APPROVAL SHEET

Title of Thesis:
Text and Ontology Driven Clinical Decision Support System

Name of Candidate: Deepal Dhariwal
Master of Science, CS, 2013

Thesis and Abstract Approved: _____

Dr. Anupam Joshi
Professor
Department of Computer
Science and Electrical Engineering

Date Approved: 4 / 29 / 2013

Curriculum Vitae

Name: Deepal Dhariwal.

Degree and date to be conferred: Masters in Computer Science, 2013.

Secondary education:

- Dr. Kalmadi Shamarao High School, Pune, 2005
- Fergusson College, Pune 2007

Collegiate institutions attended:

- University of Maryland Baltimore County, M.S. Computer Science, 2013.
- College of Engineering Pune, B.Tech (Information Technology), 2011

Major: Computer Science.

Professional positions held:

- Research Assistant, Ebiquity Lab, UMBC.
- Software Engineering Intern, Amazon Inc., USA
- Summer Intern, PTC Software (India) Ltd. , India

ABSTRACT

Title: TEXT AND ONTOLOGY DRIVEN CLINICAL DECISION SUPPORT SYSTEM.

Deepal Dhariwal, Masters in Computer Science, 2013.

Directed By: Professor Anupam Joshi, Department of Computer Science and Electrical Engineering.

In this work, we discuss our ongoing research in the domain of text and ontology driven clinical decision support system. The proposed framework uses text analytics to extract clinical entities from electronic health records and semantic web analytics to generate a domain specific knowledge base (KB) of patients' clinical facts. Clinical Rules expressed in the Semantic Web Language OWL are used to reason over the KB to infer additional facts about the patient. The KB is then queried to provide clinically relevant information to the physicians

We propose a generic text and ontology driven information extraction framework which will be useful in clinical decision support systems. In the first phase, standard preprocessing techniques such as section tagging, dependency parsing, gazetteer lists are used filter clinical terms from the raw data. In the second phase, a domain specific medical ontology is used to establish relation between the extracted clinical terms. The output of this phase is a Resource Description Framework KB that stores all possible medical facts about the patient. In the final phase, an OWL reasoner and clinical rules are used to infer additional facts about patient and generate a richer KB. This KB can then be queried for a variety of clinical tasks. To demonstrate a proof of concept of this framework, we have used discharge summaries from the cardiovascular domain and determined the TIMI Risk Score and San Francisco Syncope Score for a patient.

The goal of this research is to combine factual knowledge about patients, procedural knowledge (clinical rules), and structured knowledge (medical ontologies) to develop a clinical decision support system.

TEXT AND ONTOLOGY DRIVEN CLINICAL DECISION SUPPORT SYSTEM

By

DEEPAL DHARIWAL

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County, in partial fulfillment
Of the requirements for the degree of
MS
2013

Dedicated to my grandparents

Acknowledgements

I would sincerely like to acknowledge the following people who have played a very important role in the success of this work. I thank Dr. Anupam Joshi for giving his valuable time and advice for the past two years. Without his constant support and guidance this project would have been a distant reality. Thanks to Dr. Grasso for his invaluable inputs throughout my work. I would like to thank his team for evaluating the results of my work. Thanks to Dr. Finin and Dr. Yesha for reviewing my work and agreeing to be on my thesis committee. I would also like to thank all Ebiquity members for their support at every juncture of my thesis. It was great fun working with them and sharing knowledge. Finally I would like to thank the Oros Professorship endowment for funding my work.

List of Figures

List of Tables

Chapter I
INTRODUCTION

1.1 Motivation

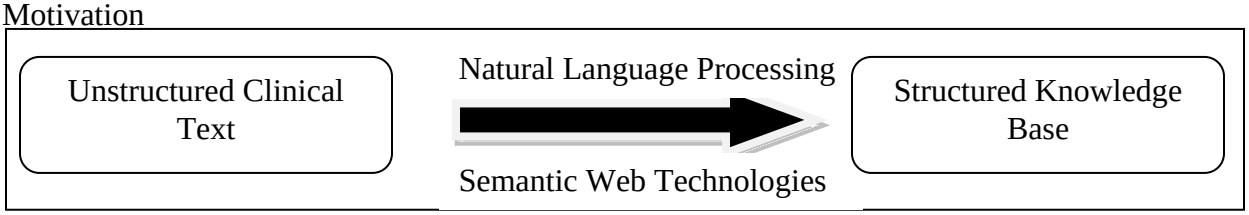

Fig. 1.1 Transformation from unstructured to structured data

Clinical narratives which include discharge summaries, admission notes, progress notes, physical exam results, form a significant proportion of electronic health records. They include a wealth of clinical information about patient such as medical history, observations during hospital course, medications, treatments and their outcomes, laboratory data in narrative format making this data highly unstructured. Hence one can't apply machine learning or data mining algorithms directly on such data. Though clinical narratives can be easily read by a physician, however a machine would require significant processing to understand its content. Hence there is a need to transform this data into a more structured format so that it could be used in decision support or information management tasks. Extracting clinically relevant and useful information from such data is highly difficult. This transformation can be done using natural language processing (NLP) and semantic web technologies. The NLP techniques would do lexical and syntactical analysis of text whereas ontologies would allow semantic analysis of the content. Applying only NLP techniques would result in extracting clinical concepts, however the true knowledge would be extracted when we are able to establish relationship between the extracted clinical concepts using semantic web technologies. Ontologies allow for automatic reasoning using subclass – super class relation and transitive properties.

Clinical entities extracted from records form the factual knowledge about the patient. When this factual knowledge is combined with known medical knowledge expressed in clinical rules, it would unearth hidden facts about patient which are either not mentioned or not evident at a first glance. Capturing those unstated facts about patients using known facts and medical knowledge is what we attempt to do in this work.

We do a proof concept of this idea by computing cardiovascular clinical scores of patient from the discharge summaries. The motivation stems from the need to risk stratify patients admitted to ER. With increasing number of patients it is imperative to identify patients with high risk for adverse outcomes such as syncope or hemorrhage. Currently a physician needs to manually read a discharge summary ; identify whether the patient has coronary artery disease risk factors, consider the patient's medical history and symptoms before taking any decision. In such scenarios clinical prediction tools such as TIMI[4] or CHESS[5] scores prove to be useful since they identify patients with adverse outcomes in 14 days. A high score indicates a patient is at higher risk for ischemic event and needs immediate medical attention. We intend to automate this process of computing the clinical scores. However we don't restrict ourselves to a limited set of clinical scores rather we attempt to automatically extract all possible facts about patient which would help differential diagnostic procedures.

Perhaps the following scenario would better explain the need of such automated clinical decision support systems.

A 78 year old man comes to ER for chest pain associated with nausea and shortness of breath. He has symptoms of hypertension and hypercholesterolemia and a positive family history of coronary artery disease. His daily medications include pravastatin, aspirin etc. The physical exam results reveal blood pressure 84/62 mm Hg, hematocrit 29.4, respiration rate 20 /min.

The physician has to now do a tedious task of reading the entire admission note for differential diagnosis. What if a system could assimilate all this information and provide the clinician with the one key information that he was looking for; the CHESS and TIMI score of the patient. Further if all these facts are stored in structured manner the physician could query for simple facts such as whether the patient has taken aspirin to ensure that he was not administering a conflicting medication or whether the patient has any other allergy or the simplest query is - does the patient has symptom X? The physicians could also ask complex queries such as whether the patient has symptom X, allergic to Y, diagnosed with Z, history of W. This is just a sample of queries that could be answered if the unstructured clinical text is stored in a structured manner using ontologies.

1.2 Problem Statement

Given an unstructured clinical narrative, compute clinically relevant physician queries. So in this work we intend to compute the TIMI and San Francisco Syncope Risk score of patients from the evidence present in discharge summaries.

A discharge summary is split into number of clinical sections. The patient's medical history might indicate whether he has coronary artery disease risk factors or congestive heart failure history. His current, past and suggested medications could be reported in Medication section. The laboratory results section would include physical exam results such as electrocardiogram, hematocrit and blood pressure levels. Thus the discharge summary would give factual knowledge about patients. When these facts are combined with the known medical knowledge i.e. medical concepts and relations between them and a set of clinical rules are given as input we could infer additional facts about patient which are not directly evident or mentioned in the discharge summary.

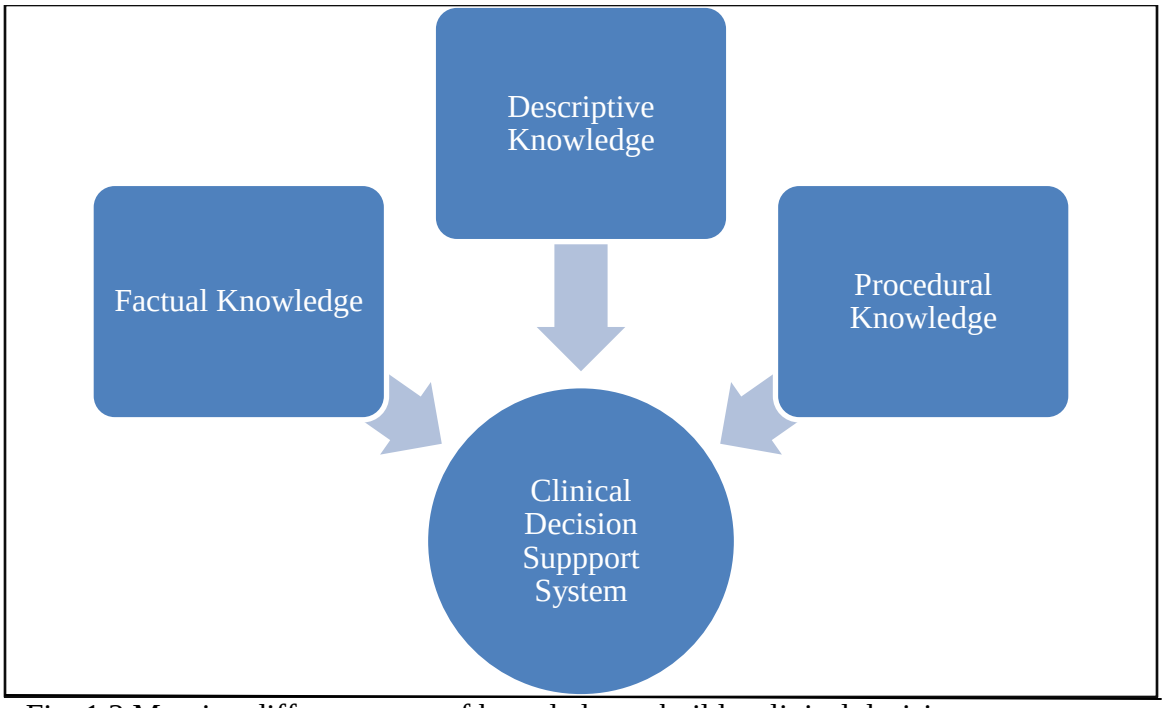

Fig. 1.2 Merging different types of knowledge to build a clinical decision support system[2]

For instance:
Cean Leach is a 38 year old male who is admitted in transfer from the Morehegron Valley Hospital in Ville , Virginia for further evaluation and management of accelerating chest pains syndrome . His medications that had been used individually or in combination include beta blockers and Lodapine cardiazem , nifidepine and coumadin , aspirin , cyclopadine and seratonin antagonists .
From the above description we get two facts about patient regarding his age and that he had been on aspirin dosage recently. This forms the factual known knowledge about patient. According to a medical ontology age is data type property associated with a patient and aspirin is a medication taken by a patient. This forms known descriptive medical knowledge. Now if we plug in a clinical rule which states that, 'if the age of patient is greater than sixty-five years and he has taken aspirin recently then his TIMI score = 2'. This forms the procedural knowledge. Hence after combining all these three types of knowledge we conclude that the patient has TIMI score = 1 and the patient is at very low risk for serious outcomes.
Thus given a discharge summary, domain specific ontology and a set of clinical rules, develop a system that would extract clinical entities (terms, values and additional context), store them in a structured format such as Resource Description Framework[35] and allow a reasoner to iterate over the knowledge base to infer additional facts about patient according to the given clinical rules. The specific problem that we are addressing is that how we can use text processing steps such chunking, parts of speech tagging, named entity recognition, regular expression in clinical domain and medical ontologies for extracting facts about patients. We are investigating how the structure of ontology and clinical rules could be utilized to uncover hidden facts about patient.

1.3 Use Cases

Some of the applications of the proposed system are:

1.3.1 Automatic event surveillance systems[8]: Based on the evidence present in discharge summaries we could write rules where certain combination of facts indicates adverse events and flags warnings. For instance combining all medication related facts we could set triggers when the patient has been administered conflicting drugs.

1.3.2 Decision Support System[8]: Building an expert system for diagnostic warnings and suggestions using facts reported in the discharge summary, clinical rules and a semantic reasoner. While a physician is assessing the patient's condition based on admission note a decision support system would spit suggested course of action and expedite the decision making process.

1.3.3 Research Purposes[8]: If we had rules of the form 'X is characterized by Y and results in Z'[1] then based on evidence suggested by records we could find plausible relations between symptoms, treatment, drug dosages and diagnoses.

1.3.4 Summary of Patient medical facts[8]: When a patient is admitted to ER the clinician is foremost interested in the cardiovascular findings about the patient. Instead of reading the entire discharge summary, the automated system would provide a compact representation of important and domain specific facts which would summarize patient's present, previous diagnoses, allergies etc.

1.4 Thesis Overview

In the next chapter we describe the related work in the domain of clinical text processing and decision support systems and literature review done for this thesis. In Chapter 3 we describe the proposed framework highlighting the main phases of the system. Chapter 4 delves deeper into the implementation details and explain the steps in each phase. Chapter 5 describes the experiments undertaken to validate the system and the evaluation results. Chapter 6 discusses certain limitations of the existing system and highlights the challenges involved in designing the system. Chapter 7 briefly describes some of the future extensions.

CHAPTER 2
RELATED WORK AND LITERATURE REVIEW

In this chapter we provide an overview of the domain of clinical text processing, medical knowledge representation and clinical decision support systems. We also discuss structure of electronic health records, clinical scores and medical ontologies. The chapter also introduces key technologies used in this research.

2.1 Clinical Text Processing

2.1.1 Related Work

2.1.1.1 Textractor: Hybrid System for medication and reason for their prescription extraction from clinical text documents[22]

In this paper the authors extract medication and related information from unstructured clinical text using machine learning, regular expressions and rules. The framework consists of analyzing the document structure, sentence and section segmentation, parts of speech tagging. They consider the context in which the medication occurs to extract the dosage, frequency, duration and reason for medication. The authors have built a dedicated system for i2b2 (Informatics for Integrating Biology and Bedside) medication challenge. The paper gave key insights into the preprocessing steps required to analyze unstructured clinical narratives especially the i2b2 records.

2.1.1.2 Extracting medical information from clinical text[19]

In this paper as well, the authors propose a framework to extract medication and related information from i2b2 records which are mainly discharge summaries written by physicians in plain text. The structure of a discharge summary would be discussed later in this chapter. Text filtering was done to eliminate content not related to medication of patients and clinical terms were identified by building a vocabulary of Unified Medical Language System terms. In our approach we consider the UMLS vocabulary as an upper bound for clinical terms. All other terms or words occurring in the document except numeric values are not considered in the decision support module.

2.1.1.3 Extracting medical information from narrative patient records : the case of medication related information[18]

In this paper, the authors develop a framework to extract medication information from clinical records. The preprocessing steps include sentence and section segmentation using rules. A lexicon is used to identify drug names and additional rules are written to extract associated information. In our approach we use SecTag for identifying clinical section headings and cTAKES Sentence annotator for sentence segmentation. Further cardiovascular related knowledge is extracted using heart failure ontology and combination of cTAKES annotators is used to extract the additional information associated with the clinical entities.

2.1.1.4 Study of machine learning based approaches to extract clinical entities and their assertions from discharge summaries[32]

In this paper the authors extract clinical entities namely medical problems, tests, treatments and their asserted status from discharge summaries and evaluate machine learning approaches for the tasks. Briefly their approach is to identify the boundary of a clinical entity i.e. terms at the beginning of the entity, at the end and within the entity. A multi class classifier built using Conditional Random Fields (CRF) and Support Vector Machine (SVM) is used to assign a label to every word in the document. The text processing phase includes a bag of words model, extracting orthographic information for every word such as prefix, suffix, capitalization, parts of speech tagging, lexical and

semantic analysis. The text processing phase in our work includes parts of speech tagging, sentence detection, chunking. However certain clinical entities such as 'ST segment changes' or 'family history' require additional text processing to identify the polarity of entity, the clinical section in which the entity occurs, the numerical terms present in the proximity of the clinical term

2.1.1.5 High Accuracy Information extraction of medication information from clinical records[20]

In this paper the authors have developed an information extraction engine using a cascade of machine learners and a set of rules to automatically extract medication information from clinical records. The main components of the system include a preprocessing engine, CRF generator, SVM classifier that determines the relation between the two extracted clinical entities and a rule based engine for section tagging. So the difference in approach comes from the fact that the authors intend to extract medication name, dosage, mode of administration, frequency, duration, reason associated with a medication term found in document whereas we aim to extract only the value associated with a cardiovascular medical term found in document. The authors use additional context surrounding the medication and apply SVM and CRF algorithms to extract these attributes. We intend to use dependency parsing and lookup window of the concept to extract the value associated with the concept.

2.1.1.6 Automated extraction of ejection fraction from quality measurements using regular expressions in UIMA for heart failure.[7]

In this paper the authors have built a Natural Language Processing system using regular expressions and rules to extract ejection fraction from free text echocardiogram reports. The medical concepts are identified using multi part regular expressions and numerical values are extracted using number token patterns. A binary classifier is used to determine whether the patient is eligible for life prolonging treatment based on its ejection fraction value. The goal of our work is to develop a generic framework to extract medical facts from electronic health records. We don't intend to develop a specialized NLP system for accurately extracting value of every clinical entity. For instance there has been considerable research in identifying the smoking status of a patient from the discharge summaries. However in our work, to determine whether the patient is an active smoker, we check whether the term 'smoker' occurs in document and if so whether it is negated and the clinical section in which the term occurs.

2.1.1.7 Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text[14]

In this paper the authors, compare supervised and unsupervised dimensional reduction techniques for analyzing unstructured clinical text. A proof of concept of their framework is to predict whether a patient who has sepsis is likely to develop an infection and whether the patient is likely to be admitted to ICU. They use limited preprocessing and focus more on dimensional reduction for extracting clinically relevant information. However we try to maximally exploit text analytics such as negation detection, dependency parsing to extract clinically relevant information.

