediction but few of them have been investigated in the medical domain. Those widely used NLP systems such as cTAKES and MetaMap are mainly based on dictionary lookup approaches for concept mapping at this time.

Furthermore, as previously illustrated in Table 1-1, most existing NLP systems are mapping entities to the UMLS concepts. Although the UMLS contains comprehensive medical vocabularies, its noisiness and inconsistency also make it less desirable for reliable inference based on hierarchy [47]. Therefore, encoding clinical entities to a single, comprehensive medical ontology that has consistent hierarchy is more appealing, and SNOMED CT is such a good candidate ontology. The study indicates that about 80% of itemized entries for the summary level information in EHRs can be encoded with SMOMED CT normalized phrases (pre-coordinated concepts) [93]. It also allows compositional encoding of clinical concepts with semantic relations between them, so that multiple concepts can be combined to form a more detailed representation of the clinical information (post-coordinated concepts). Compositional expressions allow more complex descriptions and therefore provide a complete representation of medical concepts.

Despite growing interests to incorporate SNOMED CT as a reference terminology into the clinical information systems, there are few detailed encoding instructions and examples available [17]. The existing methodologies for mapping clinical text in EHR to SNOMED CT concepts range from manual to semi-automatic and automatic methods [17,94,95]. In a manual encoding method, the majority of the effort was spent on data cleaning and generating the data items to be encoded. The exact matching algorithm was used for the batch process and the matching results were manually verified. Data items that cannot be encoded using the batch process were searched for manually using a

SNOMED CT browser [17]. The general approach for automated clinical coding is to transform code descriptions and narrative text into an internal representation. Text is matched to codes based on the similarity between the text's and the code's internal representation. Internal representations normalize raw forms and generally capture linguistic information used in matching and scoring. Barrett et al. developed a token-based approach that codes narrative tokens and manipulates token-level encodings by mapping linguistic structures to topological operations [94]. Most of these methods convert text to pre-coordinated SNOMED CT concepts. Studies that have used post-coordination was completed manually [17] or did not include the detailed description of the approach [96].

In this study, we mainly focus on the encoding of "clinical findings" and their relations with modifiers (e.g., body location, negation) in SNOMED CT. Considering that SNOMED CT may not have a full coverage of clinical concepts in practical clinical settings, the encoding is carried out at three levels of granularities: (1) the mentions of the clinical findings; (2) binary relations: the phrase containing a clinical finding concept and one of its modifier; (3) multiple relations: the phrase containing a clinical finding concept and all of its modifiers. To obtain the optimal encoding performance, we propose a novel learning-to-rank based method that incorporates multiple features to capture the similarity between concept mentions and standard terms from different linguistic aspects. Particularly, a translational language model is generated based on synonyms in

SNOMED CT to alleviate the severe problem of surface text discrepancy of clinical expressions.

5.2 Method

As illustrated in Figure 5-1, the encoding process contains three steps: (1) Candidate generation and ranking: Firstly, a search engine is built, in which the terms inside the description file of SNOMED CT are trimmed and indexed. The description file contains all the synonyms of the same semantic concept in SNOMED CT. Given a mention of clinical finding, or a combination of a modifier and the clinical finding as of the query, its top 10 candidate terms are retrieved from the index using the common information retrieval model of BM25 [97]. (2) Candidate re-ranking: After that, the initial set of candidate terms is re-ranked using the learning to ranking [98–101] method; (3) Candidate determination: Finally, the corresponding SNOMED CT concept code of the top-ranked SNOMED CT term is assigned.

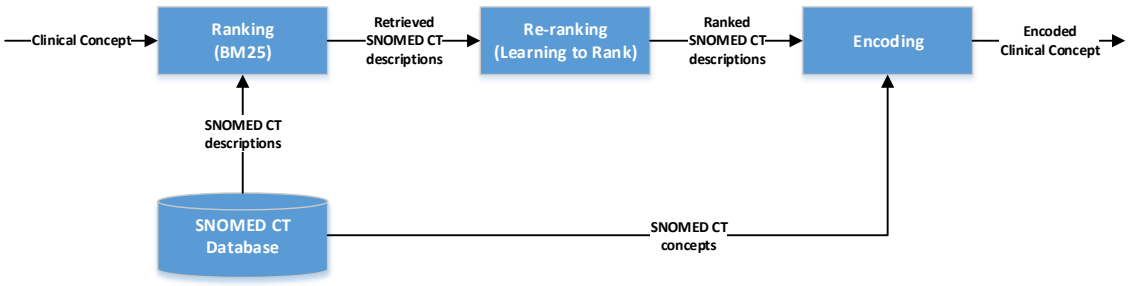

Figure 5-1. System Architecture for Encoding

5.2.1 Gold Standard Annotation for Encoding Evaluation

Gold standard encoding of clinical concepts and relations were created and used for development and evaluations. Our focus is clinical finding related concepts and their relations. There are totally 5,531 concept mentions in our annotation corpus of discharge summaries; 2,916 concept mentions are annotated with the semantic types of "finding" or "disorder", both of which were included in this study as "clinical finding" type.