2.1.2 Tools

2.1.2.1 Clinical Text Analysis and Knowledge Extraction System (cTAKES)[28]

Apache cTAKES is an open source natural language processing tool to extract information from unstructured clinical text. Built on top of Unstructured Information Management Architecture (UIMA) and OpenNLP, cTAKES identifies clinical entities such as drugs, diseases, signs, symptoms. Every clinical entity is associated with the text

span, UMLS Ontology mapping code, polarity. Some of the cTAKES components that have been used in this research are sentence boundary detector, context dependent tokenizer, dictionary lookup annotator, and negation detector.

2.1.2.2SecTag – Tagging Clinical Note Section Headers[38]

SecTag developed at biomedical language processing lab at Vanderbilt University provides an index of clinical section headings. SecTag maps clinical sections with different names but similar meaning to a root clinical section header. For example past medical history or history of past illness is mapped to past_medical_history. SecTag algorithm uses NLP, Bayesian model, spelling correction and scoring technique to identify sections in clinical notes. We don't use SecTag algorithm in our work, however make use of the index of clinical section headings.

2.1.2.3 Other Clinical Text Processing Tools

Some of the other widely used clinical text processing tools are MedLee – text processor that extracts and structures clinical information from textual reports and translates the extracted terms to a controlled vocabulary; HiTex – health information text extraction system to identify principle diagnoses , medication administered and smoking status of patient; Linguistic String Project which couples language specific NLP modules with domain specific information extraction modules.

2.2 Medical Ontologies

2.2.1 Related Work

2.2.1.1 Representing UMLS Semantic Network using OWL[6]

In this paper the author puts forth certain limitations in transforming UMLS semantic network to an ontology represented in OWL. In their proposed hierarchical approach the nodes correspond to UMLS semantic types, links represent the semantic relationships where each link could be a spatial, physical, temporal, functional and conceptual relationship. The UMLS Semantic Network (SN) semantic types are mapped to a description logic concept, the SN relationship corresponds to Description Logic Role or OWL property. The type or relationship hierarchy corresponds to subclass or sub property axiom in the ontology. This paper put forth certain challenges in developing a domain specific ontology using UMLS vocabulary. Hence we decided to use an off the shelf ontology for heart failure domain. However in future we intend to extend the heart failure ontology with additional cardiovascular terms using UMLS Rich Resource Format Browser. The additional terms would be mapped as instances of classes defined in UML Semantic Network.

2.2.1.2 Ontologies in Medical Knowledge Representation[33]

In this paper the authors discuss the structure of medical ontologies, how they are constructed based on the scope of ontology, domain specific knowledge and the ontology creator tool. The authors also introduce the heart failure ontology developed as part of Heartfaid Union project. The authors state that while creating a medical ontology, all relevant terms should be represented in a hierarchical manner; classes should be known concepts in some existing medical terminology. As mentioned before, because of limited domain expertise we intend to use an off the shelf domain specific ontology and extend it as required.

2.2.1.3 Building ontology of cardiovascular diseases for concept based information retrieval[34]

In this paper, authors develop a specialized ontology using NLP techniques and existing medical knowledge bases such as UMLS, MESH. The main classes in this ontology

are cardiovascular diseases, cardiovascular agents, cardiovascular physiology, cardiovascular surgical procedures, cardiovascular systems, cardiovascular abnormalities, cardiovascular diagnostic and techniques, cardiology and cardiovascular models. The ontology correlates different cardiovascular concepts, however doesn't define properties that relate a patient to cardiovascular concepts. For instance the heart failure ontology has a Patient class and set of properties such as HasDiagnosis or HasSignOrSymptom or TakenMedication that relate an instance of type Patient to a heart failure concept. Hence after careful deliberation we plan to use Heart Failure ontology in our work.

2.2.2 Ontologies used in this research

2.2.2.1Unified Medical Language System[1]

The Unified Medical Language System developed by US National Library of Medicine combines a number of biomedical vocabularies such as ICD-10, MeSH, SNOMED-CT, RxNorm etc. The UMLS Metathesaurus includes 1 million biomedical concepts and 5 million concept names with each concept having a concept unique identifier and links to equivalent terms in other vocabularies. The UMLS also includes clinical test processing tools such as RRF browser that allows creating a subset of UMLS terms for a specific domain. We initially intended to use UMLS SN as the base ontology for reasoning, however lack of well defined classes for cardiovascular concepts, absence of Patient class and properties that correlate patient with medical concepts prompted us to move to a different and perhaps more domain specific ontology. Further the UMLS SN is inconsistent and has faulty relations due to which OWL reasoner isn't able to draw appropriate inferences. However the Heart Failure (HF) Ontology has been implemented using Ontology Web language and Protégé editing tool. The authors of HF ontology have also developed an expert reasoning system using HF Ontology, SWRL rules and Jess reasoner. Hence considering all these factors we decided to use heart failure ontology in our work. However we intend to use UMLS Metamorphsys RRF browser to update the Heart Failure ontology with additional cardiovascular terms. Further the cTAKES UMLS dictionary lookup annotator maps cTAKES annotations to UMLS dictionaries. It looks for matches where words in dictionaries appear in same order as in document or exact matches and some limited canonical forms.

2.2.2.2 Heart Failure Ontology[2]

The heart failure ontology has been developed specifically for the heart failure domain and the paper also discusses how the ontology can be used for a variety of clinical tasks. The HF ontology includes 200 classes, 100 properties and 2000 instances. The ontology has been built in accordance with the latest guidelines by the European society of cardiology for diagnosis and treatment of acute and chronic heart failure. Each instance of the ontology is associated with a concept unique identifier from the UMLS dictionary. The ontology also includes the Patient class which is populated with data extracted from a patient's medical record. This class allows storing factual knowledge about patient. The patient class has around 40 (including data and object) properties. The paper also put forth certain limitations of the ontology. For example the ontology doesn't allow associating a value with a medical concept such as hematocrit. The ontology has instances such as Cardiac_output_less_than_2_l_per_min; however the ontology lacks a datatype property that associates a numerical value with a clinical concept. Hence we

propose to add object and datatype properties to ontology depending upon the needs of the decision support system.

2.3 Clinical Scores

Cardiovascular risk scores collect information about patient and estimate cardiac risk. Cardiac risk stratification allows identifying patients which are at higher risk for sudden outcome such as cardiac arrest or death. Thus using laboratory test results along with vital signs such as blood pressure, heart rate, presence or absence of coronary risk factors such history of coronary artery disease, hypertensive blood pressure, and family history would allow identifying patients at high risk for ischemic events and thus protect them from sudden cardiac arrest. For example the Detsky and Goldman scores tell the risk for cardiac death by considering the patient's medical history, recent angina episodes, and echocardiogram test results. Following are the two clinical scores that we compute to demonstrate a proof of concept of our work:

2.3.1 Thrombolysis in Myocardial Infarction (TIMI) Risk Score[4]

TIMI risk is to categorize the risk of death and ischemic events in patients experiencing unstable angina. The TIMI score is calculated as follows:

- Whether Age of patient greater than 65 years
- Whether Aspirin was administered in past 7 days
- Whether the patient had at least 2 angina episodes within last 24 hours.
- Whether ST changes were of at least 0.5 mm
- Whether the patient has elevated cardiac biomarkers
- Whether patient has known coronary artery disease i.e. whether patient has percentage of stenosis greater than 50 %
- Whether the patient has at least 3 coronary risk factors from amongst diabetes, hypertension, current smoker, family history of CAD, hypercholesterolemia

2.3.2 San Francisco Syncope Rule to compute CHESS score[5]

The San Francisco Syncope rule is used for evaluating the risk of adverse outcomes in patients with syncope. The Chess score is calculated as:

- Congestive Heart Failure History
- Whether Hematocrit value is less than 30 %
- Whether patient has abnormal EKG
- Whether patient has a history of shortness of breath
- Whether Systolic Blood Pressure is less than 90 mm of Hg

A patient who has any of the above features has high risk for serious outcomes such as death, myocardial infarction, arrthymia, pulmonary embolism, stroke or even hemorrhage.

2.3.3 Clinical Decision Rules

A clinical decision rule combines a set of factors such as patient's social and medical history, laboratory results, signs and symptoms to help clinician with diagnostic and prognostic assessments. Such rules allow clinicians to group patients into various risk categories commonly known as risk stratification. For example the ROSE clinical decision rule predicts serious outcome or death in patients with syncope. A clinical decision rule in cardiovascular domain will allow emergency physicians to accurately identify patients with chest pain who are safe for early discharge. Thus a clinical rule

takes patient's signs, symptoms and other findings to predict the probability of a specific disease or outcome.

2.4 Electronic Health Records

2.4.1 Related Work

2.4.1.1 Processing Electronic Medical Record[8]

In this Master's thesis the author has developed a framework using cTAKES, UIMA, and Weka that uses structured output to identify whether a patient has obesity or correlated morbidities. The text processing phase includes XML parser, section segmentation and a morbidity annotator that associates every medical term found in document with one of the sixteen morbidities. The machine learning phase uses decision trees to associate one of the sixteen morbidities with patient. In our approach we use semantic web analytics along with text analytics to identify medical facts present in the document. Further they use a machine learning algorithm in the decision making process with the features provided by the text processing phase. However we store the results of text processing phase in a structured knowledge base and use SPARQL queries and Clinical Decision Rules to answer variety clinical queries.

2.4.2 i2b2 – Discharge Summaries[3]

Clinical Narrative is a first person story written by clinician that describes specific clinical event or situation. It includes the admission and discharge date, demographic information such as age, race, gender, present as well as past medical history, signs and symptoms observed during the course of hospital stay, treatments and surgeries undertaken and their outcomes, laboratory tests and their results, medications administered. Sometimes discharge summaries also include the social history of patient, detailed explanation of the event due to which patient was admitted, background conditions, conversations with family members. The i2b2 records include sections such as history of present illness, past medical history, medications administered on admissions, social history of patient, physical examination results, laboratory data, observations made during the hospital course. For every i2b2 medication challenge a new data set is released. Cumulatively the data set includes around 889 medical records. The following figure shows a sample discharge summary from i2b2 data set.

<doc id="330">
<text>
435791816 | CULHS | 80771325 | | 428219 | 1/12/1994 12:00:00 AM | Discharge Summary | Unsigned | DIS | Admission Date: 1/12/1994 Report Status: Unsigned

Discharge Date: 9/15/1994
PRINCIPAL DIAGNOSIS: CORONARY ARTERY DISEASE.
HISTORY OF PRESENT ILLNESS: Mr. Weddel is a 52-year-old man status post coronary artery bypass graft x1 , with saphenous vein graft to the left anterior descending in 1976. He presents with a 5-6 week history of unstable angina. He had an echocardiogram which revealed recurrent coronary artery disease. The patient was referred to Dr. Finstad for reoperative coronary artery bypass graft.
PAST MEDICAL HISTORY: The patient has a history of coronary artery disease and peripheral vascular disease.
PAST SURGICAL HISTORY: The patient is status post a radical prostatectomy complicated by osteitis pubis and a urethral colonic fistula.
CURRENT MEDICATIONS: Aspirin , Lopressor , Procardia XL , intravenous heparin and intravenous nitroglycerin.
ALLERGIES: No known drug allergies.
LABORATORY DATA: BUN and creatinine 16/1.2 , white count 5500 , hematocrit 46.
HOSPITAL COURSE: The patient was admitted to the Cardiology Service on September , 1994 , where he was stabilized and finally taken to the Operating Room on July , 1994 , at which time he underwent a reoperative coronary artery bypass graft x3 , with a left interior mammary artery to the left anterior descending. The patient tolerated the procedure well and was hemodynamically stable in the unit. He was weaned overnight and extubated on postoperative day 1. He was started on aspirin and Lopressor , weaned off his O2 and started on diuretics. He was transferred to the unit on postoperative day 1. Once on the unit the patient was gradually advanced to a regular diet and weaned off his O2. He was diuresed down to his preoperative weight. He had a routine postoperative course.
DISPOSITION: Home on August , 1994.
CONDITION ON DISCHARGE: Good.
DISCHARGE MEDICATIONS: Aspirin one tablet p.o. q. day; Lopressor 25 mg p.o. t.i.d.; and Percocet p.r.n.
FOLLOWUP: The patient will follow up with Dr. Mcgoff and will contact his office for an appointment after discharge.
Dictated By: CLINTON H. BARRET , M.D. TM2
Attending: DANIAL C. DELL , M.D. SY3 RZ941/4768
Batch: 8363 Index No. K1QM0GBOY D: 1/13/94
T: 6/15/94</text>
</doc>

Fig. 2.1 Sample Discharge Summary from i2b2 data set

2.5 Knowledge Representation

Any clinical decision support system needs a knowledge base (KB) layer which stores domain knowledge. Some of the popular medical knowledge bases include SNOMED CT ontology, UMLS, Medline Metathesaurus. This KB can be stored in structured manner using Resource Description Framework (RDF) triple store or UIMA Typed Feature Structure such as UIMA CAS object depending upon the end user application.

2.5.1 Related Work

2.5.1.1 Ontology Driven Information Extraction(ODIE)[47]

In this paper, authors give an overview of how ontology driven information extraction systems could be used for a variety of clinical tasks. Information Extraction means extracting clinical entities, filtering negated entities, finding relations between the entities and storing it in structured format. Electronic Health records are free text present in clinical notes, discharge summaries, admission and progress notes. The clinical entities include diagnoses, symptoms, drugs, events such as procedures and outcomes. The information extraction phase in an ODIE system includes chunking, word sense disambiguation, sentence detection, tokenization, parts of speech tagging, identifying negation and temporality. The extracted information is stored in structured manner for efficient retrieval and reasoning. The paper puts forth certain challenges in extracting clinical entities from unstructured text such as correct spellings which can be resolved using Levenshtein score, or using Metaphone or Tolentina spell cleaners. Similarly word

sense disambiguation problem is addressed by developing an unsupervised classifier over Medline and UMLS metathesaurus. They also discuss how abbreviations occurring in clinical text can be resolved by following the approach of Xu et. Al. In our proposed system we address some of these issues viz. abbreviations and synonyms. The heart failure ontology gives a list of UMLS synonyms for all instances in the ontology. They also include the commonly used abbreviations for these instances.

2.5.1.2 Ontology Based Information Extraction Survey of Current Approaches[12]

In this paper the authors discuss several approaches of using domain specific ontologies for information extraction. An OBIE system comprises of a text preprocessing module, ontology generator, semantic lexicon and an information extraction system built using regular expressions. In our work we use an off the shelf ontology and extend the ontology depending on the needs of the decision support system. Using an off the shelf ontology allows to build a structured knowledge base of patient facts with respect to a particular domain such as cardiovascular or cancer or neurology. Further another disadvantage in having an ontology generator module is our limited domain knowledge. With an ontology generator one needs to have a domain expert as well as an ontology updation module. In the proposed approach we plan to extend the heart failure ontology by building a gazetteer list of cardiovascular concepts using UMLS Metamorphsys RRF browser.

2.5.1.3 Towards an OWL based framework for extracting information from clinical text[15]

In this paper the authors develop an owl based framework to extract clinical entities. In the first phase the text is parsed into an owl knowledge base using generic ontology and standard text processing techniques such as parts of speech tagger, gazetteer lists, and forest chart parsers. Since they use general purpose ontology they require SQWRL rules to extract evidence of medical conditions. For example a combination of terms indicate the word following the group of terms, is a medical event. However in our approach we use a domain specific ontology to extract clinical entities and establish relations between them.

2.5.1.4 Developing Factual Knowledge from Medical Data by Composing Ontology Structures[30]

In this paper the authors discuss how to populate ontology, from a database, which can be subsequently used for reasoning in decision support. Ontology is used to represent both concepts from medical domain as well as factual data about patients. The reasoner takes the knowledge and rules that connect the concepts as input and infers additional knowledge about patient. In this paper the factual data about patient is gathered from a structured database where the columns are patient id, first or last name, date of birth, age, gender etc. Hence they suggest additional modifications, to handle those columns which don't have a corresponding class in the ontology or are missing values. Further they also discuss how to handle additional information attached with an event such as if patient has taken a test then the name of test, test results missing values and accordingly update the ontology. The last three cases are some of issues that even we try to address in this work. Further the input data set used in our work is highly unstructured.

2.5.2 Ontology Population Tools

2.5.2.1 GATE – OntoRoot Gazetteer[44]

GATE allows ontology based information extraction. Terms occurring text are mapped to classes or instances or properties from ontology. GATE also allows for ontology population. Given ontology and a text document GATE populates the ontology concepts with instances derived automatically from a text. The GATE OntoRoot Gazetteer finds mentions in text matching classes, instances, data property values in ontology. OntoRoot

Gazetteer handles morphological, typographical variants, camel case names, hyphens, underscores. After multiple attempts we weren't able to load Heart Failure ontology in GATE. HF ontology has been tested for Protégé, however due SWRL name space problems; it couldn't be loaded in GATE.

2.5.2.2 ODIE[47]

ODIE allows to find terms in text that correspond to instances of classes which constitutes the information extraction task, generate an inference model using ontologies and find the relation between the terms extracted. ODIE is currently not in active use. Further since we aim to extract maximum information from text we don't restrict to instances. Rather we search for presence of classes as well as instances. Also for every term found in document we consider all the properties whose range is the super class of the term and the domain of the property is the Patient class and accordingly assert a fact.

2.6 Clinical Decision Support Systems

The goal of a clinical decision support system is to create an automated knowledge base of patient facts based on evidence present in the discharge summaries and use this KB for decision support. This mainly involves extracting clinical concepts from electronic health records, mapping them to concepts in medical ontologies and using clinical rules to reason over the generated KB and deducing additional facts.

2.6.1 Related Work

2.6.1.1 Ontology Based Distributed Health Record Management System[11]

In this paper the authors have developed a distributed health record management system using heart failure ontology. The knowledge base layer includes patient records and medical ontologies. This KB along with clinical rules is used for reasoning and decision support. Their framework also covers data security of medical records and interaction between general practitioners and patients and laboratory personnel. Unlike the multi layered system proposed in the paper we stick to a single layer comprising of factual knowledge, procedural knowledge and descriptive knowledge which forms the basis for decision support.

2.6.1.2 Patient Surveillance Algorithm for Emergency Department[14]

In this paper the authors have developed an algorithm based on Naïve Bayes model to detect likelihood of sepsis using the patient's medical history, laboratory results, demographic information and real time data as input. They propose a probabilistic model that uses information from discharge summaries to reason whether a patient is likely to develop sepsis. In our approach we use semantic reasoning to make a decision and use text and semantic web analytics to extract medical facts. However the input for decision making task is similar to what we do in our work.

2.6.1.3 Semantic Web Ontology Utilization Heart Failure Expert System Design[48]

In this paper the authors propose a clinical decision support system based on heart failure ontology. The paper describes three kinds of knowledge to develop an expert system based on instance checking – descriptive knowledge which is derived from medical ontologies, procedural knowledge provided in terms of clinical rules and the factual knowledge gathered from clinical notes. The paper also explains the heart failure ontology used in the proposed system and how reasoning could be done over procedural and descriptive knowledge. In our approach we combine text analytics with knowledge generation to develop a clinical decision support system. The reasoner takes the patient

facts represented as RDF triples, clinical rules and ontology as input and makes diagnostic suggestions

2.7 Semantic Web Technologies

2.7.1 Resource Description Framework[35]

A RDF triple is a labeled connection between two resources and consist of subject, predicate, object where subject and predicate are URI i.e. resources on web and object is a URI or literal. Collections of RDF statements form a RDF graph. In other words a RDF graph is directed labeled graph connecting different RDF nodes. RDF triples can be serialized in a variety of formats such as XML, N-Triples, and Turtle. In our work we use the RDF-XML serialization format. Further, an instance of class Patient forms the root of the RDF graph and classes and instances of Heart Failure Ontology form the child nodes and the data and object properties form the predicates. A semantic reasoner uses the patient's facts expressed as RDF triples and clinical rules as input and infers additional facts about patient. An example of full blown RDF graph after applying bunch of clinical rules is explained in the implementation section.

2.7.2 Web Ontology Language[34]

OWL (web ontology language) a W3C standard has been developed for representing ontologies in RDF/XML serialization format. The main motivation behind developing OWL is to process information on web. Ontology allows defining set of individuals and properties that relate the individuals, constraints that govern the definition of classes and properties. OWL has 3 sublanguages OWL Lite, OWL Description Logic, OWL Full. Each of these sublanguages is a syntactic extension of its predecessor.

2.7.3 Jena framework[46]

Apache Jena is an open source framework to build semantic web applications in Java. In our work we use Jena-OWL API to navigate heart failure ontology, Jena rule based inference engine for reasoning over RDF graph, Jena-RDF API to store patient facts as RDF triples and Jena-SPAQRL API for query RDF KB for clinical scores.

2.7.4 Reasoner

A semantic reasoner infers logical consequences from a set of asserted facts or axioms. The inference rules are specified by the ontology language and the reasoner. Some of the current state of art reasoner includes Pellet, Fact++, RACER, Jena, Bossam, owlim, CWM. In our work we use the Jena OWL reasoner to generate an inference model based on patient facts, ontology and clinical rules.

Chapter III

PROPOSED FRAMEWORK

The proposed framework consists of three phases: first in which text analytics are applied to discharge summaries to extract clinical entities, second in which semantic web analytics is used to store extracted information in a structured knowledge base and the in the third phase in which the KB is queried for clinical scores and additional facts are inferred using clinical rules.

3.1 Phase I: Information Extraction:

In this phase a combination of annotators is used to extract various clinical score parameters. This phase is divided into a number of steps beginning with processing the raw i2b2[3] records using a XML parser. Every individual record is then parsed using cTAKES[36] Aggregate UMLS[1] Document Processor. The cTAKES processor includes text preprocessing steps such as parts of speech tagging, chunking, dependency parsing, sentence segmentation, medical named entity recognition, numeric, measurement , date entity recognition. The cTAKES annotation results are then analyzed and combined using an annotation parser to extract clinical score parameters. For instance the result of measurement, polarity and dictionary annotators is combined to extract the clinical term 'ST segment' and the associated value in mm. Another example is combining the dictionary lookup annotator with section tagger to extract the clinical term 'aspirin' and the clinical section in which it is mentioned. Thus combination of annotators is used to extract clinical entities and the context in which they occur. For instance the number token annotator is combined with section tagger to extract the age of patient from the section - history of previous illness.

While working with unstructured clinical narratives just the present or absence of a term is not sufficient to assert a fact. Additional processing needs to be done to understand whether the entity is conditionally present or whether it is negated. The text surrounding a word can change assertion being made about the term for instance 'patient has symptom of breathlessness' and 'patient didn't experience shortness of breath' are two different facts. Hence there are four types of assertions that we consider positive, negative, uncertain, and conditional. Analyzing the assertion changes the interpretation of a statement. Hence in this phase we also extract the polarity of a clinical term using cTAKES Assertion Annotator. For instance results of assertion annotator, sentence annotator and dictionary lookup annotator are combined to extract various signs and symptoms observed in patient.

Further depending upon the target clinical parameter, additional context (such as associated value or clinical section or polarity) is extracted as well. To give a final example, consider the clinical parameter systolic blood pressure. In the information extraction phase the term 'blood pressure' would go through the following syntax analysis: The UMLS dictionary annotator returns blood pressure as a clinical term, the sentence annotator returns the sentence in which the term blood pressure occurs, the number token annotator returns all the numeric entities mentioned in the extracted sentence and the text spans of entities help to identify the value closest to the term blood pressure.

The information extraction phase mainly deals with syntactic analysis of unstructured clinical text using parts of speech tagging, shallow parsing, chunking and extracts diagnoses, observations, test results, demographic information from clinical narratives.

3.2 Phase II: Knowledge Generation

In this phase, a structured knowledge base is generated using ontology. Semantic analysis of the extracted terms is done using ontology that attaches a meaning and semantic type to the term allowing further reasoning. Ontologies identify relations between extracted entities and reasoner infers additional information. A domain specific ontology is used to map UMLS medical terms returned by cTAKES to a semantic type from Heart Failure $(HF)^2$ ontology. This is similar to building a gazetteer list of cardiac terms but with the additional information about the relationship between the terms. Identifying classes and instances of an ontology from a text allows querying the text using an ontology query language and building an inference model.

To query the text one needs to build a structured knowledge base. In our current work the KB is a RDF graph with each node being a clinical term and the edges being the properties from HF ontology. A clinical fact is stored as a RDF triple and set of RDF triples forms a RDF graph. So we construct a RDF graph with the root being an instance of the Patient class from HF ontology. All other clinical terms form the child nodes of this graph. Thus the age of patient gets mapped as a data type property while the medication aspirin is mapped as an object property. Each cardiovascular term which corresponds to an instance or class of ontology is associated with the patient using either a data or an object type property. Individually the clinical terms 'hypertension', '75 years', 'abnormal EKG' doesn't contribute to the clinical findings about patient. However when we build a RDF graph which asserts the patient's age is 75, patient has been diagnosed with hypertension and he is showing signs of abnormal EKG one can conclude set of other facts about the patient such as he has two coronary artery risk factors. Thus in this phase, we identify instances, properties, classes of an ontology that could be mapped to terms occurring in text. The output of this phase is populating the class Patient with factual knowledge extracted from the records. The categorical data is mapped using object properties and the numerical data is asserted using data properties.

Clinical sections are used to assert the appropriate predicates. For instance there is a huge difference between asserting aspirin as Suggested_Medication as compared to Taken_Medication or between patient hasDiagnosis or hadDiagnosis of chronic heart failure. Another parameter which is handled in this phase are synonyms – i.e. semantically equivalent clinical terms but lexically different for example bradycardia and low heart rate need to be treated in same manner. Similarly CHF, chronic heart failure refers to the same medical condition. We also handle the canonical forms and abbreviations of clinical terms in this phase based on the knowledge ontology provides.

A single clinical entity such as hematocrit value = 30 % is decomposed into a number of triples such as patient has taken hematology and biochemistry test, which measures hematocrit level in percentage and whose value is 30.

Thus in knowledge generation phase we build a KB of patient facts using heart failure ontology and the data extracted from the information extraction phase. This KB is then queried in the next phase for a variety of clinically relevant information such as clinical scores, predicting the likelihood of adverse outcome in patients coming to ER with syncope or unstable angina.

3.3 Phase III: Inference and Querying:

In this phase, the KB generated in previous phase is queried for clinically relevant information and clinical rules are used for inferring additional facts about patients. We can query the RDF graph of patient facts to answer intelligent queries such as whether the patient has systolic blood pressure less than 90 mm of Hg. This can be answered using two RDF triples viz. whether the

patient has been diagnosed with blood pressure and if so the value associated with the blood pressure.

Using naive clinical rules co relating medical terms we infer additional facts. For instance if there is a clinical rule that states that , 'aspirin dosage is related to myocardial infarction'[2] then if the RDF graph has a triple (Patient, TakenMedication, Aspirin) the reasoner concludes that the person has been diagnosed with myocardial infarction and adds a triple

(Patient, HasDiagnosis, MyocardialInfarction) to the RDF graph. Consider another example, if we have a rule[2] that links bradycardia and sick sinus syndrome then if we assert that patient has been diagnosed with bradycardia the reasoner infers that the cause might be sick sinus syndrome.

The above rules were examples of rules co relating medical concepts. We could also exploit the hierarchical relationship between the instances and classes of ontology to infer additional facts.

For instance the knowledge generation phase asserts that the patient has been diagnosed with 'systolic_hypertension' and links it to an instance of class 'hypertension'. The ontology defines is-a relation between 'systolic_hypertension' and 'hypertension' and hence the reasoner infers that patient is diagnosed with hypertension.

Applying set of clinical rules and iterating multiple times over the RDF graph the reasoner creates a rich KB of patient facts. For instance if the patient had symptom of dyspnea we could possibly infer that it could be related to pulmonary disease or mitral valve regurgitation. If we had a clinical rule and some additional context we could assert this to be true or false. Based on that evidence, we could further conclude that if the cause is pulmonary disease then the patient had been previously admitted to hospital for heart failure. In such a way we could have an extended RDF graph of patient facts.

Once we have the entire knowledge base, we can answer complex queries such as 'does the patient have hematocrit value less than 30 %' and 'symptom of dyspnea' and 'congestive heart failure history' and 'abnormal EKG' and 'systolic blood pressure less than 90' which is essentially the CHESS score of patient. In our work we query the knowledge base for TIMI and CHESS score parameters.

Chapter IV
SYSTEM IMPLEMENTATION

HISTORY OF PRESENT ILLNESS:
*The patient is a **76** year old female with a history of schizotypal personality disorder , congestive heart failure , peripheral vascular disease , **diabetes mellitus** , chronic renal insufficiency , and a documented ejection fraction of 18% , who presented to the Heaonboburg Linpack Grant Medical Center Emergency Ward with shortness of breath . The patient **denies** any fevers , chest pain , **and shortness of breath** , cough , and is generally uncooperative with the history .*
***PAST MEDICAL HISTORY** :*
*The patient was admitted in December of 1992 , at that time with anasarca and **congestive heart failure** , responsive to diuretics and ACE inhibitors .She has a diabetic cardiomyopathy with an ejection fraction of 18% , likely due to alcohol abuse in the past .*
MEDICATIONS ON ADMISSION :
1. Lasix 80 mg PO q.day ,
2. Colace 100 mg PO t.i.d.
*3. **Aspirin** 40 mg*
PHYSICAL EXAMINATION :
On physical examination , an uncooperative white female sitting up in bed in no acute distress .
*Her vital signs included a **blood pressure of 166/108** , heart rate 100 , temperature refused , respirations 24 , oxygen saturation on room air 92% , 99% on two liters .*
LABORATORY DATA :
*significant for a sodium of 140 , potassium 5.7 , chloride 108 , bicarbonate 21.2 , BUN and creatinine 66 and 2.9 , glucose 198 , calcium 8.4 , phosphate 4.0 , magnesium 1.7 , total bilirubin 1.0 , alkaline phosphatase 153 , SGOT 21 , LDH 236 , amylase 32 , urate 7.2 , albumen 2.4 , **hematocrit 39.9** , white blood count 10.7 , platelet count 197,000 , mean corpuscular volume 79 , coagulation studies within normal limits .*
*Her chest X-ray showed cardiomegaly with bilateral pleural effusions and moderate interstitial / air space pattern consistent with congestive heart failure . The **electrocardiogram showed sinus tachycardia at 100 beats per minute with left bundle branch block** .*

→

Clinical Score	Value
TIMI	1
SYNCOPE	2

Fig. 4.1 Input: Discharge Summary and Output: Clinical Scores

In this chapter we explain the detailed implementation of the system that extracts clinical scores and cardiovascular facts from a discharge summary (unstructured) by transforming it to a RDF knowledge base of clinical facts (structured) using natural language processing and semantic web technologies.

4.1 Preprocessing raw i2b2 data set

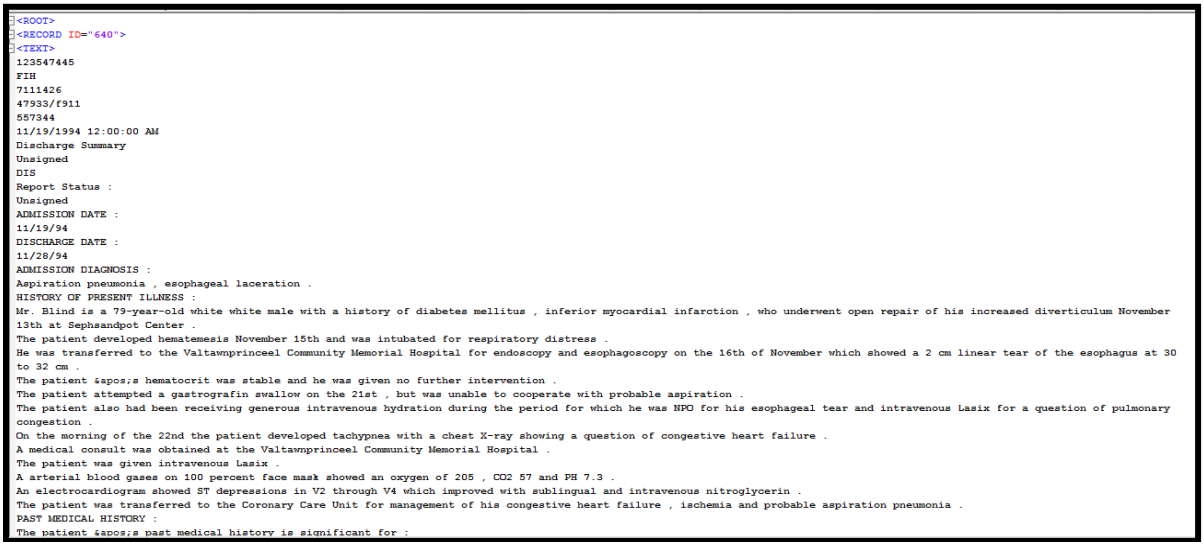

Fig. 4.2 Raw i2b2 data set

The raw i2b2 data set includes 889 medical records in an unstructured format as a single file as shown in the figure above. To efficiently analyze and evaluate these records they need to be split using XML parser with only the content of the 'text' section to be considered for processing. There were two main reasons for splitting the data set:

- cTAKES Aggregate Plaintext UMLS Processor was unable to parse the entire data set as a single record. Further with every record the cTAKES Processor generated an annotation result file of minimum size 100 KB and maximum size 7 MB. Since while computing clinical scores we needed both the original medical record and the annotation results file processing such huge files was slowing the entire computation.
- Further one of the current limitations is the time taken by the semantic web reasoner to infer additional patient facts based on ontology (900 KB), clinical rules and the medical record as input. A single medical record file of size 3 MB slows down the reasoner considerably.

4.2 Processing a discharge summary

4.2.1 cTAKES Aggregate Plaintext UMLS Processor

Every individual record is then annotated using cTAKES Aggregate Plaintext UMLS Processor which includes sentence detection, parts of speech tagging, chunking, UMLS Named Entity recognition, context detection and negation detection. The dependency between various cTAKES components is as follows:

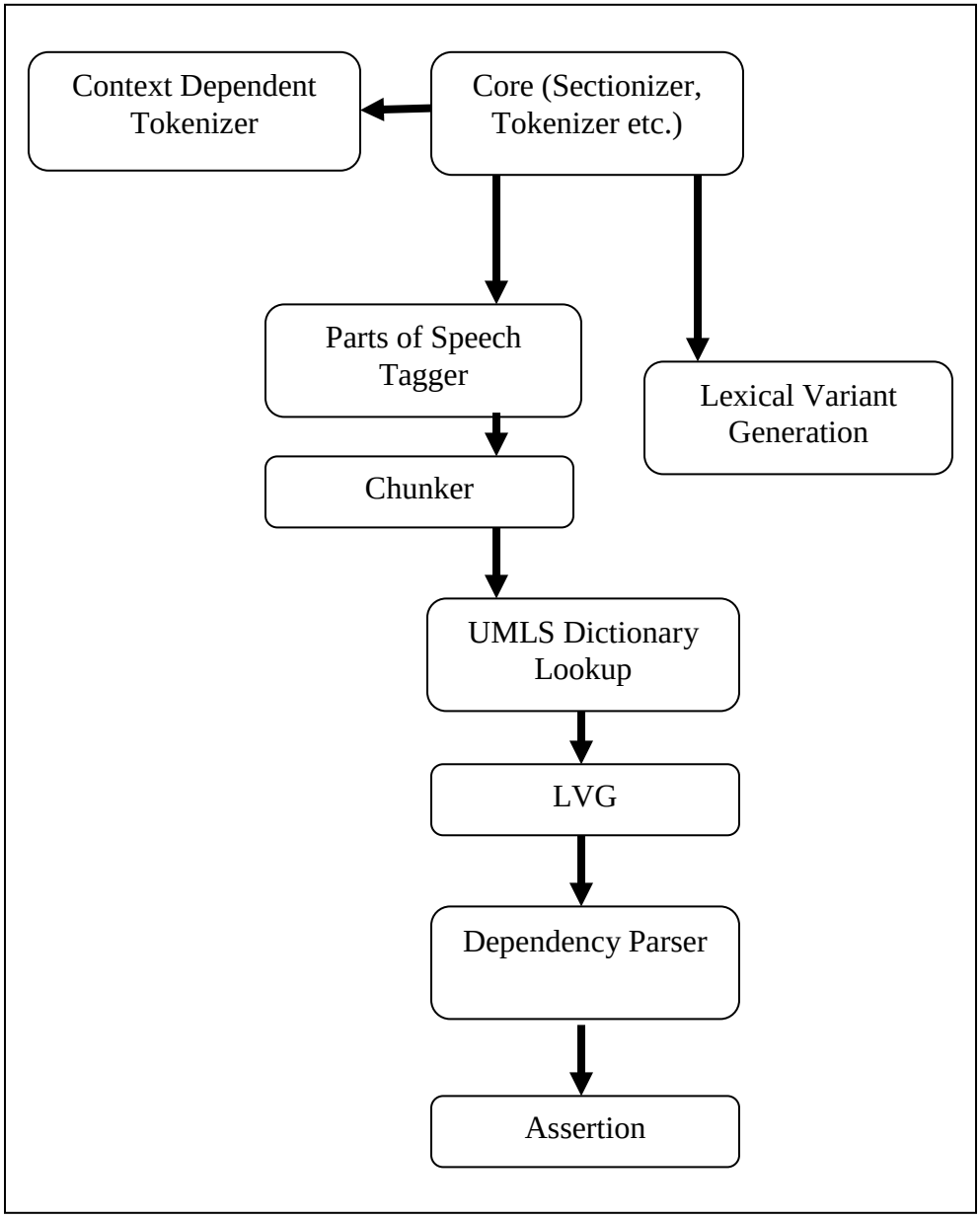

Fig. 4.3 cTAKES Component Dependencies

Each record goes through the following stages:

- Core Processor: This includes a sentence detector and tokenizer. Sentences are detected using end-of-line characters. The cTAKES sentence detector is a wrapper around OpenNLP sentence detector.
- Context Dependent Tokenizer: This is used to create numeric and measurement annotations. It combines one or more tokens and creates a new number token or measurement token. For instance tokens '1', 'mm' are combined to annotate the text '1mm' of type measurement.