For relation encoding, first we coded binary relations, which are the relations between one clinical finding and one of its modifier. Next, we coded complex relations, which are the most granular SNOMED CT concept codes for combining the clinical finding and all of its modifiers. For example in Figure 5-2, clinical finding "injury" has two binary relations: a "Has location" relation with body structure "head" and a "Has modifier" relation with qualifier value "closed". Our final results will have three sets of SNOMED CT codes:

- Clinical concepts: "417746004 | Traumatic injury (disorder)", "29179001 | Closed (qualifier value)", "69536005 | Head structure (body structure)"
- Concepts contain binary relation: "264513002 | Closed injury (qualifier value)", "82271004 | Injury of head (disorder)"
- Concepts contain complex relations: "451000119106 | Closed injury of head (disorder)"

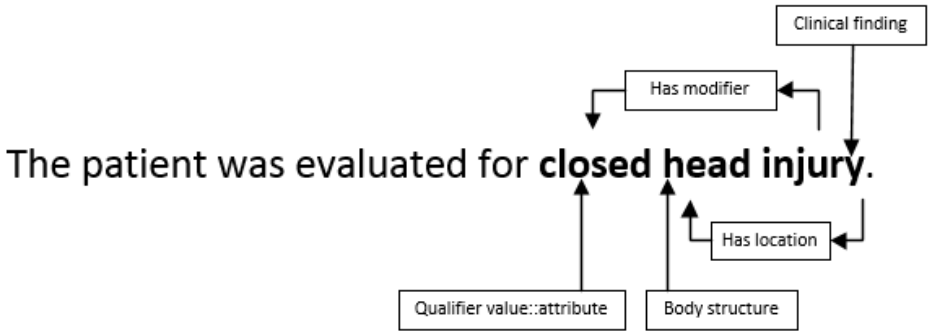

Figure 5-2. Concepts and relations encoding

The annotation of gold standard concepts for encoding is a semi-automatic process: First, we applied a pooling process to find the candidate list of SNOMED CT terms by combining the candidates from five different sources: firstly, the BM25 algorithm was used as an information retrieval model to match and rank SNOMED CT terms based on lexicon similarity and distribution; in addition, the encoding modules in three clinical NLP software, CLAMP, cTAKES and MetaMap were used to map a clinical concept to UMLS CUI. The UMLS CUI is then mapped to a SNOMED CT concept using UMLS's mapping file MRCONSO.RRF. Furthermore, the UMLS API [102] was also applied to retrieve UMLS CUIs which are mapped to SNOMED CT concepts.

Next, a physician manually reviewed the candidate concepts in the pool and assigned the correct SNOMED CT codes. On average, each clinical concept mention had 14.18 candidates after pooling. Using automatically identified candidates greatly reduced the amount of work for manually searching SNOMED CT and assigning codes. Correct

candidates were selected and labeled as the gold standard. If none of the candidates were correct, the SNOMED CT codes would be manually searched and assigned. Some concepts and relations cannot be located in SNOMED CT codes and we assigned "Nil" as the code. Table 5-1 shows the number of SNOMED CT codes in the gold standard data. In our data set, there are 2,916 clinical concepts with clinical finding semantic type, 3,501 binary relations and 2,916 complex relations for these concepts. Table 5-2 shows some examples of the gold standard data.

Table 5-1

Gold Standard Data for Encoding

	Has SNOMED CT code	**No SNOMED CT code**	**Total**
Clinical concept	2,746 (94.17%)	170 (5.83%)	2,916
Concept contain binary relation	1,344 (38.39%)	2,157 (61.61%)	3,501
Concept contain complex relation	786 (26.95%)	2,130 (73.05%)	2,916

Table 5-2

Gold Standard Data Examples

Mention	**Concept Type**	**SNOMED CT Code**	**SNOMED CT Concept**
cancer	Clinical concept	363346000	Malignant neoplastic disease

			(disorder)
diarrhea	Clinical concept	62315008	Diarrhea (finding)
mild diarrhea	Concept contain binary relation	Nil	
abnormalities	Clinical concept	Nil	
congenital abnormalities	Concept contain binary relation	276654001	Congenital malformation (disorder)
congenital genitourinary abnormalities	Concept contain complex relation	287085006	Genitourinary congenital anomalies (disorder)
fevers	Clinical concept	386661006	Fever (finding)
persistent fevers	Concept contain binary relation	271751000	Continuous fever (finding)
persistent high fevers	Concept contain complex relation	Nil	

5.2.2 Models of Learning to Rank

Training Dataset

The training dataset for the learning to rank model was constructed from the top 10 BM25 matching results that we generated from the previous pooling process.

Algorithm

We used linear Ranking SVM [103], a state-of-art method of learning to rank, to train the ranking model.

The Ranking SVM algorithm is a learning retrieval function that employs pair-wise ranking methods to adaptively sort results based on how relevant they are for a specific query. From the gold standard data, we derive pairwise preference data (m, c) such that score(m, c^+) > score(m, c^-), where m is the clinical concept mention and c is the SNOMED CT term candidate. Specifically, (m, c^+) are selected from the instances c labeled as positive with respect to m, while (m, c^-) are selected from the instances labeled as negative.

The Ranking SVM function uses a mapping function to describe the match between a clinical concept mention and the features of each of the possible SNOMED CT term candidates. This mapping function projects each mention and candidate data pair onto a feature space φ. These features of the labeled data are then used to train an automatic ranking system. As illustrated in the following equation, the final score $score(m, c)$ of each pair (m, c) are a linear interpolation of the feature functions $\varphi(m, c)$, multiplied by their weights w_i.

$$score(m, c) = \sum w_i \varphi_i(m, c)$$

Ranking Features

Three basic matching models are first implemented as the baseline features. Then, a translation-based language model (TransLM) is included for alleviating the lexical gap problem.