- Parts of Speech Tagger: The cTAKES POS tagger wraps around the OpenNLP parts of speech tagger. The tagger has been specially trained on medical resources such as Mayo parts-of-speech corpus, GENIA and Penn Treebank.
- Chunker: Similar to normal chunkers, this one does shallow parsing and identifies noun and verb phrases from a sentence. The cTAKES Chunker is an UIMA wrapper around the OpenNLP chunker.
- UMLS Dictionary Lookup: This annotator finds terms from UMLS Metathesaurus that match with the terms from text. The annotator looks for both exact matches as well as canonical forms of words.
- Lexical Variant Generator: LVG generates canonical forms of words. The cTAKES LVG wraps the NLM Specialist Lexical and uses the Penn Treebank tags.
- Dependency Parser: The dependency parser finds dependencies between individual tokens. For instance, in the phrase 'hormone replacement therapy' it annotates that the token 'hormone' depends on 'replacement' and the token 'replacement' depends on 'therapy'.
- Assertion: The assertion annotator identifies whether a clinical entity is negated or uncertain or conditionally present in the document. A polarity of +1 implies that the term is asserted positively and polarity of -1 implies that the entity is negated. For instance in the sentence, 'the patient denied breathlessness', the polarity of token 'breathlessness' is '-1'.

The figure below shows how a single sentence 'Family History of Obesity but no family history of coronary artery disease' is processed by cTAKES Aggregate Plaintext UMLS Processor.

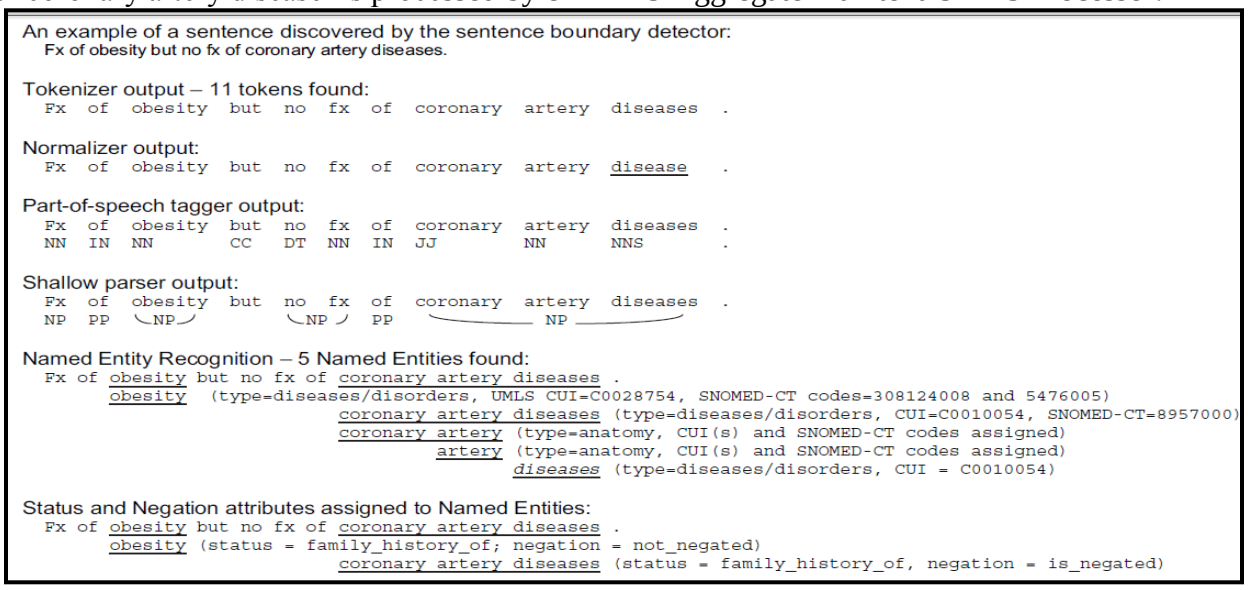

Fig. 4.4 A single sentence parsed by cTAKES Aggregate UMLS Document Processor

The result of cTAKES Aggregate Plaintext UMLS Processor is a single file with all the annotations such as Sentence, Clinical Entity, Number Token, Measurement, Medical Concept, Date, Roman Numeral, Fraction, Medication Event, Lookup Windows, Noun Phrases, Punctuation, Word Token and Verb Phrase. The following figure shows an annotation results file viewed in UIMA Annotation Viewer. UIMA Annotation Viewer allows seeing single or multiple annotations associated with every token. The section on the left allows expanding every annotation and viewing the associated attributes and their values. In the figure below the Entity Mention and Concept annotations have been highlighted for the given document. 'Hypotension'

is a medical entity as well as a medical concept. The Entity annotation gives the UMLS Identifier and the polarity of the token ‘hypotension’ whereas the Concept annotation asserts that ‘hypotension’ is a UMLS Medical concept.

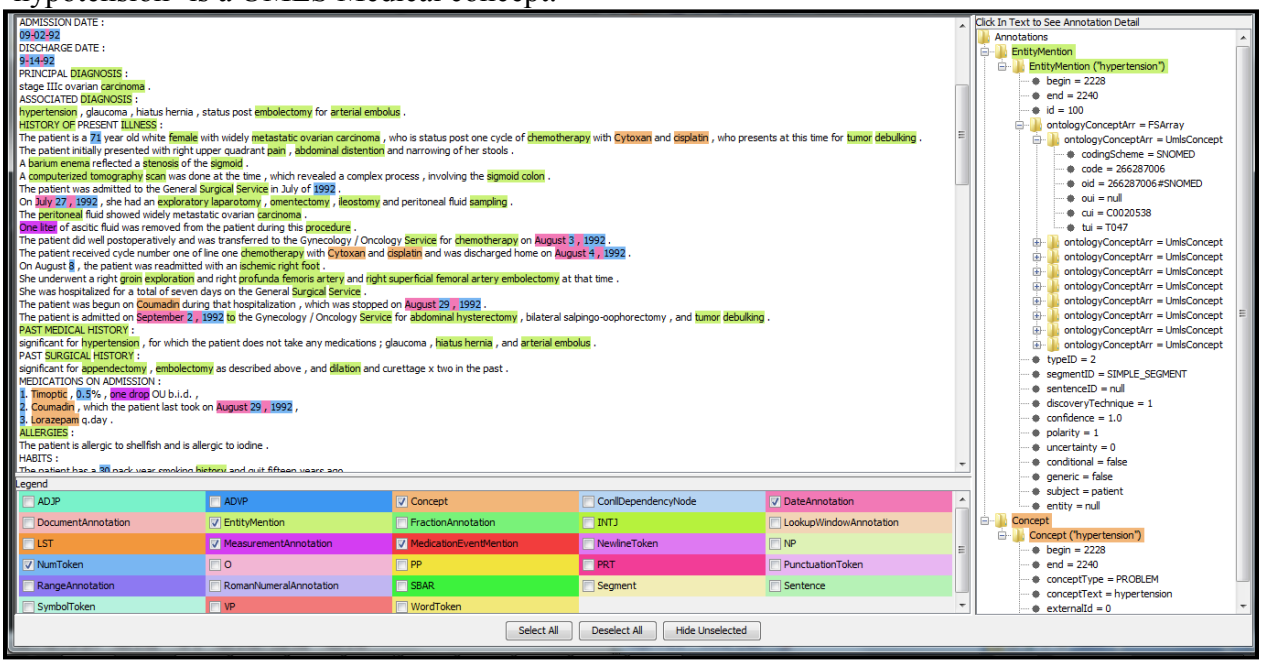

Fig 4.5 cTAKES annotation results file in UIMA Annotation Viewer.

4.2.2 UIMA Collection Processing Engine

UIMA CPE allows processing a set of records using cTAKES Aggregate UMLS Plaintext Processor. The inputs to the UIMA CPE are a File System Collection Reader that parses XML files, cTAKES Aggregate Plaintext Processor to process medical records and XMI CAS Consumer that outputs the results of Analysis Engine in XMI format. An analysis engine is an annotator that analyses documents and infers information from them. For instance in the above screenshot, ‘chemotherapy’ is a medical entity annotation over the span of text ‘chemotherapy’ and its span in document is from 2229 to 2240.

4.3 Information Extraction Phase: The information extraction phase involves parsing annotation results in XMI format produced in previous step and combining them to extract a clinical score parameter. Section Tagging of clinical documents is also done in this phase.

4.3.1 Parsing cTAKES annotations result

XMI Writer CAS Consumer outputs the annotation results in XMI format. Theses XMI annotation files were parsed using XML Dom Parser. Relevant attributes of the annotations were extracted such as polarity from Entity Mention Annotation, spans of tokens, concept Text attribute from Concept Annotation to be used in the next step i.e. extracting clinical score parameters. A sample annotation result file viewed using XML Grid is shown in the screenshot below. The Concept annotation has been expanded. The attributes begin and end corresponds to the span of the token in the document. The spans are mainly used while combining attributes from different annotations. The attribute concept Type corresponds to UMLS Semantic type of the clinical term.

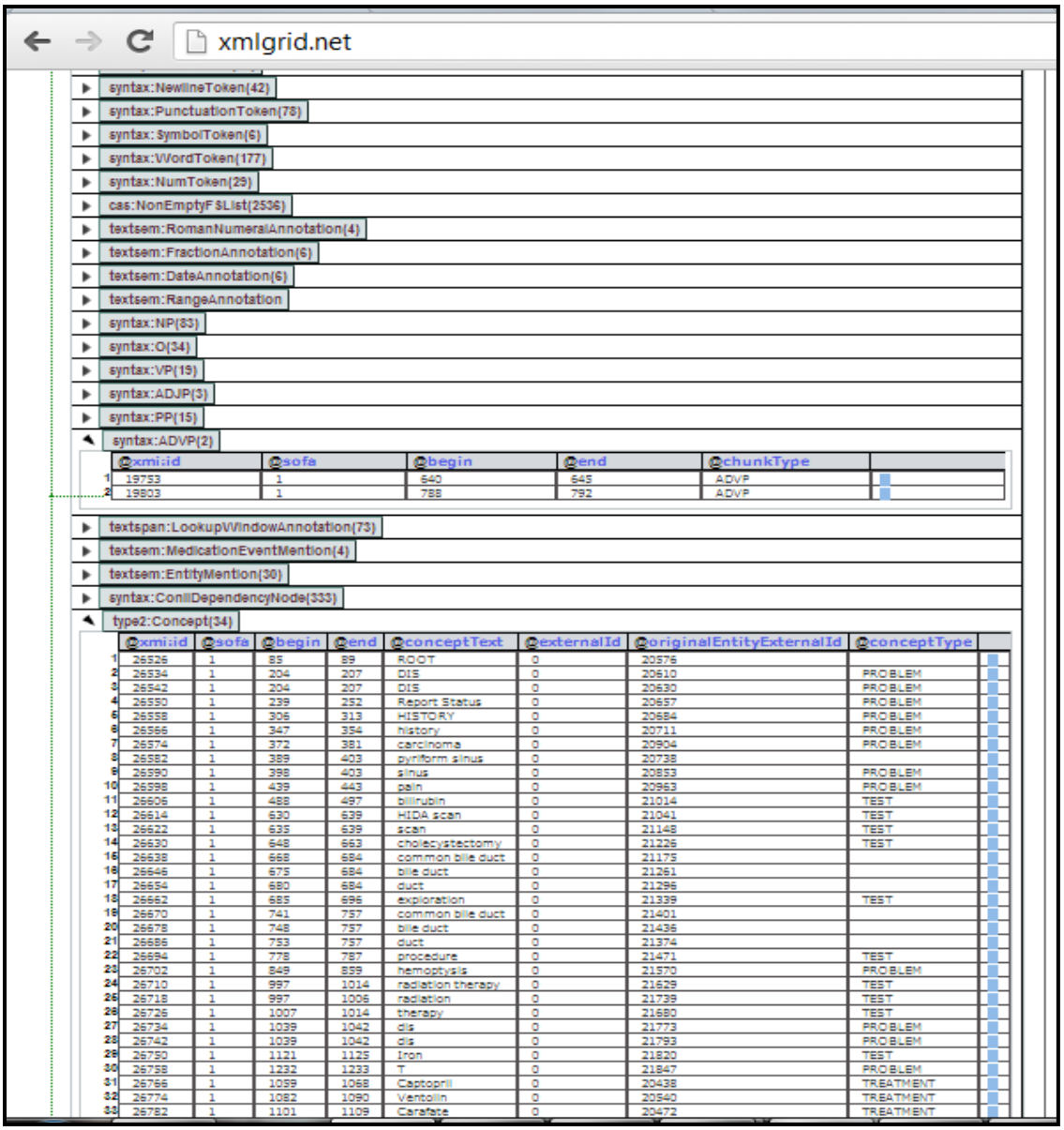

Fig. 4.6 Concept and Verb Phrase Annotation viewed in XMLGrid

The following screenshot shows the attributes for Sentence and Numeric Token Annotation. The annotations only include the span of the text that has been annotated and hence you need the original input file to get the content covered by those spans.

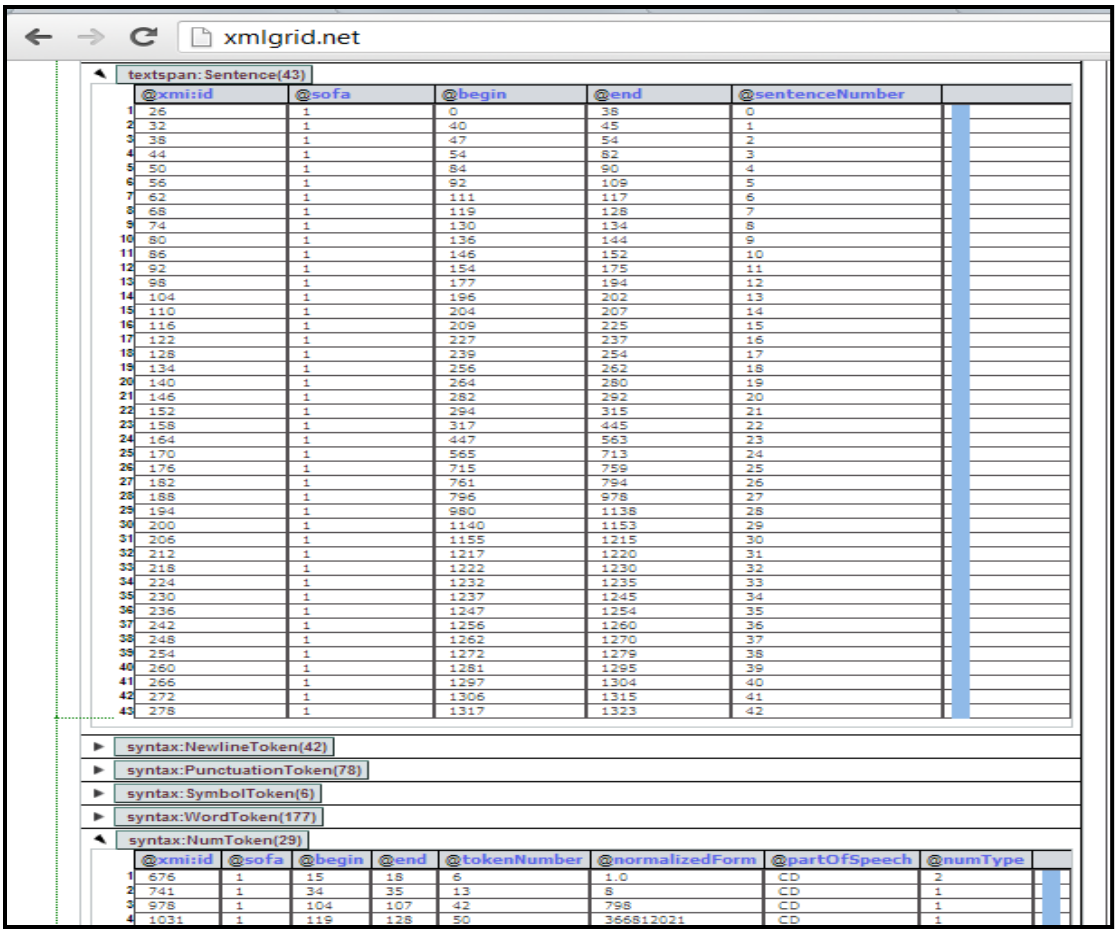

Fig. 4.7 Sentence and NumToken Annotation viewed in XML Gridnet

4.3.2 Extracting Clinical Score Parameters

This section explains how different annotations and their attributes are used to extract a single clinical parameter.

4.3.2.1 Example 1 - Age:

A clinical score parameter such as age of the patient is extracted by combining the results of Section Tagger, Sentence Annotation and Number Token Annotation. The following figure shows how a single sentence is analyzed by each of the above annotators and how results are merged to extract the age of patient.

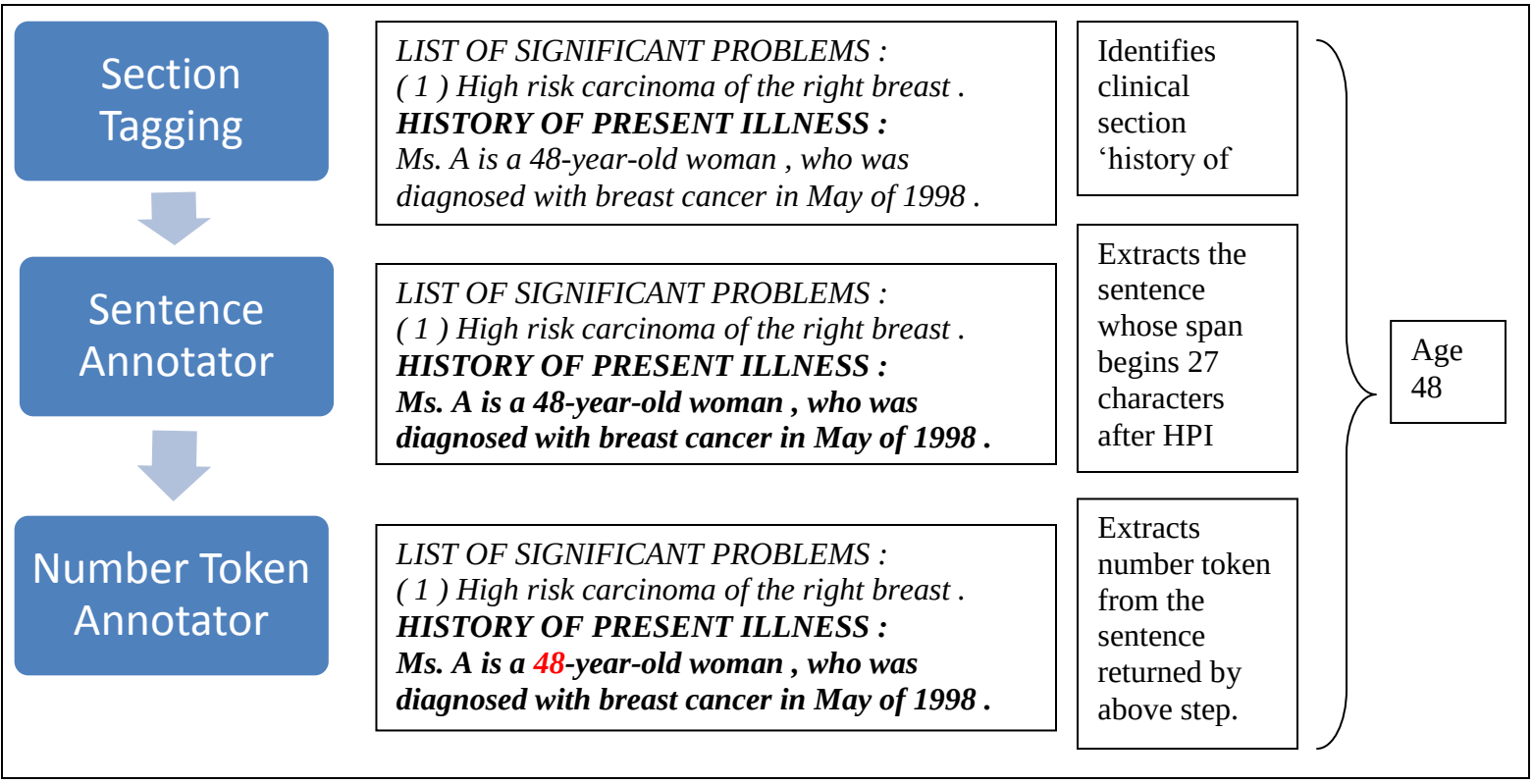

Fig. 4.8 Extracting Age of Patient using cTAKES Annotators and Section Tagger

4.3.2.2 Example 2 – Systolic Blood Pressure Value

The clinical parameter systolic blood pressure value is extracted by combining Dictionary Lookup Annotator, Sentence Annotator and Number Token Annotator.

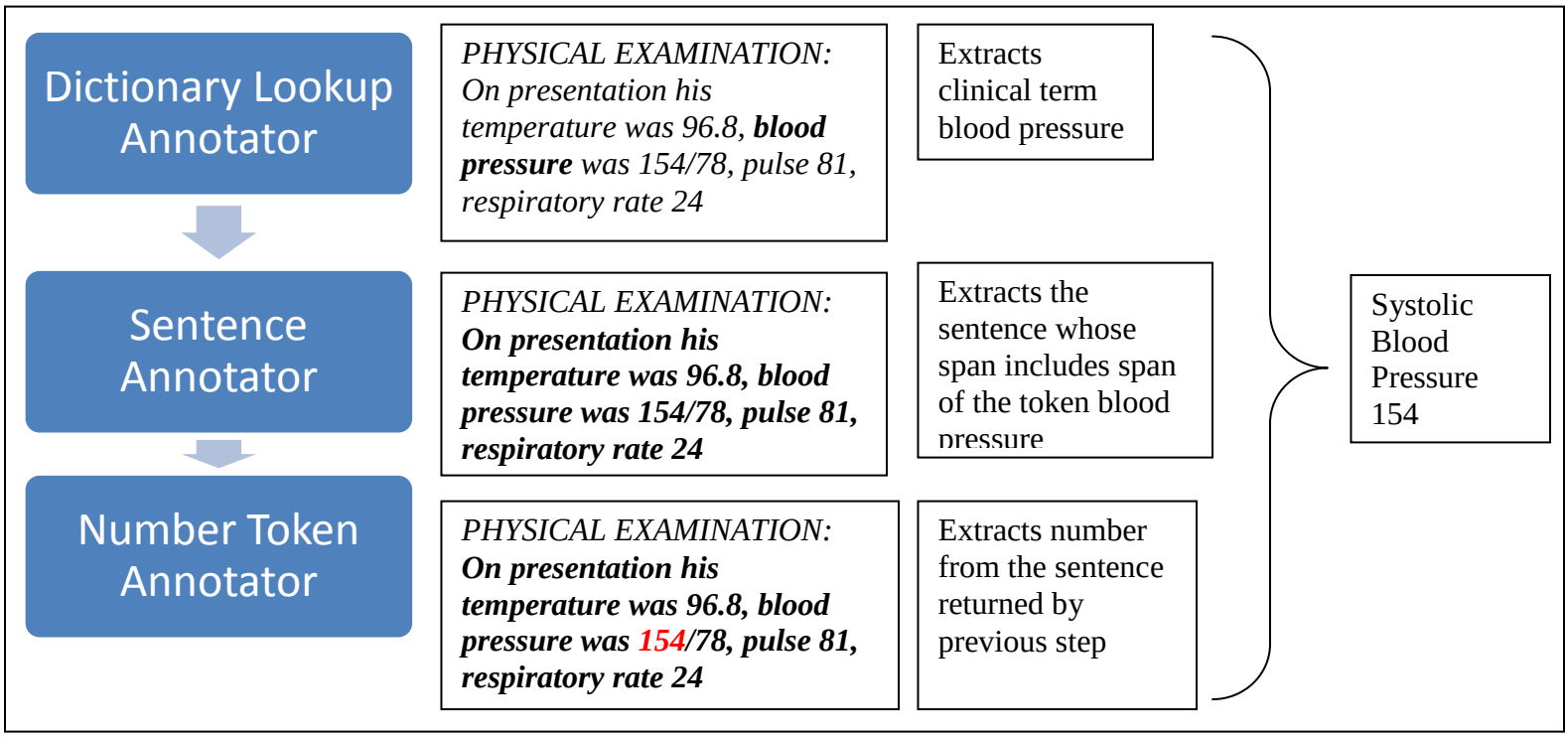

Fig. 4.9 Extracting the value of Systolic Blood Pressure using cTAKES Annotators

The following screenshots would give a better idea regarding how multiple annotations associated with the same span of text can be merged to extract clinically relevant entities.
In the first screenshot all the medical concepts have been highlighted and the attributes associated with the concept blood pressure are shown. In the second screenshot, all the sentence annotations have been highlighted. The sentence which includes the clinical term 'blood pressure' is expanded. The last screenshot shows all number tokens in the given document.

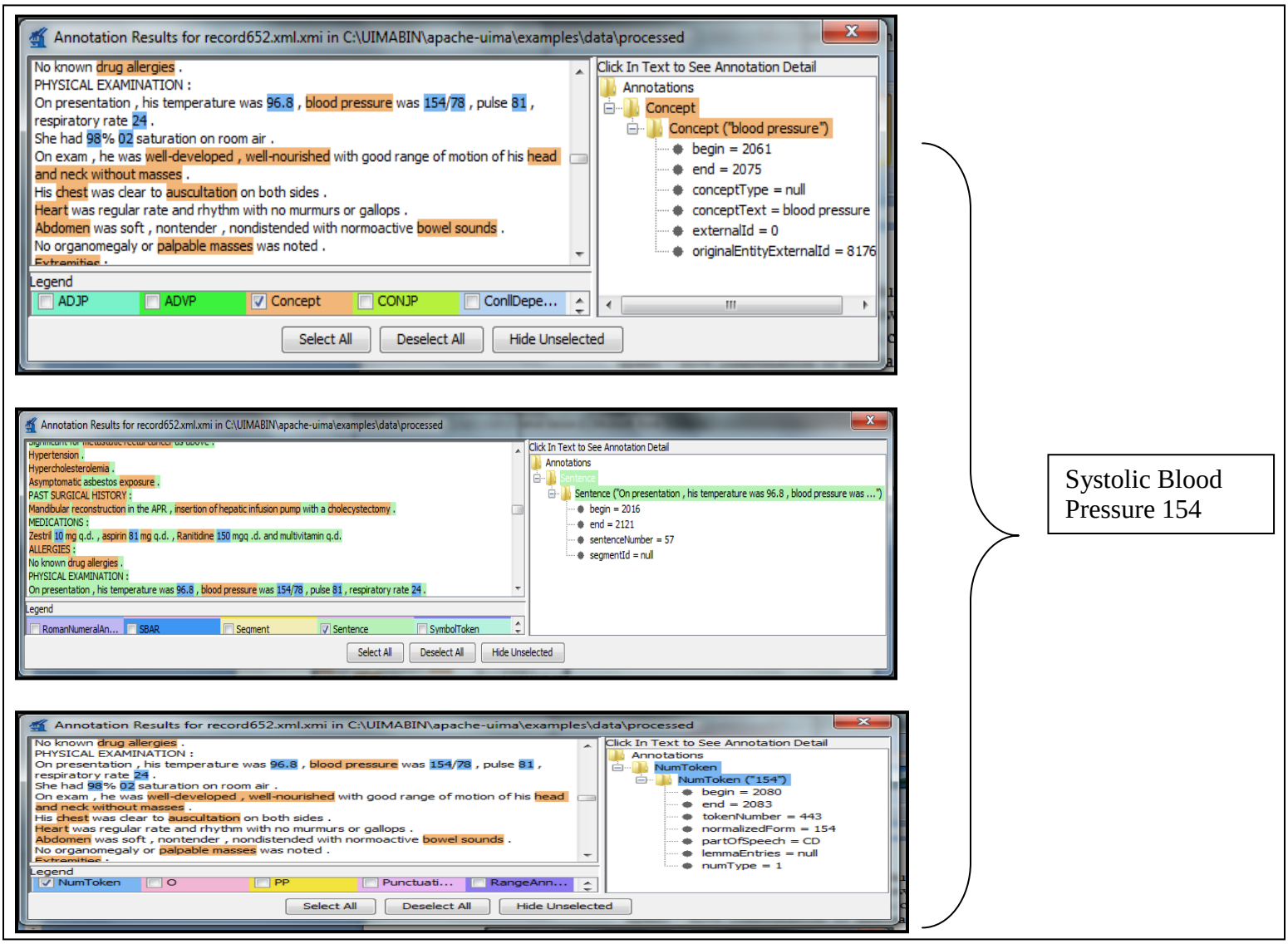

Fig. 4.10 Extracting value of Systolic Blood Pressure viewed in UIMA Annotation Viewer

Thus the results of annotations are combined to determine the value of Systolic Blood Pressure. The Dictionary Lookup Annotator returns blood pressure as a medical concept, the Sentence Annotator returns the span of sentence than contains the token 'blood pressure' and the Number Token Annotator returns all numeric entities from the extracted sentence. The spans help in eliminating false positives.

4.3.2.3 Example 3 – Whether Aspirin was administered in past 7 days

A clinical parameter such as 'whether the medication aspirin was administered to patient in past seven days' is extracted using section tagging and dictionary lookup annotator. The section tagger helps to identify whether the medication was suggested to be taken after discharge, or had been taken recently or was administered in past.

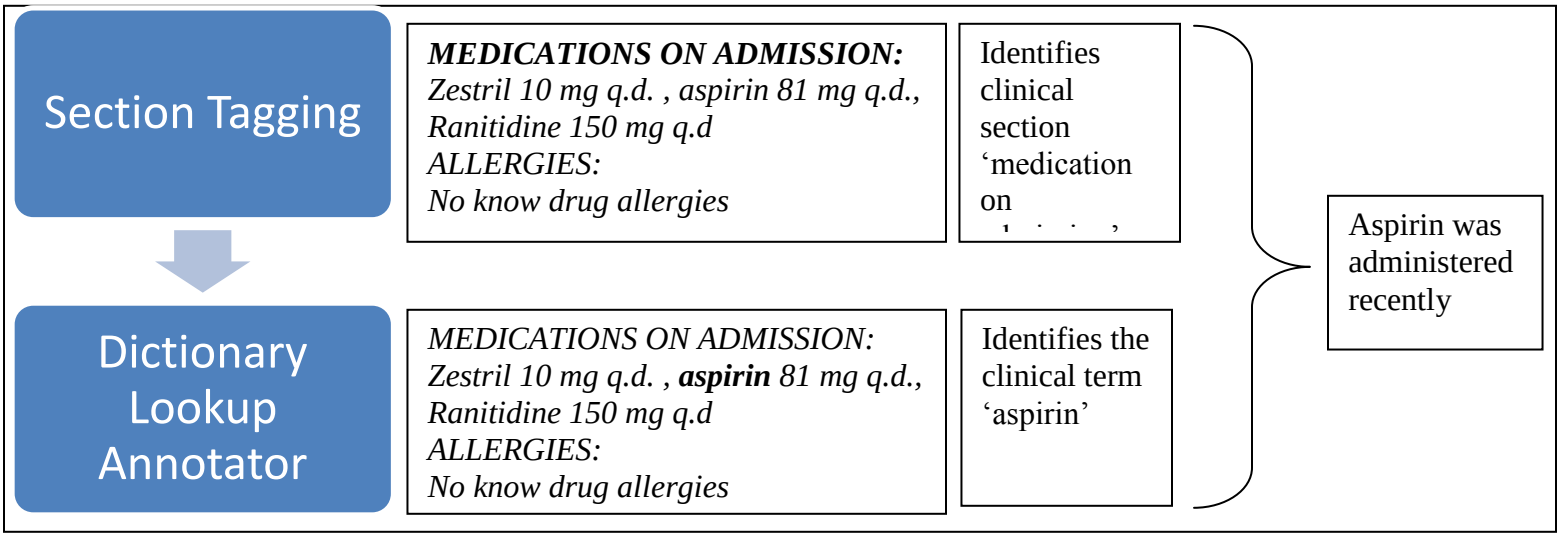

Fig. 4.11 Extracting whether aspirin was administered recently to patient using cTAKES Annotators and Section Tagger

4.3.2.4 Example 4 – 'Shortness of breath history'

The clinical parameter 'shortness of breath history' is extracted by combining Entity Mention Annotator and the Concept Annotator. The polarity attribute of Entity mention annotator determines whether the clinical term was negated in the given sentence. The Concept Annotator internally uses UMLS Dictionary Lookup Annotator.

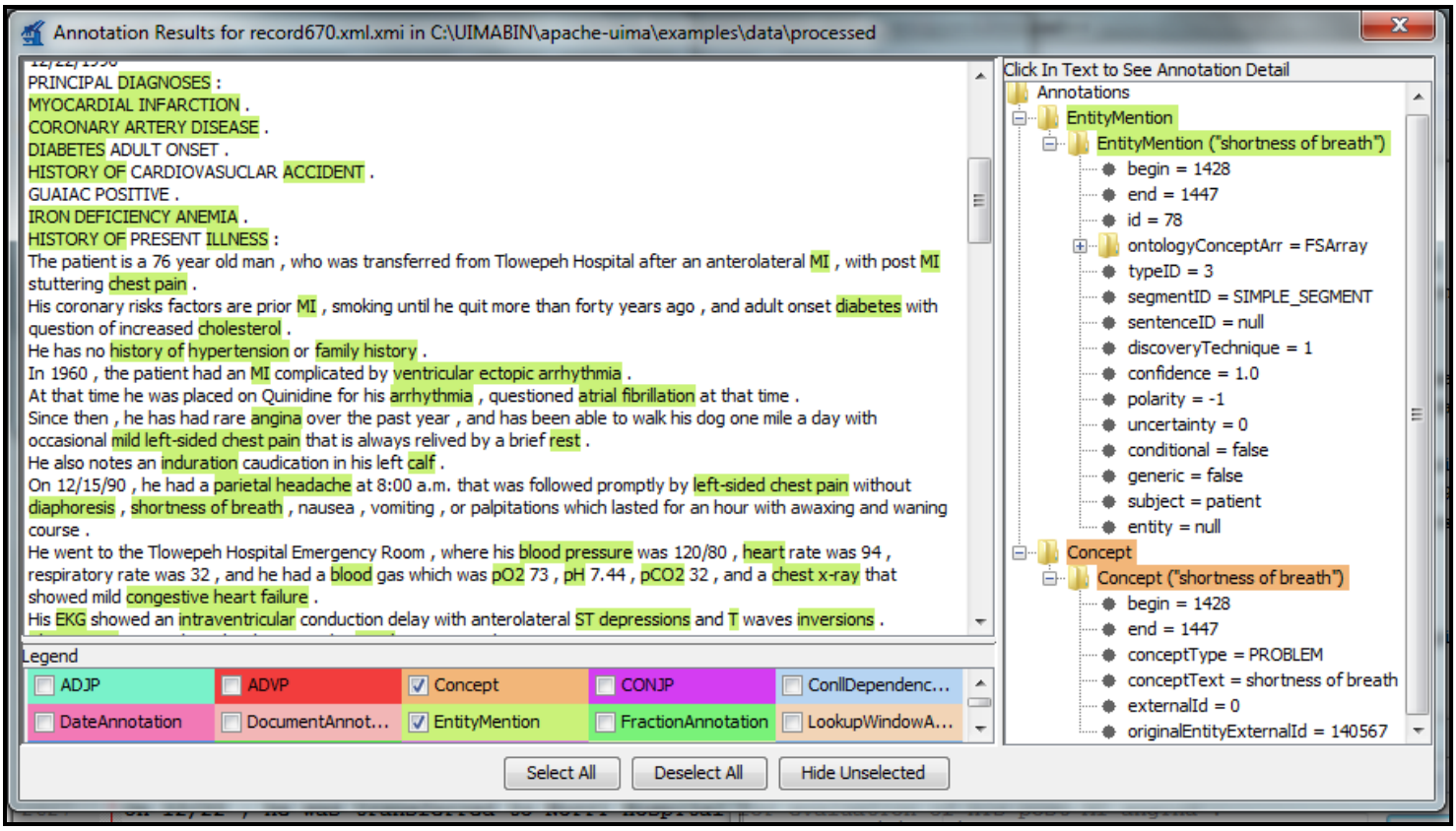

Fig. 4.12 Determining whether shortness of breath symptom was present or absent in patient

As shown in the figure above all the Clinical Entity Mention Annotations and the Clinical Concept Annotations have been highlighted. 'Shortness of breath' is a clinical term occurring in the section history of present illness with polarity = -1. This implies that the patient had a history of breathlessness.