BM25 Score

Given a concept mention *M*, containing words m_1, …, m_n, the BM25 score of a SNOMED CT term *T* is:

$$score(T, M) = \sum_{i=1}^{n} IDF(m_i) \times \frac{f(m_i, T) \times (k_1 + 1)}{f(m_i, T) + k_1 \times (1 - b + b \times \frac{|T|}{avgtl}}$$

where $f(m_i, T)$ is m_i's term frequency in the term *T*, |*T*| is the length of the term *T* in words, and *avgtl* is the average term length of all SNOMED CT terms. k_1 and *b* are free parameters, usually chosen, in the absence of an advanced optimization, as $k_1 \in [1.2, 2.0]$ and $b = 0.75$. $IDF(m_i)$ is the IDF (inverse document frequency) weight of the mention word m_i:

$$IDF(m_i) = log \frac{N - n(m_i) + 0.5}{n(m_i) + 0.5}$$

where *N* is the total number of SNOMED CT terms, and $n(m_i)$ is the number of SNOMED CT terms containing m_i.

Exact Match

From the pairwise data (*m*, *c*) in the ranking algorithm, where *m* is the clinical concept mention and *c* is the SNOMED CT term candidate, we built a feature set based on whether *m* and *c* are exact matches. We also built a feature set based on whether normalized *m* and normalized *c* are exact matches. The normalization process involves changing the term to lower case, removing punctuation and prefixes, as well as stemming.

Jaccard Similarity Score

The Jaccard similarity score measures string similarity between a concept mention *M* and a SNOMED CT term *T*, and is defined as the size of the intersection words in *M* and *T* divided by the size of the union words in *M* and *T*:

$$J(M,T) = \frac{|M \cap T|}{|M \cup T|} = \frac{|M \cap T|}{|M| + |T| - |M \cap T|}$$

For example, the Jaccard similarity score for a mention "closed head injury" and a SNOMED CT term "closed injury of head" is

$$J(M,T) = \frac{|M \cap T|}{|M \cup T|} = \frac{|\{\text{closed", "head", "injury"}\}|}{|\{\text{closed", "head", "injury", "of"}\}|} = \frac{3}{4} = 0.75$$

The lexical mismatch is common in the usage of natural languages. It occurs when different people name the same thing or concept differently. The lexical mismatch between clinical concept mentions and SNOMED CT terms causes the mismatch problem in our encoding process. For example, the correct SNOMED CT code for the mention "cancer" is "363346000 | Malignant neoplastic disease (disorder)", while the word "cancer" is not a part of the fully specified concept name "Malignant neoplastic disease" in SNOMED CT.

Translation-based Language Model

To alleviate the word mismatch problem, we employ the state-of-art translation-based language model (TransLM) [104]. Given a query (mention) *q* and a document (concept) *d*, the ranking function based on TransLM is written as

$$P(q|d) \propto \sum_{w \in V} c(w,q) log\, P(w|d)$$

$$P(w|d) = (1-\alpha)\sum_{t\epsilon d} P(w|t)P(t|d) + \alpha P(w|C)$$

where $P(w|d)$ and $P(w|C)$ are the unigram language models (LM), which are estimated with the maximum likelihood for the concept *d* and the whole collection *C*, respectively. $P(w|t)$ is the probability of translating a word *t* in concept *d* into a word *w* in mention *q*. It bridges the gap between different words.

The performance of the translation-based language model relies on the quality of the word-to-word translation probabilities. We followed the method of Xue et al. [104] and used GIZA++ toolkit [105,106] to learn the word translation probabilities. To train the translation-based language model, two types of data were used to construct the parallel corpus:

(1) The synonyms from SNOMED CT descriptions. For example, "Cancer" has five synonyms "CA - Cancer", "Malignant neoplasm", "Malignant neoplastic disease", "Malignant tumor", and "Malignant tumour". We pair these synonyms to get the collection ("Cancer", "CA - Cancer"), ("Cancer", "Malignant neoplasm"), …, ("Malignant tumour", "Malignant neoplastic disease"), ("Malignant tumour", "Malignant tumor").

(2) The gold standard in the training data. For example, mention "cancer" is mapped to SNOMED CT concept id "363346000" in the gold standard. We pair the mention "cancer" with all the SNOMED CT terms which have concept id "363346000" to get the collection ("cancer", "Cancer"), ("cancer", "CA - Cancer"), …, ("cancer", "Malignant tumor"), ("cancer", "Malignant tumour").

5.3 Experiments and Evaluation

Baselines for Encoding

We evaluated the baseline performance from the pooling results of the five different approaches in the last section. The top candidate of each approach was used in the evaluation.

Evaluation Criteria

We measured the performance of different retrieval methods using the following metrics:

$$Accuracy(ACC) = \frac{TruePositives(TP) + TrueNegatives(TN)}{TotalPopulation}$$

where the terms True Positives, True Negatives, False Positives, and False Negatives are used to compare the results of the encoding system under test with trusted gold standard data. The terms positive and negative refer to the encoding system's prediction result, and the terms true and false refer to whether that prediction corresponds to the trusted gold standard data.