4.4 Text Analytics to extract TIMI and San Francisco Syncope Score Parameters

Similarly other clinical parameters needed to compute the TIMI and San Francisco Syncope score are extracted as shown in the table below.

Clinical Parameter	Sample Sentence	Text Analytics	Result of IE phase
Age > 65	HISTORY OF PRESENT ILLNESS Mr. Blind is a 79-year-old white male with a history of diabetes mellitus. Age : 82y	Section Tagging Numeric Annotation Sentence Annotation Regex	Age – 65 Age – 82
Hypertension	She has a history of hypertension	Dictionary Lookup Annotator Assertion Annotator	CAD Risk Factor – Hypertension Present in the patient

	He has no history of hypertension or family history.		CAD Risk Factor Hypertension not found.
Diabetes	There was no cardiac disease, hypertension or diabetes mellitus.	Dictionary Lookup Assertion Annotator UMLS Synonyms Lexical Variants	CAD Risk Factor – Diabetes Mellitus Absent
Hypercholesterolemia	PAST MEDICAL HISTORY : Ovarian cancer Non-insulin dependent diabetes mellitus Hypercholesterolemia	Dictionary Lookup Assertion Annotator UMLS Synonym	CAD Risk Factor – Hypercholesterolemia present
Current Smoker	SOCIAL HISTORY : He smoked a pack and a half for five years. He is not a current smoker. HISTORY OF PRESENT ILLNESS : He is a heavy smoker and drinks 2-3 shots per day at times	Section Tagging Dictionary Lookup Assertion Annotator	Patient is not an active smoker. Patient is Current Smoker.
Family history of CAD	FAMILY HISTORY : Not obtained. FAMILY HISTORY : The patient reports no family history of cancer. The father of died of coronary artery disease.	Section Tagging Dictionary Lookup Assertion Annotator	Negative Family History for CAD Positive Family History for CAD.
Known CAD (Stenosis >= 50 %)	Severe right external iliac stenosis was noted. There was a 30% stenosis of the main left coronary artery. His circumflex had a 50% obstruction prior to the obtuse marginal branch with a 60%	Dictionary Lookup Numeric Token Annotator Lookup Window With stenosis lookup window was found more effective than sentence annotator since there would be multiple numeric tokens in the same	No numeric term found in the window of term ‘stenosis’ Stenosis = 30 %

	stenosis of the obtuse marginal branch	sentence.	
Aspirin use in past 7 days	MEDICATIONS ON ADMISSION : The patient ' s medications at home were Micronase, Isordil and aspirin. HOSPITAL COURSE: His aspirin was discontinued. DISCHARGE MEDICATIONS : enteric coated aspirin 325 milligrams p.o. q.d.	Dictionary Lookup Annotator Section Tagging Assertion Annotator	Patient has been on aspirin dosage recently. Patient has not taken aspirin in past few days. Patient has been suggested to take aspirin.
Severe Angina (>= 2 episodes in past 24 hrs)	He has had no recurrent heart angina pectoris or heart failure . HISTORY OF PRESENT ILLNESS : Mr. Steve was admitted to Medical Center with three vessel coronary disease after having been admitted there on 11/17/98 for CHF and unstable angina .	Clinical terms that indicate unstable angina or multiple episodes of angina were identified. Dictionary Lookup Section Tagging	Patient did not have severe angina episode recently. Patient recently had multiple episodes of angina as indicated by the terms unstable angina.
ST changes >= 0.5 mm	The patient in September of 1993 had 3-4 millimeter ST segment elevation in V1 through V4.	Dictionary Lookup Measurement Annotator Sentence Annotator Synonyms Abbreviations Canonical forms With ST changes it was observed that the measurement values don’t always lie in the	ST changes = 3 mm ST changes = 4 mm

		lookup window of the term 'ST segment'. Hence sentence annotation and distance of the numeric token from the term is considered.	
Cardiac Markers – Troponin	The Troponin I was less than .4 and the creatine kinase was 114 .	Dictionary Lookup Annotator Assertion Annotator	Patient has taken troponin test which indicates elevated cardiac markers.
CHF History	PAST MEDICAL HISTORY : congestive heart failure	Dictionary lookup Abbreviations Synonyms Assertion	Patient has a positive CHF history
Hematocrit < 30 %	Hematocrit was 46.1	Dictionary Lookup Number Token Annotator Lookup Window	Hematocrit level = 46.1
ECG Abnormal	Electrocardiogram shows sinus bradycardia with Digitalis effect .	Clinically terms that indicate abnormal EKG were identified such as sinus bradycardia. Assertion Annotator Dictionary Lookup Annotator Synonyms Abbreviations	Patient has abnormal EKG
Shortness of Breath History	HISTORY OF PRESENT ILLNESS She has no chest pain , shortness of breath or fever .	Dictionary Lookup Annotator Assertion Annotator Synonyms Section Tagging	Patient doesn't have shortness of breath.
Systolic BP < 90mm	PHYSICAL EXAMINATION : pulse 77 , blood pressure 125/71 , respiratory rate 18 .	Dictionary Lookup Annotator Number Token Annotator Lookup Window Sentence Annotator	Systolic BP = 125

Fig. 4.13 Text Analytics required to extract various Clinical Score parameter

Apart from clinical scores we also aim to extract other cardiac facts from a medical record using Assertion, Dictionary Lookup, Number Token, Measurement, Section Tagging and heart failure ontology. This would be explained in the next section.

The data set used in this research is a publicly available data set and includes medical records from different domains such cancer, kidney along with cardiac. The discharge summaries are

not the ideal chest pain progress or admission notes that would have all the clinical data needed to compute above scores. Missing values was one of the main challenges while analyzing this data.

4.4 Knowledge Generation

In this phase the clinical parameters extracted from previous step are translated to a structured format using heart failure ontology. We first navigate the ontology identifying relevant properties and then translate the patient facts into RDF format.

4.4.1 Building index of cardiovascular terms using Heart Failure Ontology

In this step the heart failure ontology is navigated to build an index of cardio vascular terms and the properties that co relate the term with the patient class. We first identified properties that whose domain is the Patient class and then build an index of terms that constitute the range of property. All the synonyms, abbreviations, classes and instances were combined to build an exhaustive list of cardiovascular terms.

The following screenshots shows the range of properties which were considered in this work.

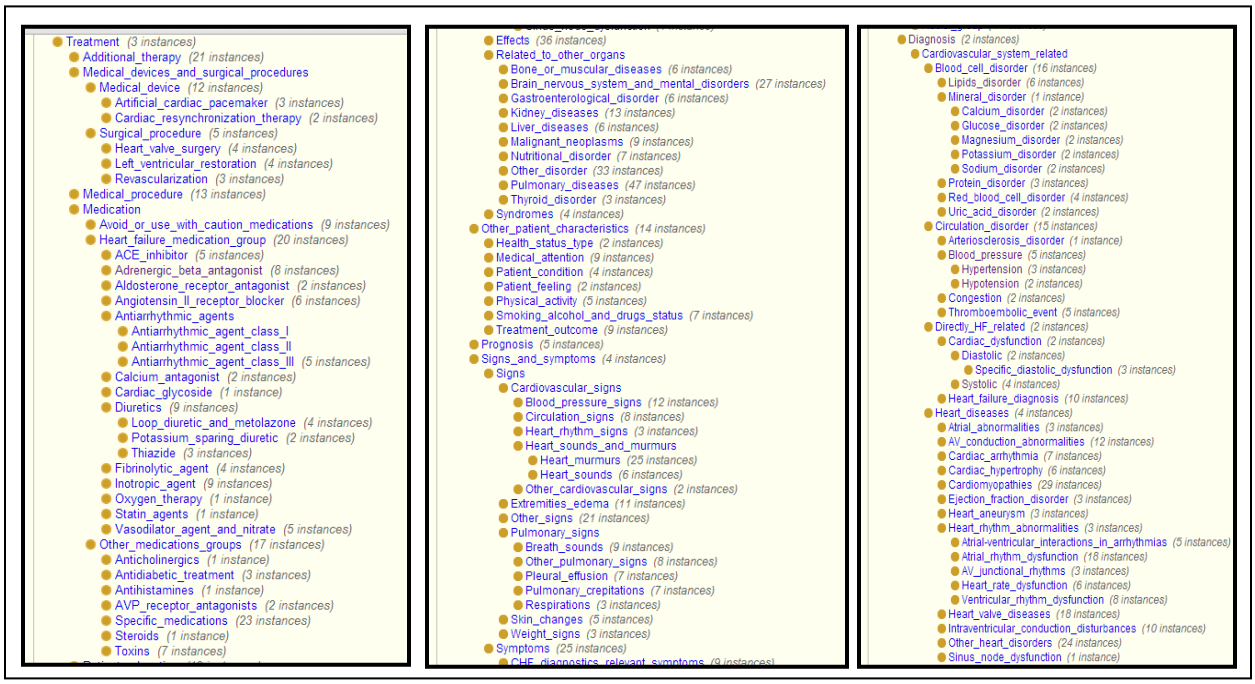

Fig. 4.14 Heart Failure Ontology Important Classes: Treatment, Signs and Symptoms, Diagnosis

For instance consider the property ‘Taken Medication’ whose domain is class Patient and range is class Medication. Hence all subclasses of ‘Medication’ such as Fibronolytic Agent as well as instances such as ‘Aspirin’ and their synonyms such as ‘Acetylsalicylic Acid’ and abbreviations such as ‘ASA’ were added to the range of the property ‘Taken Medication’. Similarly consider another property ‘has Diagnosis’ whose range included classes such as lipids disorder , corresponding instances such as hyperlipidemia and their synonyms such as hypercholesterolemia, high cholesterol levels. All of these terms were added to the domain of the property of ‘Taken Medication’

We identified 19 relevant properties and added ‘has Value’ object property to certain classes. The ‘has Value’ property allows storing a value associated with a clinical concept.

The following figure shows the properties used in this work and the range of every of property.

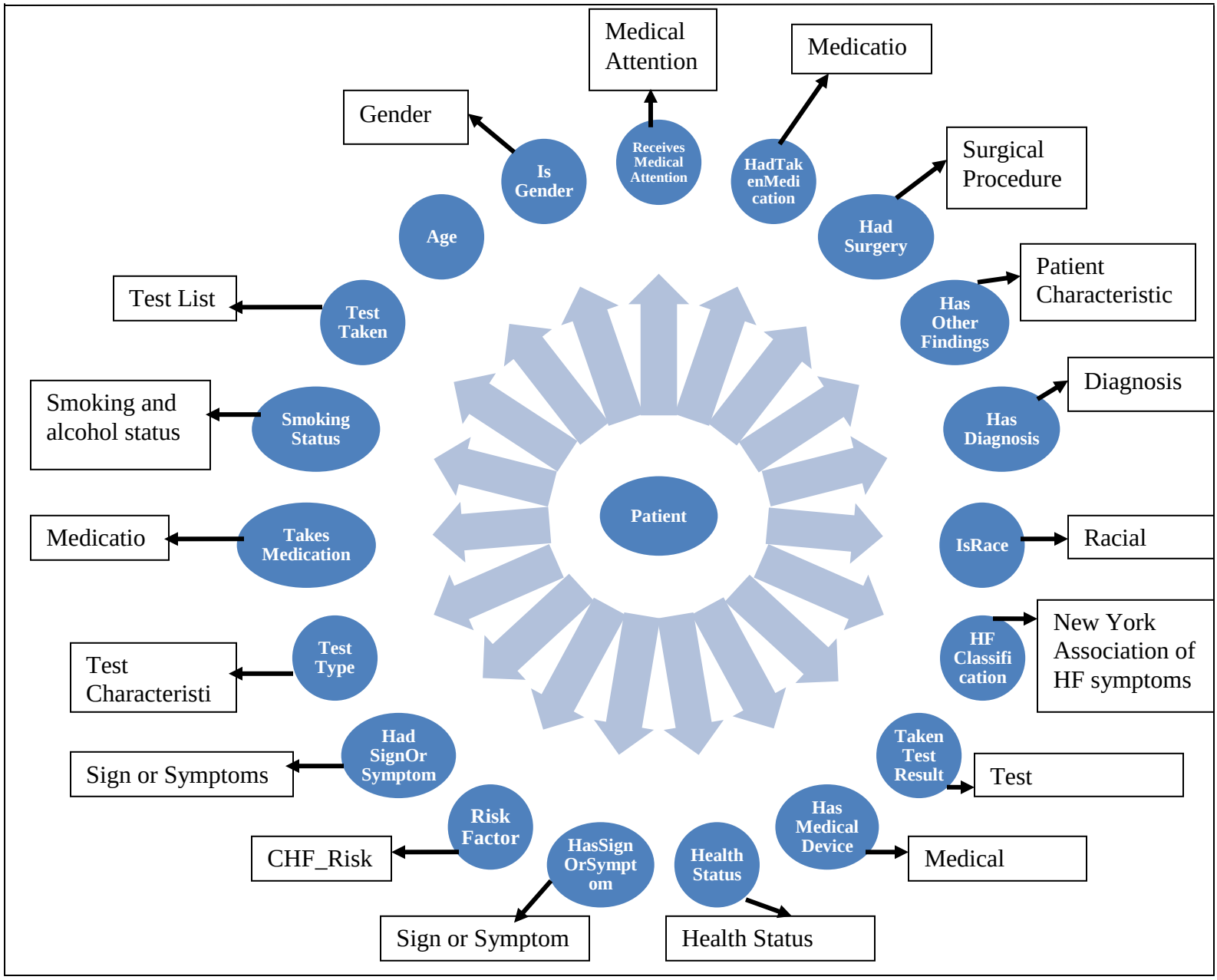

Fig. 4.15 HF Ontology Properties used to populate Patient class

4.4.2 Form RDF triples from clinical parameters

Every clinical parameter extracted or clinical entity returned by information extraction phase is translated to a RDF triple with subject being the instance of Patient class, predicate being one of the above properties and the object being the clinical entity. When required a clinical parameter is split into set of RDF triples. The following table shows a set of patient facts and the corresponding RDF triples followed by the RDF graph for the same.

Fact	Predicate	Object
Patient has taken Troponin Test	Test Taken	S Troponin
Chronic Heart Failure History	Has Diagnosis	Chronic Heart Failure

Shortness of Breath	Has Sign or Symptom	Shortness of Breath
Age 81	Age	81
Hematocrit Level 42	Test Taken	Erythrocyte Count
	Has Value	42
Systolic Blood Pressure 155	Has Diagnosis	Systolic Blood Pressure
	Has Value	155

Fig. 4.16 Patient Facts as RDF Triples

The corresponding RDF graph is as follows:

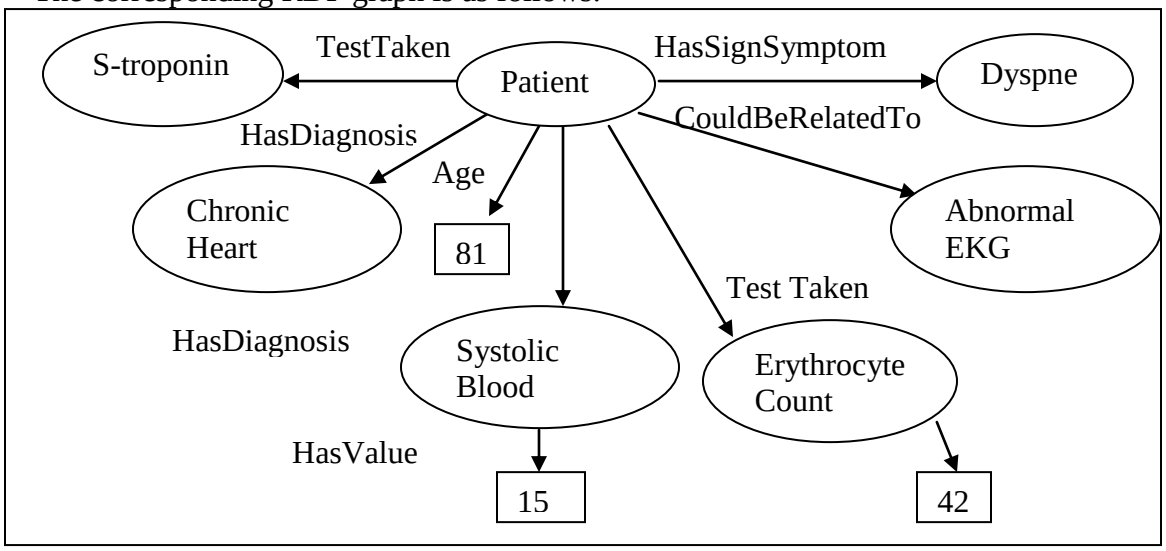

Fig. 4.17 Patient facts represented as RDF graph

Further the following table shows how other cardiovascular facts extracted from the discharge summaries were translated as RDF triples.

Fact	Predicate	Object
Aspiration	Has Diagnosis	Aspiration
Pneumonia	Has Diagnosis	Pneumonia
Male	Is Gender	Male
Diabetes	Has Diagnosis	Diabetes Mellitus
Myocardial Infarction	Has Diagnosis	Myocardial Infarction
Pulmonary Congestion	Has Sign or Symptom	Pulmonary Congestion
Tachypnea	Has Sign or Symptom	Tachypnea
Chest X Ray	Test Taken	Chest X Ray
Heart Failure	Has Diagnosis	Heart Failure
Electrocardiogram	Test Taken	Electrocardiogram
Insulin	Had Taken Medication	Insulin
Morphine	Had Taken Medication	Morphine
Hypotension	Has Diagnosis	Hypotension
Bleeding	Has Diagnosis	Bleeding
Pneumothorax	Has Diagnosis	Pneumothorax
Renal Failure	Has Diagnosis	Renal Failure
Nitroglycerin	Had Taken Medication	Nitroglycerin

Aspirin	Had Taken Medication	Aspirin

Fig. 4.18 Cardiovascular facts as RDF triples

4.4.3 TIMI and Syncope Score RDF Triples

The RDF triples corresponding to the clinical parameters needed to compute TIMI and Syncope scores are as follows:

Clinical Parameter	Parent Class	Predicate	Object
Age – 65		Age	65
CAD Risk Factor – Hypertension Present in the patient	Diagnosis	Has Diagnosis	Hypertension
CAD Risk Factor Diabetes Mellitus	Diagnosis	Has Diagnosis	Diabetes Mellitus
CAD Risk Factor – Hypercholesterolemia present	Diagnosis	Has Diagnosis	Hypercholesterolemia
Patient is Current Smoker.	Smoking alcohol and drugs status	Has Smoking Status	Smoker
Positive Family History for CAD.	Patient Characteristics	Could Be Related To	Family history of CAD
Stenosis = 80 %	Diagnosis	Has Diagnosis	Coronary Heart Failure
Patient has been on aspirin dosage recently.	Medication	Takes Medication	Aspirin
Patient recently had multiple episodes of angina as indicated by the terms unstable angina.	Signs and Symptoms	Has Sign Or Symptom	Unstable Angina
ST changes = 3 mm	Test List	Test Taken	Electrocardiogram at rest
		Measures	ST segment changes
		Has Value	3
Patient has taken troponin test which indicates elevated cardiac markers.	Test List	Test Taken	S-troponin
Patient has a positive CHF history	Diagnosis	Has Diagnosis	Chronic Heart Failure
Hematocrit level = 46.1	Test List	Test Taken	Erythrocyte Count
		Measures	Hematocrit Level
		Has Value	46.1
Patient has abnormal EKG	Patient Characteristic	Could Be Related To	abnormal electrocardiogram

Patient has shortness of breath.	Signs or Symptoms	Has Sign Or Symptom	Dyspnea

Fig. 4.19 Clinical Score Parameters expressed as RDF Triples

4.5 Inference and Query

The RDF graph obtained in above phase can be made richer by applying a set of inference and clinical rules . The resulting knowledge base can then be queried for a variety of clinical scores.

4.5.1 Inference using Clinical Rules

Consider following set of rules:

- Class I : Ontology Inference Properties
  If Class A is subclass of Class B and if Patient Has Diagnosis of A then Patient Has Diagnosis of B as well.
- Class II: Clinical Rules Connecting two Clinical Terms
  Diagnosis of Chronic Heart Failure could be related to Acute Myocardial Infarction or Hypertrophic Cardiomyopathy
  A relevant sign or symptom for Chronic Heart Failure is Dyspnea or Fatigue
  Chronic Heart Failure is commonly accompanied by atrial fibrillation or atrial flutter
- Class III: Complex Clinical Rules combining different clinical terms
  If patient has done echocardiogram test, has low EA ratio and HF symptoms then the suggested diagnoses is diastolic Heart Failure
  A person has systolic heart failure if he has performed echocardiography and has either decreased left ventricular contractility or left ventricular ejection fraction < 40 % and the patient has some HF signs or symptoms

When you apply the above rules to the RDF graph in the previous section, the extended graph is as follows:

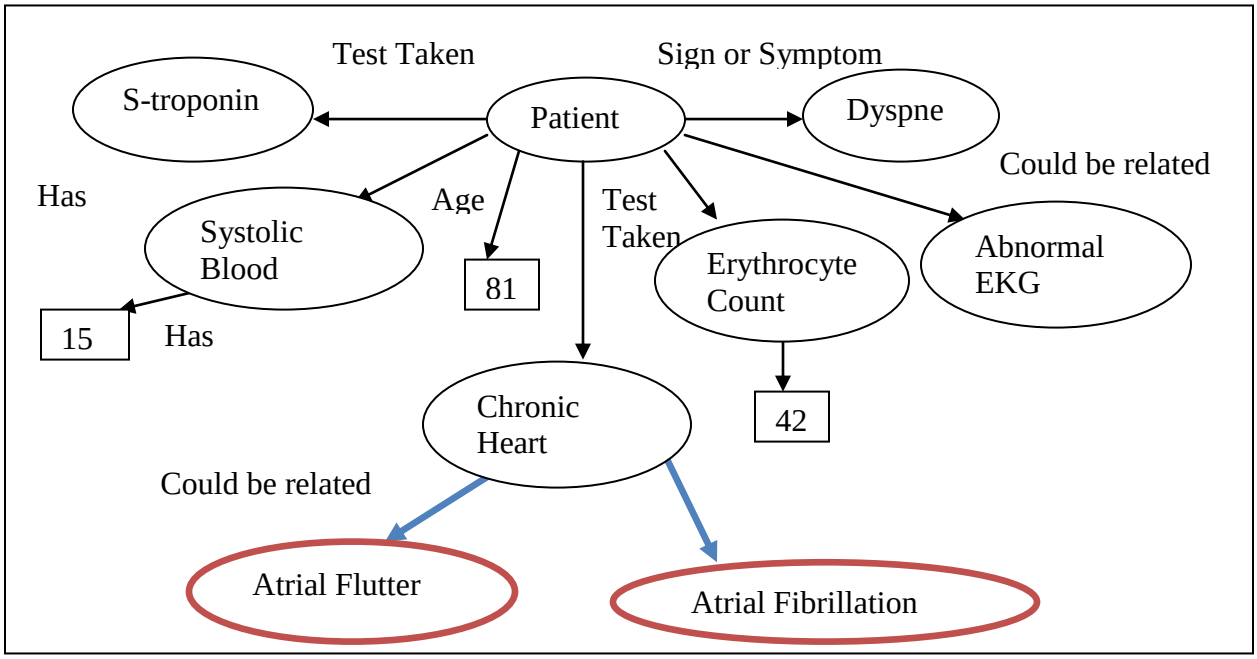

Fig. 4.20 Expansion of RDF graph after applying a clinical rule

The graph extends further with the addition of the following rules:

- Atrial Fibrillation is caused by Hyperthyroidism
- Dyspnea could be related to increases pulmonary capillary wedge.

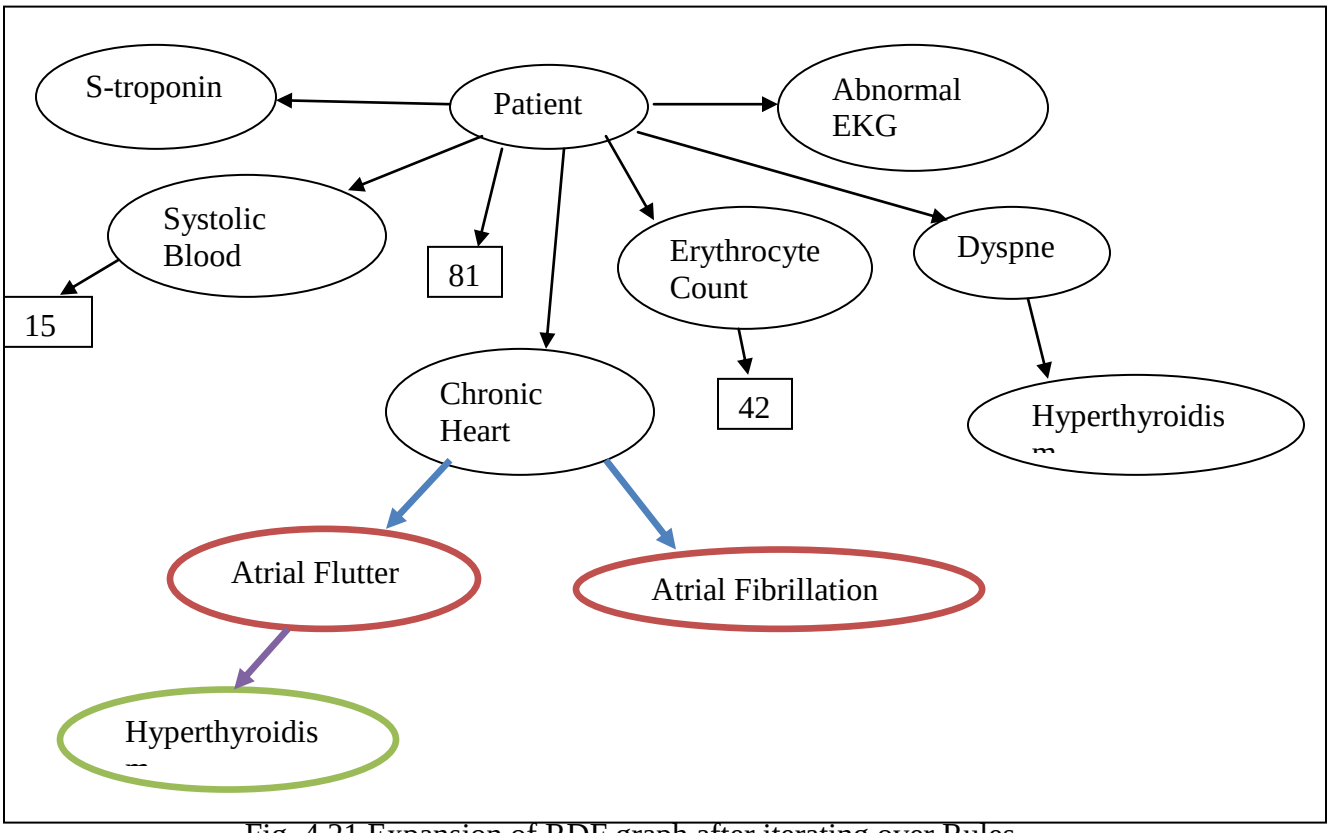

Fig. 4.21 Expansion of RDF graph after iterating over Rules

The above examples demonstrate how the RDF graph of Patient facts keeps expanding as we add more clinical rules to the knowledge base. Applying a chain of clinical rules and using OWL properties we generate a rich knowledge base of patient facts from a given discharge summary.
The following diagram shows how structured knowledge base of patient facts is built using discharge summaries which provide factual data, medical ontologies which define relations between clinical entities , clinical rules and a semantic web reasoner. The knowledge base is then queried for a variety of clinical scores or other clinically relevant physician queries.

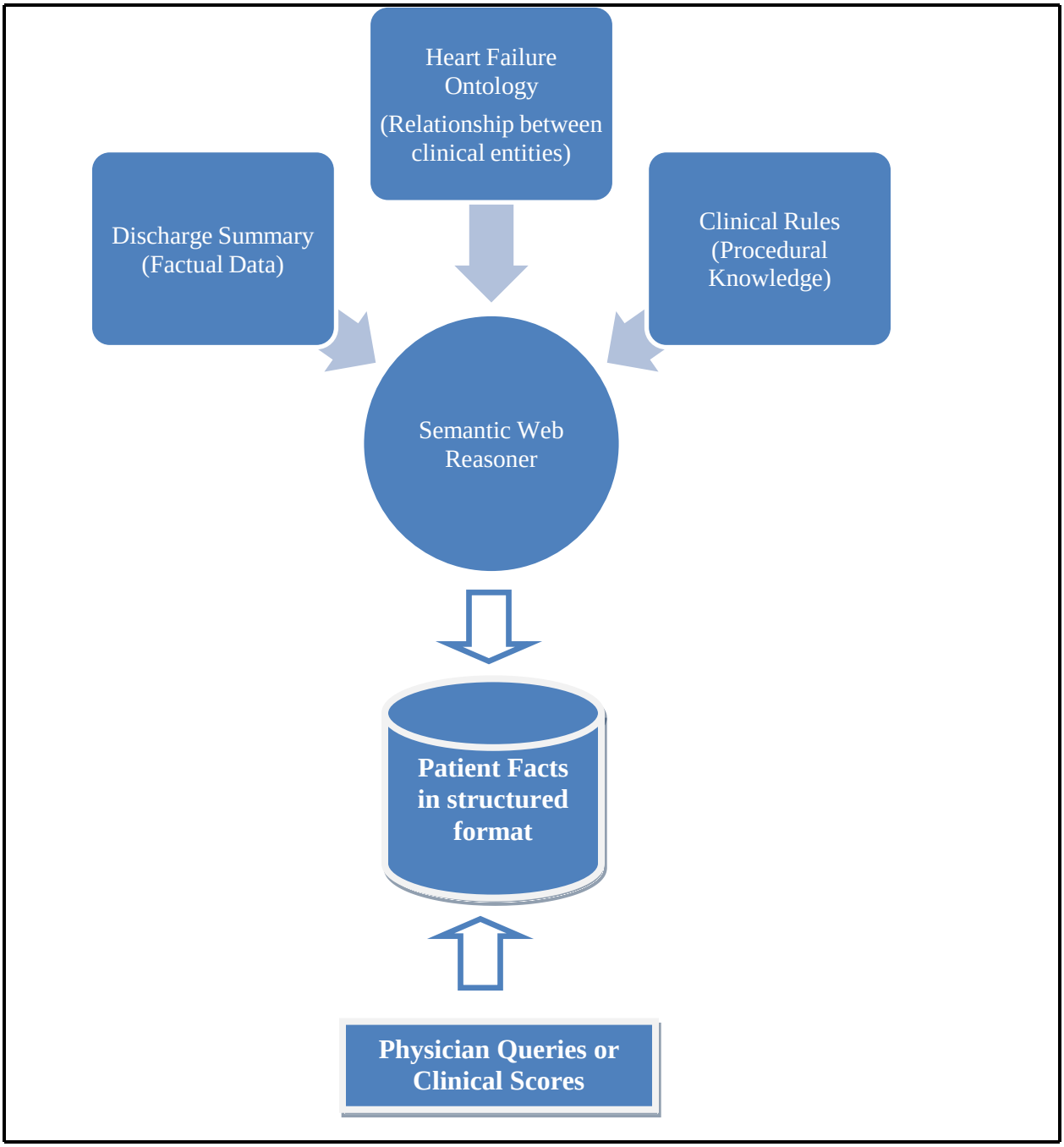

Fig. 4.22 Proposed Clinical Decision Support System

4.5.2 Querying for Clinical Scores

The SPARQL Queries for computing TIMI and Syncope score are as follows:

Clinical Parameter	SPARQL Query
Age	SELECT ?age WHERE { ?patient hf:Age ?age . }
CAD Risk Factor Hypertension	SELECT ?patient WHERE {?patient hf:HasDiagnosis hf:Hypertension.}

CAD Risk Factor Hypercholesterolemia	SELECT ?patient WHERE {?patient hf:HasDiagnosis hf:Hypercholesterolemia.}
CAD Risk Factor Diabetes	SELECT ?patient WHERE {?patient hf:HasDiagnosis hf:Diabetes_Mellitus.}
CAD Risk Factor Current Smoker	SELECT ?patient WHERE {?patient hf:HasSmokingStatus hf:Smoking.}
CAD Risk Factor Family History CAD	SELECT ?patient WHERE {?patient hf:CouldBeRelatedTo hf:FamilyHistoryofCAD.}
Known CAD	SELECT ?patient WHERE {?patient hf:HasDiagnosis hf:Coronary_heart_disease.}
Whether Aspirin was administered recently	SELECT ?patient WHERE {?patient hf:TakesMedication hf:Aspirin.}
Whether >2 angina episodes recently	SELECT ?patient WHERE {?patient hf:HasSignOrSymptom hf:Angina_Pectoris.}
Whether ST segment changes >0.5 mm	SELECT ?value WHERE { hf:ST_segment_changes hf:HasValue ?value .}
Known CHF History	SELECT ?patient WHERE {?patient hf:HasDiagnosis hf:Chronic_Heart_Failure.}
Whether Hematocrit < 30 %	SELECT ?value WHERE { hf:Erythrocyte_Count hf:HasValue ?value .}
Shortness of Breath History	SELECT ?patient WHERE {?patient hf:HasSignOrSymptom hf:Dyspnea.}
Whether Systolic Blood Pressure < 90mm Hg	SELECT ?value WHERE { hf:Systolic_blood_pressure hf:HasValue ?value .}
Abnormal EKG	SELECT ?patient WHERE {?patient hf:CouldBeRelatedTo hf:Abnormal_electrocardiogram.}
Elevated Cardiac markers	SELECT ?patient WHERE {?patient hf:TestTaken hf:S-troponin.}

Fig. 4.23 Sparql Queries to extract clinical score parameters

4.6 Example showing all outputs of all stages and the final Result

Towards the end we present yet another example explaining different stages of system implementation. We refer to the discharge summary presented at the beginning of the chapter.

The first figure shows results of information extraction phase. The second figure shows the RDF graph generated at the end of knowledge generation phase. Lastly we show how the knowledge base can be queried for clinical scores.