5.4 Results

Table 5-3 shows the performance of each baseline encoding approach: BM25, CLAMP, cTakes, MetaMap, and UMLS API. Our proposed approaches: Learning to Rank and Learning to Rank with translation-based language model (TransLM) were also reported.

The best baseline performance was achieved by BM25 and its accuracy value for encoding clinical concepts, binary relations, and complex relations were, 75.0%, 60.0%, and 67.3% respectively. Our Learning to Rank approach improved the accuracy in all three categories by 6.1%, 6.2%, and 6.2% respectively. After applied the translation-based language model (TransLM), the accuracy was further improved by 1.3%, 4.4%, and 3.9% respectively. Learning to Rank with translation-based language model (TransLM) achieved the best accuracy value in all three categories: 82.4% for encoding the clinical concepts, 70.6% for encoding the concepts which contain the binary relations, and 77.4% for encoding the concepts which contain complex relations.

Table 5-3

SNOMED CT Encoding Performance (Accuracy)

	Clinical Concept (%)	Concept contain binary relation (%)	Concept contain complex relation (%)
Learning to Rank	81.1	66.2	73.5
Learning to Rank (TransLM)	**82.4**	**70.6**	**77.4**
BM25	75.0	60.0	67.3
CLAMP	52.3	47.0	49.4
cTakes	40.6	33.1	35.9
MetaMap	50.0	44.7	46.5
UMLS API	43.1	34.4	37.5

5.5 Discussion

In the five baseline encoding approach, BM25 reached the best performance in all three encoding categories. CLAMP also uses the BM25 algorithm. However, similar to other clinical NLP systems (cTAKES, MeataMap, and the UMLS API), it did not perform well, probably because all these systems' search space is bigger (the entire UMLS rather than the SNOMED CT terms only). This finding indicates the importance of candidate generation by limiting the search space.

Our Learning to Rank approach added features other than the BM25 score. Experiments show that we were able to achieve much better accuracy value by taking other similarity measures into account. The performance gain from applying the translation-based language model was not trivial as well, indicating the potential of this approach.

Previous automated encoding studies [34,107–109] focus on mapping narrative phrases to terminological descriptions. These methods make little or no use of the additional semantic information available through ontology. Our approach exploited additional semantic information available in SNOMED CT and encoded clinical concepts as well as their relations.

It is possible to represent the same information in multiple ways while using standard terminologies and information models. The same information can be represented using one or several concepts. In other words, the coding of concepts can be achieved by using

pre-coordination or post-coordination [110]. These methodologies have both advantages and disadvantages [111]. Studies have concluded that pre-coordination is easier and ensures consistency [110]. For post-coordination, rules must exist for the consistent use of SNOMED CT. Moreover, transforming SNOMED CT concepts into normal forms can achieve consistency and support selective retrieval [111]. The SNOMED CT implementation guide is limited. It suggests that each hierarchy has a particular purpose [112]. However, the study found overlaps between ‘‘clinical finding” and ‘‘morphologic abnormality” hierarchies [17]. As a result, the encoding by using post-coordination has many problems. There is no complete and uniform methodology for achieving it. SNOMED CT pre-coordination has been proved sufficient for coding clinical data, and local concepts can extend its coverage [113]. Therefore, our study only used pre-coordination for encoding.

5.6 Conclusion

We annotated clinical concepts, binary relations, and complex relations by manually assigning the corresponding SNOMED CT codes. Using the annotated data, we developed new SNOMED CT encoding approaches using Learning to Rank with traditional BM25 model and translation-based language model. We compared the performance of our approaches with other clinical NLP systems and demonstrated the superior performance of our approach on encoding clinical concepts as well as their relations into SNOMED CT concepts.

Chapter 6: Conclusions

6.1 Summary of Key Findings

Extracting important clinical entities and relations embedded in unstructured clinical narratives and encoding them with standard medical ontologies is vital to enable the secondary use of EHRs. In this study, we developed a fine granular information model based on the SNOMED CT ontology. Based on this information model, we developed state-of-the-art approaches to recognize the clinical entities and relations, which were then mapped to SNOMED CT concepts.

The work of our study in each chapter are summarized as follows:

In Chapter 1, we did a survey of current applications, common tasks, tools and systems of clinical NLP, which indicate the importance of information extraction and encoding in the medical domain. Although medical knowledge is available in comprehensive ontologies such as SNOMED CT, they have not been leveraged to guide the development of clinical information extraction and encoding systems. Therefore, we proposed to design an information model based on the SNOMED CT, and build clinical NLP systems following the SNOMED-based information model.

In Chapter 2, we designed a fine granular information model based on SNOMED CT, for flexible encoding of clinical concepts with different granularities. The most important clinical concepts in SNOMED CT such as clinical findings and procedures and their relations were included in the information model. Following an annotation guideline, a corpus of discharge summaries was annotated using the information model, which serves as the basis for developing ontology information extraction systems using SNOMED CT.

In Chapter 3, we investigated dictionary-based, conventional machine learning-based, and deep learning based methods for clinical entity recognition. In the dictionary lookup method, both SNOMED CT lexicons and corpus specific lexicons were used for comparing the performance. Our machine learning-based CRF method and LSTM-CRF method achieved better performance than the dictionary-based method. The evaluation demonstrated that the performances of recognizing important clinical entities are promising for practical applications.

In Chapter 4, we investigated a feature-based approach, a graph kernel-based approach, and a joint learning based approach for the task of clinical relation extraction. The performances were evaluated by using the gold-standard entity mentions as well as automatically recognized entity mentions (i.e., the end-to-end system). Experimental results demonstrated that the joint learning based method outperformed the other two methods on the end-to-end performance, indicating that this method can reduce the errors propagated from the entity recognition step.

In Chapter 5, we first built a gold-standard corpus for SNOMED CT encoding, by annotating clinical finding concepts with different granularities. Next, we investigated Learning to Rank based algorithms for automatic encoding, with traditional IR model of BM25 and translation-based language model. We compared the performance of our approaches with five other encoding systems such as MetaMap and cTAKES. Experimental results demonstrated that our proposed new methods were able to achieve higher performance on encoding clinical concepts as well as their relations.

6.2 Innovations and Contributions

To the best of our knowledge, this is one of the first studies to recognize a comprehensive set of clinical concepts and their relations guided by the SNOMED CT ontology.

In this study, we designed a fine granular information model based on the SNOMED CT ontology, and built an annotation corpus of discharge summary notes with clinical concepts, relations and encoding based on the information model. The information model and gold-standard corpus can be reused in other related clinical applications.

We systematically implemented and compared different approaches for clinical entity recognition and relation extraction, ranging from basic dictionary-based methods to more cutting-edge deep learning based methods. Moreover, a novel Learning to Rank based approach was proposed with multiple features for encoding clinical finding entities to

SNOMED concepts. With the feature obtained from a translation-based language model of synonym pairs, our approach significantly outperformed other existing encoding systems, demonstrating the novelty of this approach.

Overall, we built a state-of-the-art NLP system, guided by the SNOMED CT ontology, to process the clinical text and map them to standard concepts in SNOMED CT. The output information includes a comprehensive set of important clinical entities, relations and standard concept codes mapped to SNOMED CT.

6.3 Future Work

Due to the rich set of clinical entities and relations in the information model, it is very time consuming and labor intensive to annotate a clinical dataset with a high inter-annotator agreement. Currently, our domain experts successfully annotated 100 discharge summary notes. Some less frequent concepts especially modifiers do not have instances sufficient enough for the NLP system to recognize automatically. We will annotate more clinical notes and explore semi-automatic methods such as pre-annotation to enhance the annotation efficiency in the next step.

The fine granular information model could be further refined and expanded. It needs to be adapted to different clinical settings. The NLP pipeline system also needs to be tested using real clinical data from different domains or institutions. Moreover, in addition to learn clinical entities and their relations jointly, we could further investigate the system

performance by jointly learning all the three tasks, entity recognition, relation extraction and the encoding in a single framework.

6.4 Conclusion

In this dissertation research, we took the initiative to develop a fine granular information model based on the SNOMED CT ontology and used it to guide our information extraction process for clinical entities and their relations. We built an NLP system that can recognize a comprehensive set of clinical entities and relations, and finally map them to standardized SNOMED CT codes, which would benefit many clinical applications that rely on the SNOMED CT ontology.

References

1 Meystre SM, Savova GK, Kipper-Schuler KC, *et al.* Extracting Information from Textual Documents in the Electronic Health Record: A Review of Recent Research. *IMIA Yearb Med Inform* 2008;:128–44.

2 Doan S, Conway M, Phuong TM. Natural Language Processing in Biomedicine: A Unified System Architecture Overview. In: *Clinical Bioinformatics*. 2014. 275–94.

3 Jensen PB, Jensen LJ, Brunak S. Mining electronic health records: towards better research applications and clinical care. *Nat Rev Genet* 2012;**13**:395–405.

4 Friedman C, Hripcsak G. Natural language processing and its future in medicine. *Acad Med* 1999;**74**:890–895.

5 Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. *J Am Med Inf Assoc* 2011;**18**:544–51.

6 Jurafsky D, Martin JH. *Speech and Language Processing*. 2nd ed. Prentice-Hall 2008.

7 Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001;:17–21.

8 Zou Q, Chu WW, Morioka C, *et al.* IndexFinder: A Method of Extracting Key Concepts from Clinical Texts for Indexing. *AMIA Annu Symp Proc* 2003;:763–7.

9 Rindflesch TC, Aronson AR. Ambiguity resolution while mapping free text to the UMLS Metathesaurus. *Proc Annu Symp Comput Appl Med Care* 1994;:240–4.

10 Weeber M, Mork JG, Aronson AR. Developing a test collection for biomedical word sense disambiguation. *Proc AMIA Symp* 2001;:746–50.

11 Tao C, Solbrig HR, Sharma DK. Time-Oriented Question Answering from Clinical Narratives Using Semantic-Web Techniques. In: *The Semantic Web – ISWC 2010*. 2010. 241–56.

12 Hripcsak G, Elhadad N, Chen Y, *et al.* Using Empiric Semantic Correlation to Interpret Temporal Assertions in Clinical Texts. *J Am Med Inform Assoc* 2009;**16**:220–7.

13 Chapman WW, Bridewell W, Hanbury P, *et al.* A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. *J Biomed Inform* 2001;**34**:301–10.

14 Mutalik PG, Deshpande A, Nadkarni PM. Use of General-purpose Negation Detection to Augment Concept Indexing of Medical Documents: A Quantitative Study Using the UMLS. *J Am Med Inform Assoc* 2001;**8**:598–609.

15 Huang Y, Lowe HJ. A Novel Hybrid Approach to Automated Negation Detection in Clinical Radiology Reports. *J Am Med Inform Assoc* 2007;**14**:304–11.

16 Chiang MF, Hwang JC, Yu AC. Reliability of SNOMED-CT Coding by Three Physicians using Two Terminology Browsers. *AMIA Annu Symp Proc* 2006;:131–5.

17 Lee DH, Lau FY, Quan H. A method for encoding clinical datasets with SNOMED CT. *BMC Med Inform Decis Mak* 2010;:53.

18 El-Sappagh S, Elmogy M. An encoding methodology for medical knowledge using SNOMED CT ontology. *J King Saud Univ - Comput Inf Sci* 2016;**28**:311–29.

19 Tatham A. The increasing importance of clinical coding. *Br J Hosp Med* 2008;**69**:372–3.

20 Friedman C, Elhadad N. *Biomedical Informatics*. 2013.

21 Chapman WC, Dowling JN, Wagner MM. Fever detection from free-text clinical records for biosurveillance. *J Biomed Inform* 2004;**37**:120–7.

22 Hripcsak G, Soulakis ND, Li L, *et al.* Syndromic surveillance using ambulatory electronic health records. *J Am Med Inform Assoc* 2009;**16**:354–61.

23 Maroto M, Reshef R, Munsterberg AE, *et al.* Ectopic Pax-3 activates MyoD and Myf-5 expression in embryonic mesoderm and neural tissue. *Cell* 1997;**89**:139–48.

24 Wang X, Hripcsak G, Markatou M, *et al.* Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: A feasibility study. *J Am Med Inform Assoc* 2009;**16**:328–37.

25 Elhadad N, Kan MY, Klavans JL, *et al.* Customization in a unified framework for summarizing medical literature. *Artif Intell Med* 2005;**33**:179–98.

26 Zhang H, Fiszman M, Shin D, *et al.* Degree centrality for semantic abstraction summarization of therapeutic studies. *J Biomed Inform* 2011;**44**:830–8.

27 Demner-Fushman D, Lin J. Answering clinical questions with knowledge-based and statistical techniques. *Comput Linguist* 2007;**33**:63–103.

28 Sager N, Friedman C, Lyman MS. *Medical Language Processing: Computer Management of Narrative Data*. Addison-Wesley 1987.

29 Sager N, Lyman M, Bucknall C, *et al.* Natural Language Processing and the Representation of Clinical Data. *J Am Med Inform Assoc* 1994;**1**:142–60.

30 Friedman C, Alderson PO, Austin J, *et al.* A General Natural-language Text Processor for Clinical Radiology. *J Am Med Inform Assoc* 1994;**1**:161–74.

31 Friedman C, Cimino JJ, Johnson SB. A Schema for Representing Medical Language Applied to Clinical Radiology. *J Am Med Inform Assoc* 1994;**1**:233–48.

32 Haug P, Koehler S, Lau LM, *et al.* A natural language understanding system combining syntactic and semantic techniques. *Proc Annu Symp Comput Appl Med Care* 1994;:247–51.

33 Christensen LM, Haug PJ, Fiszman M. MPLUS: a probabilistic medical language understanding system. In: *Proceeding BioMed '02 Proceedings of the ACL-02 workshop on Natural language processing in the biomedical domain*. Phildadelphia, Pennsylvania: 2002. 29–36.

34 Aronson AR, Lang F. An overview of MetaMap: historical perspective and recent advances. *J Am Med Inform Assoc* 2010;**17**:229–36.

35 Savova GK, Masanz JJ, Ogren PV, *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;**17**:507–13.

36 Soysal E, Wang J, Jiang M. CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018;**25**:331–6.

37 Gruber TR. Toward principles for the design of ontologies used for knowledge sharing. *Int J Hum-Comput Stud* 1995;**43**:907–28.

38 Saripalle RK. Current status of ontologies in Biomedical and Clinical informatics. *Int J Sci Inf* 2010.

39 Lindberg D, Humphreys B, McCray AT. The Unified Medical Language System. *IMIA Yearb* 1993;:41–51.

40 Cote RA, Robboy S. Progress in Medical Information Management Systematized Nomenclature of Medicine (SNOMED). *JAMA* 1980;**243**:756–62.

41 McCray AT, Srinivasan S, Browne AC. Lexical Methods for Managing Variation in Biomedical Terminologies. *Proc Annu Symp Comput Appl Med Care* 1994;:235–9.

42 IHTSDO. *SNOMED CT Starter Guide*. International Health Terminology Standards Development Organisation 2017.

43 Wimalasuriya DC, Dou D. Ontology-Based Information Extraction: An Introduction and a Survey of Current Approaches. *J Inf Sci* 2010;**36**:306–23.

44 Cunningham H. GATE, A General Architecture for Text Engineering. *Comput Humanit* 2002;**36**:223–54.

45 Müller H, Kenny EE, Sternberg PW. Textpresso: An Ontology-Based Information Retrieval and Extraction System for Biological Literature. *PLoS Biol* 2004;**2**.

46 Hina S, Atwell E, Johnson O, *et al.* Extracting the concepts in Clinical Documents using SNOMED-CT and GATE. *Fourth I2b2VA Shar-Task Workshop Chall Nat Lang Process Clin Data* 2010.

47 Campbell JR, Carpenter P, Sneiderman C, *et al.* Phase II Evaluation of Clinical Coding Schemes: Completeness, Taxonomy, Mapping, Definitions, and Clarity. *J Am Med Inform Assoc* 1997;**4**:238–51.

48 Xu H, Stenner SP, Doan S, *et al.* MedEx: a medication information extraction system for clinical narratives. *J Am Med Inform Assoc* 2010;**17**:19–24.

49 Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc* 2013;**20**:806–13.

50 Friedman C, Kra P, Rzhetsky A. Two biomedical sublanguages: a description based on the theories of Zellig Harris. *J Biomed Inform* 2002;**35**:222–35.

51 Transcribed Medical Transcription Sample Reports and Examples. http://mtsamples.com/ (accessed 1 Dec 2017).

52 Clinical Language Annotation, Modelling and Processing Toolkit (CLAMP). https://sbmi.uth.edu/ccb/resources/clamp.htm (accessed 1 Dec 2017).

53 Fleiss' kappa. https://en.wikipedia.org/wiki/Fleiss%27_kappa (accessed 1 Dec 2017).

54 Long W. Extracting Diagnoses from Discharge Summaries. *AMIA Annu Symp Proc* 2005;:470–4.

55 Uzuner O, South BR, Shen S, *et al.* 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inf Assoc* 2011;**18**:552–6.

56 Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: *Proceedings of the 18th International Conference on Machine Learning*. 2001. 282–9.

57 Joachims T. Making large-scale SVM learning practical. 1998.

58 Pathak P, Goswami R, Joshi G, *et al.* CRF-based Clinical Named Entity Recognition using clinical NLP Features. 2013.

59 Habibi M, Weber L, Neves M, *et al.* Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics* 2017;**33**:37–48.

60 Wu Y, Xu J, Jiang M, *et al.* A study of neural word embeddings for named entity recognition in clinical text. *AMIA Annu Symp Proc* 2015;:1326–1333.

61 Lample G, Ballesteros M, Subramanian S, *et al.* Neural Architectures for Named Entity Recognition. *Proc NAACL-HLT 2016* 2016;:260–70.

62 Mikolov T, Sutskever I, Chen K, *et al.* Distributed Representations of Words and Phrases and their Compositionality. *Proc NIPS* 2013.

63 National Library of Medicine. SPECIALIST Lexicon and Lexical Tools. In: *UMLS Reference Manual*. 2009.

64 Aho A, Ullman J. Chapter 10. Patterns, Automata, and Regular Expressions. In: *Foundations of Computer Science*. 1992. 529–90.

65 SNOMED International. SNOMED CT Document Library - Technical Resources. 2017.https://confluence.ihtsdotools.org/display/DOC/Technical+Resources (accessed 1 Mar 2018).

66 Wallach HM. Conditional Random Fields: An Introduction. Department of Computer & Information Science, University of Pennsylvania 2004.

67 Bengio Y, Simard P, Frasconi P. Learning long-term dependencies with gradient descent is difficult. *Neural Netw IEEE Trans On* 1994;**5**:157–66.

68 Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80.

69 Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM networks. *Neural Netw 2005 IJCNN 05 Proc 2005 IEEE Int Jt Conf On* 2005.

70 Ng SK, Wong M. Toward routine automatic pathway discovery from on-line scientific text abstracts. *Proc 12th Natl Conf Artif Intell* 1999.

71 Yu H, Hatzivassiloglou V, Friedman C. Automatic extraction of gene and protein synonyms from MEDLINE and journal articles. *Proc AMIA Symp* 2002;:919–923.

72 Huang M, Zhu X, Hao Y. Discovering patterns to extract protein-protein interactions from full texts. *Bioinformatics* 2004;**20**:3604–12.

73 Pawar S, Palshikar GK, Bhattacharyya P. Relation Extraction : A Survey. 2017.

74 Zhou G, Su J, Zhang J, *et al.* Exploring Various Knowledge in Relation Extraction. *Proc 43rd Annu Meet ACL* 2005;:427–34.

75 Convolutional neural network. https://en.wikipedia.org/wiki/Convolutional_neural_network (accessed 1 May 2018).

76 Recurrent neural network. https://en.wikipedia.org/wiki/Recurrent_neural_network (accessed 1 May 2018).

77 Li Q, Ji H. Incremental Joint Extraction of Entity Mentions and Relations. *Proc 52nd Annu Meet Assoc Comput Linguist* 2014;:402–12.

78 Bethard S, Savova G, Chen W, *et al.* SemEval-2016 Task 12: Clinical TempEva. *Proc SemEval-2016* 2016;:1052–1062.

79 Elhadad N, Pradhan S, Gorman SL, *et al.* SemEval-2015 Task 14: Analysis of Clinical Text. *Proc 9th Int Workshop Semantic Eval SemEval 2015* 2015;:303–10.

80 Vapnik V. *Statistical Learning Theory*. 1998.

81 Airola A, Pyysalo S, Björne J, *et al.* All-paths graph kernel for protein-protein interaction extraction with evaluation of cross-corpus learning. *BMC Bioinformatics* 2008.

82 Airola A, Pyysalo S, Björne J, *et al.* A Graph Kernel for Protein-Protein Interaction Extraction. *Curr Trends Biomed Nat Lang Process* 2008;:1–9.

83 Zhang Y, Wu H, Xu J, *et al.* Leveraging syntactic and semantic graph kernels to extract pharmacokinetic drug drug interactions from biomedical literature. *Int Conf Intell Biol Med* 2015.

84 Bunescu RC, Mooney RJ. A shortest path dependency kernel for relation extraction. *Proc HLTEMNLP'05* 2005;:724–31.

85 Miwa M, Bansal M. End-to-End Relation Extraction using LSTMs on Sequences and Tree Structures. *Proc 54th Annu Meet Assoc Comput Linguist* 2016;:1105–16.

86 Werbos PJ. Backpropagation through time: what it does and how to do it. *Proc IEEE* 1990;**78**:1550–60.

87 De Marneffe M, MacCartney B, Manning CD. Generating Typed Dependency Parses from Phrase Structure Parses. *Proc LREC* 2006;:449–54.

88 Jiang J, Zhai C. A systematic exploration of the feature space for relation extraction. *Proc NAACL HLT 2007* 2007;:113–20.

89 Carnegie Mellon University. DyNet: The Dynamic Neural Network Toolkit. https://github.com/clab/dynet (accessed 1 May 2018).

90 Chen D, Manning CD. A Fast and Accurate Dependency Parser using Neural Networks. *Proc 2014 Conf Empir Methods Nat Lang Process* 2014;:740–50.

91 Friedman C, Shagina L, Lussier Y, *et al.* Automated Encoding of Clinical Documents Based on Natural Language Processing. *J Am Med Inform Assoc* 2004;**11**:392–402.

92 Ji H, Grishman R, Dang HT. Overview of the TAC2011 Knowledge Base Population Track. *Proc 2010 Text Anal Conf* 2010.

93 Liu H, Wagholikar K, Wu ST. Using SNOMED-CT to encode summary level data – a corpus analysis. *AMIA Jt Summits Transl Sci Proc* 2012;:30–7.

94 Barrett N, Weber-Jahnke J, Thai V. Automated clinical coding using semantic atoms and topology. *Comput-Based Med Syst* 2012;:1–6.

95 Lamy J, Tsopra R, Venot A, *et al.* A semi-automatic semantic method for mapping SNOMED CT concepts to VCM Icons. *Stud Health Technol Inf* 2013;**192**:42–6.

96 Silva T, MacDonald D, Paterson G, *et al.* Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. *Comput Methods Programs Biomed* 2011;**101**:324–9.

97 Robertson SE, Walker S, Jones S, *et al.* Okapi at TREC-3. *Proc Third Text Retr Conf* 1995.

98 Liu TY. Learning to Rank for Information Retrieval. *Found Trends Inf Retr* 2009;**3**:225–331.

99 Li H. Learning to Rank for Information Retrieval and Natural Language Processing. *Synth Lect Hum Lang Technol* 2011;**4**:1–113.

100 Li H. A Short Introduction to Learning to Rank. *IEICE Trans Inf Syst* 2011;**94**:1854–62.

101 Li H, Xu J. Semantic Matching in Search. *Found Trends Inf Retr* 2014;**7**:343–469.

102 U.S. National Library of Medicine. UMLS API Technical Documentation. 2016.https://documentation.uts.nlm.nih.gov/rest/home.html (accessed 15 Mar 2018).

103 Herbrich R. Large margin rank boundaries for ordinal regression. *Adv Large Margin Classif* 2000;:115–32.

104 Xue X, Jeon J, Croft WB. Retrieval models for question and answer archives. *Proc 31st Annu Int ACM SIGIR Conf Res Dev Inf Retr* 2008;:475–82.

105 Och FJ, Ney H. A Systematic Comparison of Various Statistical Alignment Models. *Comput Linguist* 2003;**29**:19–51.

106 Och FJ. GIZA++: Training of statistical translation models. 2001.http://www.statmt.org/moses/giza/GIZA++.html (accessed 12 Mar 2018).

107 Bashyam V, Taira RK. Incorporating syntactic dependency information towards improved coding of lengthy medical concepts in clinical reports. *Proc Workshop BioNLP* 2009;:125–32.

108 Stenzhorn H, Pacheco EJ, Nohama P, *et al.* Automatic mapping of clinical documentation to snomed ct. *Stud Health Technol Inf* 2009;:228–32.

109 Nguyen A, Moore J, Lawley M, *et al.* Automatic extraction of cancer characteristics from free-text pathology reports for cancer notifications. *Stud Health Technol Inf* 2011;:117–24.

110 Andrews J, Patrick T, Richesson R, *et al.* Comparing heterogeneous SNOMED CT coding of clinical research concepts by examining normalized expressions. *J Biomed Inf* 2008;**41**:1062–9.

111 Dolin RH, Spackman KA, Markwell D. Selective retrieval of pre- and post-coordinated SNOMED concepts. *Proc AMIA Symp* 2002;:210–4.

112 IHTSDO. *SNOMED CT Technical Implementation Guide*. International Health Terminology Standards Development Organisation 2015.

113 Wasserman H, Wang J. An Applied Evaluation of SNOMED CT as a Clinical Vocabulary for the Computerized Diagnosis and Problem List. *AMIA Annu Symp Proc* 2003;:699–703.