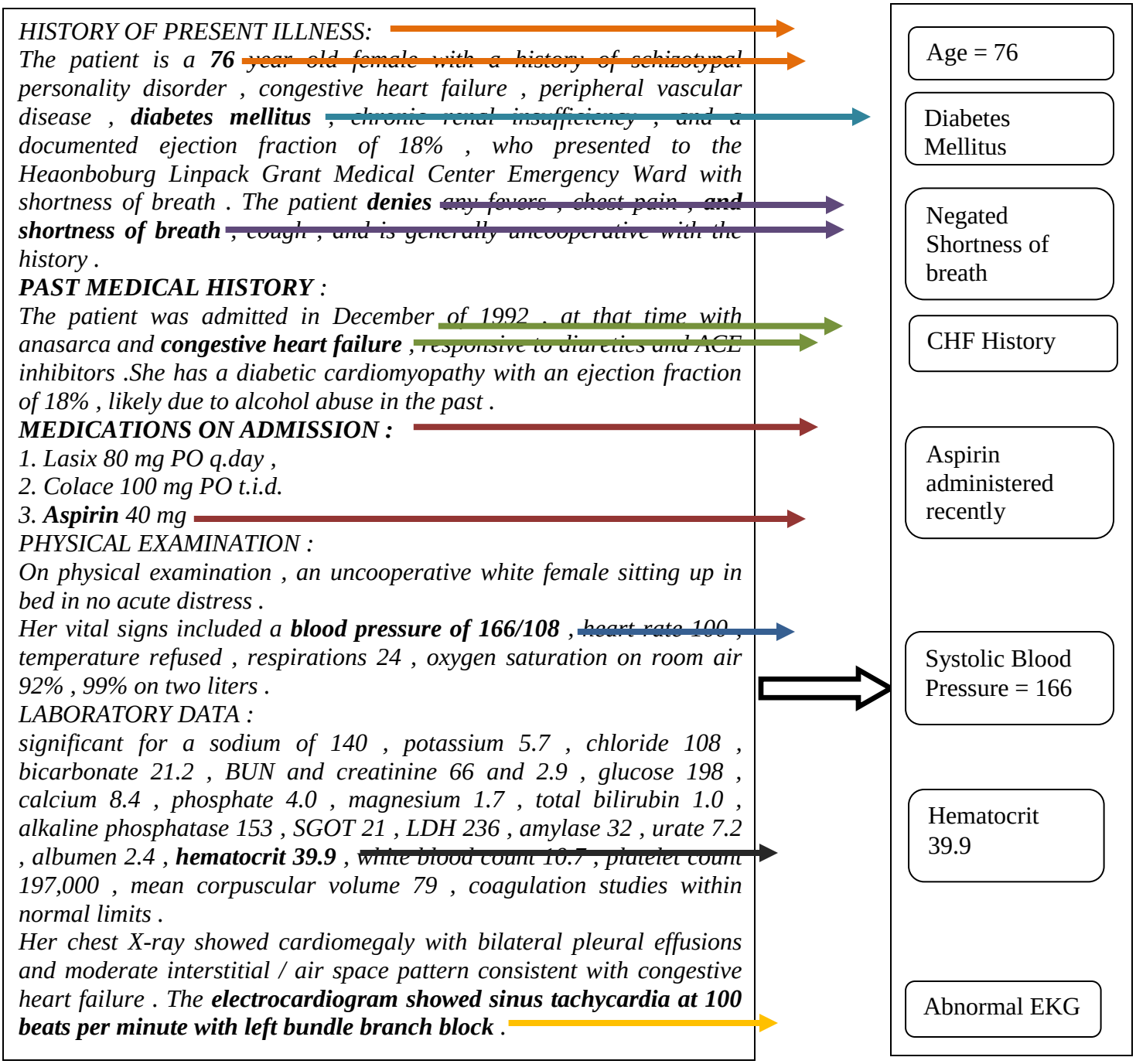

Fig. 4.24 Phase I : Information Extraction from a discharge summary

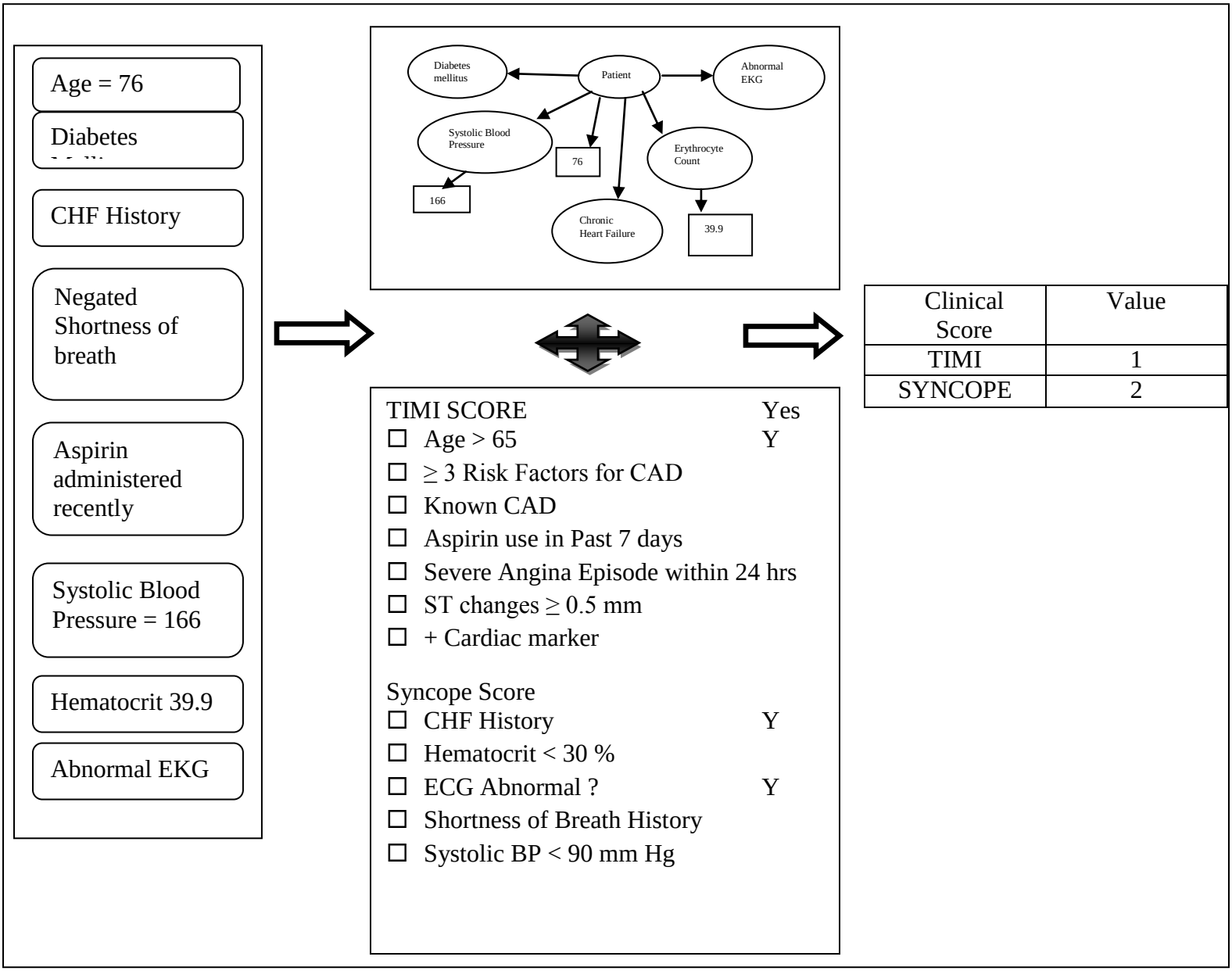

Fig. 4.25 Phase II and III: Knowledge Generation and Querying

Chapter 5
EVALUATION

In this chapter we explain the performance measures used to evaluate our work. We also explain the nature and the limitations posed by the data set. We discuss the results of evaluation, analyzing which clinical parameters were difficult to capture and subsequent enhancements made to extract those parameters.

5.1 Performance Measures

Following performance measures were used to evaluate the results:

5.1.1 Accuracy[52]

Accuracy is the degree of closeness of measurements of quantity to that quantity's actual value. In this work accuracy corresponds to exact match of the clinical scores.

5.1.2 Root Mean Square Error

While predicting clinical scores we realized that getting an exact match to what the physician thinks might not be the right assessment of the module. Hence it is more important to understand how close we are to predicting the clinical scores. When a physician reads a clinical narrative he considers the context in which terms occur, looks at the overall record, considering various parameters and assign a clinical score based on his judgment of patient's health condition. However it is difficult to build or train a module that thinks similar to a physician. Hence we consider the root mean square error to analyze how close we are to predicting clinical scores as compared to what a physician would predict.

RMS Error[49] is used to measure the difference between the values predicted by a model and the values actually observed.

RMS Error is given by

$$\sqrt{\frac{\sum_{t=1}^{n}(y_t - y'_t)^2}{n}}$$

Where

N – Number of samples in the test data set

Yt – predicted value

Y't – actual value

5.1.3 Number of mis predictions

The number of mis predictions correspond to the number of samples where the difference between the predicted and actual value was not equal to zero i.e. the residual was not zero.

5.1.3 Standard Deviation[50]

Standard deviation measures the amount of variation from the average. In our work we compute the standard deviation for the error distribution to analyze whether we are predicting higher or lower clinical scores.

5.1.4 Confusion Matrix[51]

A confusion matrix is 2x2 table which includes the number of true positives, false positives , true negatives and false negatives. The following table shows a sample confusion matrix:

True Positive	False Negative
False Positive	True Negative

Table 5.1 Confusion Matrix

In our work , we compute the confusion matrix for every clinical parameter required to compute the TIMI and San Francisco Syncope score such as whether patient had more than three CAD Risk factors or whether patient experienced abnormal EKG etc.
The true positives correspond to number of samples where both the physician and the module reported a clinical parameter to be present in patient. The true negatives correspond to number of records where both physician and the module reported that the clinical parameter is absent in the patient. The false positives correspond to number of records where the module wrongly predicted presence of a clinical parameter whereas the physician reported as the clinical parameter being absent. Lastly the false negative correspond to number of records where the module failed to capture a clinical parameter as being present in patient whereas physician reported it be present. For instance consider the clinical parameter 'shortness of breath'
If both physician and module predict it to be true then it qualifies as True Positive
If both physician and module predict to be absent in patient then it qualifies as True negative
If physician reports that patient had shortness of breath history whereas the module failed to capture it then it qualifies as False Negative.
Lastly, if module predicts that patient had shortness of breath history whereas physician denies then it qualifies as False Positive.

5.2 Data Corpus

The i2b2 data set includes 889 discharge summaries. The module was evaluated in two phases. In the first phase the discharge summaries were evaluated by the Graduate students from the Dept. of Computer Science and Engineering, University of Maryland Baltimore County. In the second phase, the clinical scores were computed by the medical professionals from the University Of Maryland School Of Medicine. The following table shows the distribution of the test data corpus.

Annotators	Cardiac Notes	Non Cardiac Notes
Graduate Students	26	74
Medical Professionals	10	15

Table 5.2 Data Corpus

The cardiac notes correspond to those discharge summaries where patient was admitted for coronary artery disease or myocardial infarction or chest pain or severe angina or severe aortic stenosis etc. The non cardiac notes correspond to those discharge summaries where the admission diagnosis could be cancer or renal or obesity problem i.e. not related to cardiovascular health issues.
The Medical professionals included two medical residents and one physician. The demographics of the subjects have been discussed in detail later.

5.3 Limitations of the data set

5.3.1 Discharge Summary versus Admission Note

A discharge summary is a document written by the patient's physician when the patient is discharged from the hospital. It contains a brief summary of all important information during the course of hospital stay including discharge diagnosis and the follow up plan for care. The scores that we are trying to predict are used for risk stratification when a patient enters Emergency Room. Hence they require information regarding patient's recent laboratory test results, patient's detailed medical history. Rather they need information related to the cause of patient's admission to ER, the initial symptoms observed, patient's health condition in the past one week. Rarely this type of information is present in a discharge summary and such information is more likely to be

found in an admission note. However for this research we were using a publicly available i2b2 data set which included only discharge summaries. It was one of the data set that closely resembled the data set required for this research. A discharge summary has a brief statement regarding the admitting diagnosis of patient whereas TIMI and Syncope scores require parameters such as whether patient had been on aspirin dosage recently or whether patient had multiple episodes of angina in past 24 hours or whether the patient's attending physician had suggested a troponin or electrocardiogram test. A discharge summary briefly discusses the medical history of patient and focuses more on the discharge diagnosis and the follow up health care plan. Hence it is less likely to find data such as whether patient had a history of congestive heart failure or dyspnea and other coronary risk factors.

According to Wikipedia, an admission note is part of medical record that documents the patient's status, reasons why the patient is being admitted to hospital and initial instructions for the patient's care. This is the data a physician looks at to determine whether the patient has chances of serious outcomes such as myocardial infarction or ischemic events. The purpose of the note corresponds to reason behind the admission such as chest pain or unstable angina which immediately allows deducing whether the patient experienced unstable angina recently or whether patient has severe stenosis. An admission note also includes history of present illness which would provide information about patient recent symptoms, medication, past medical and surgical history which indicates whether patient had coronary risk factors, family and social history which determines whether patient had a family history of CAD or whether patient is an active smoker. An admission note also has a dedicated labs and diagnostic studies section which include laboratory test results during or before admission to ER. Lastly the physical exam section includes vital signs such as blood pressure, pulse and respiratory rate. Hence an admission is more structured and has more information as compare to a discharge summary.

5.3.2 Missing Data

5.3.2.1 Electrocardiogram Test Results

As mentioned in previous section since the data set included discharge summaries the laboratory test results were missing from most of the records. One of the clinical parameters that we were predicting is abnormal EKG. Since the electrocardiogram test results were not available we had to identify a set of clinical terms which indicated abnormal EKG. The investigators in this research are not domain experts , hence currently we have only a limited set of terms such as atrial fibrillations, atrial flutter, arrhythmia, ST segment elevations, left or right bundle branch block, left or right hypertrophy etc. This is not an exhaustive list. Similarly problem occurred with analyzing whether patient had significant stenosis or whether patient had significant ST changes. Terms such as coronary artery bypass grafting or presence of stents are phrases indicating significant stenosis. Hence in scenarios where the percentage of stenosis is not explicitly mentioned, a physician searches for such clinical terms indicating presence of coronary artery disease. Significant ST changes are indicated by phrases such as 'inferior ST segment elevation' or 'ST elevation MI'. Another problem we identified that sometimes a collection of terms might indicate significant changes. For instance if the patient experienced chest pain and had abnormal EKG as indicated by term 'ST segment elevation' it indicates that patient has significant ST changes. Similarly we identified unstable angina was reported using informal language such as 'onset of chest pain' or 'substernal or atypical chest pain'. Usage of natural language instead of values for clinical parameters limits the accuracy of the prediction module. This necessitates the need to update the gazetteer list of clinical significant terms based on feedback from physicians.

5.3.2.2 Use of old Cardiac Markers

It was observed that few of the records used old cardiac markers such as CKB or ISO. To determine TIMI score we need to know whether the patient has elevated markers. The most recent cardiac markers are whether the patient has taken S Troponin test and the corresponding test results. However the terminology and the language used in the records indicate that the records were at least a decade old. One of the major challenges with this data set was that they seemed to be quite old and not written by medical professionals. At numerous places we observed usage of informal language, old medical practices such as CKG or ISO cardiac markers were reported, a lot of noisy data such as paragraphs that had no medical information, lot of grammatical and typographical mistakes.

5.3.3 Unstructured Data Set

The i2b2 data set is a publicly available data set used for Shared tasks such as identifying smoking status of patient or presence of obesity or related co morbidity in patient or extracting medication and medication related information from records etc. Hence these records are not restricted to cardiovascular domain and include a lot of noisy data. This creates problems while asserting facts such as the patient might have been on aspirin dosage, but not because he was experiencing chest pain but due some other health problem. Such kind of data skewed the accuracy of results. Similarly we were able to identify only handful of records whose admission diagnosis was chest pain or CABG or coronary artery disease or perhaps to say related to heart failure. These records are not formatted according the way standard clinical narratives are written now a day's making them less human readable. Many of medical records use medical jargon and informal language making it difficult for the system to do efficient semantic analysis. We have discussed some of the text processing issues in the Chapter 7

5.4 Evaluation Results

5.4.1 TIMI Score

5.4.1.1 Evaluations done by Computer Science Graduates

Test Data Set	Accuracy	Root Mean Square Error	Standard Deviation for Error	Mean Error
100	91 %	0.3	0.2932	-0.7

Table 5.3 TIMI Score Evaluation Results : Computer Science Graduate Students

5.4.1.2 Evaluations done Medical Residents

Test Data Set	Accuracy	Root Mean Square Error	Standard Deviation for Error	Mean Error
6	66.7	1.47	1.32	-0.8333

Table 5.4 TIMI Score Evaluation Results: Medical Residents

5.4.1.3 Evaluations done by Physician

Test Data Set	Accuracy	Root Mean Square Error	Standard Deviation for Error	Mean Error
19	42 %	1.48	1.25	-0.84211

Table 5.5 TIMI Score Evaluation Results: Physician

5.4.2 San Francisco Syncope Score Results

5.4.1.1 Evaluations done by Computer Science Graduates

Test Data	Accuracy	Root Mean	Standard Deviation for	Mean Error

Set		Square Error	Error	
100	81 %	0.5	0.47937	0.15

Table 5.6 Syncope Score Evaluation Results : Computer Science Graduate Students

5.4.1.2 Evaluations done Medical Residents

Test Data Set	Accuracy	Root Mean Square Error	Standard Deviation for Error	Mean Error
6	33 %	1.08	1.0488	0.5

Table 5.6 Syncope Score Evaluation Results : Medical Residents

5.4.1.3 Evaluations done by Physician

Test Data Set	Accuracy	Root Mean Square Error	Standard Deviation for Error	Mean Error
19	36.8 %	0.9733	0.653376	0.736882

Table 5.6 Syncope Score Evaluation Results : Physician

5.4.3 Confusion Matrices for Clinical Score Parameters

The confusion matrix for the clinical parameters in case of evaluations done by the physician is as follows. The test data set that was evaluated by a physician included 19 discharge summaries.

5.4.3.1 Age > 65

11	0
0	8

Table 5.9 Confusion Matrix: Age

5.4.3.2 Whether patient had > 3 coronary risk factors?

3	1
1	12

Table 5.10 Confusion Matrix: Whether patient has greater than 3 coronary risk factors?

5.4.3.3 Whether patient had multiple episodes of Angina in past 24 hrs?

0	7
2	10

Table 5.11 Confusion Matrix : Whether patient had multiple episodes of angina in past 24 hours

5.4.3.4 Whether patient had known coronary artery disease i.e. stenosis greater than 50 %

3	7
0	9

Table 5.12 Confusion Matrix : Whether patient had known coronary artery disease

5.4.3.5 Whether patient use aspirin in the past 7 days

2	2
3	12

Table 5.13 Confusion Matrix : Whether patient use aspirin in the past 7 days ?

5.4.3.6 Whether patient had ST changes greater than 0.5 mm

1	6
2	10

Table 5.14 Confusion Matrix : Whether patient had ST changes greater than 0.5 mm

5.4.3.7 Whether patient had elevated cardiac markers?

1	1
1	16

Table 5.15 Confusion Matrix : Whether patient had elevated cardiac markers?

5.4.3.8 Whether patient had known hematocrit less than 30 % ?

2	2
2	13

Table 5.16 Confusion Matrix : Whether patient had hematocrit less than 30 %

5.4.3.9 Whether patient had congestive heart failure history?

7	0
1	11

Table 5.17 Confusion Matrix : Whether patient had congestive heart failure history

5.4.3.10 Whether patient experience shortness of breath?

1	0
2	16

Table 5.18 Confusion Matrix : Whether patient had shortness of breath history?

5.4.3.11 Whether patient had abnormal EKG

7	1
6	5

Table 5.19 Confusion Matrix : Whether patient has abnormal EKG

5.4.3.12 Whether patient had Systolic Blood Pressure less than 90 mm of Hg

0	1
0	18

Table 5.20 Confusion Matrix : Whether patient's Systolic BP is less than 90 mm Hg

5.4.4 Binary Classification Results

The clinical scores are used to risk stratify the patients presenting to the ED. Hence many times more than the exact score we need to determine whether the patient has high or low risk for ischemic event.

According to the physician on our research panel,

Syncope score 1 or above - patient is not in low risk group for serious outcome

Syncope score of 0 - patient is in low risk group for serious outcome

TIMI score of 4 and above - patient has high risk of ischemic event

TIMI score below 4 - patient has low risk of ischemic event

Hence we computed the confusion matrix for such a binary classification.
True Positive - If both the system and medical professional thought that patient has high risk
False Positive - If system predicted patient has high risk but medical professional thought that the patient isn't at high risk
True Negative - If both the system and medical professional indicate that patient has low risk.
False Negative - If physician thinks patient has high risk , however system predicts patient has low risk
The confusion matrices for the evaluations done by Medical Professionals are as follows:

TP = 3	FN = 4
FP = 0	TN = 18

Table 5.22 Confusion Matrix : TIMI Score of 4 and above

TP = 15	FN = 2
FP =8	TN = 0

Table 5.23 Confusion Matrix: San Francisco Syncope Score of 1 and above

5.5 Discussion

5.5.1 Parameters those were difficult to annotate

While conducting evaluations with Graduate students from the Dept. of Computer Science and Engineering, we had asked them to fill the following table. We assume that computer science graduates don't have the necessary expertise to compute the clinical scores. In order to simplify the process we asked them to find the clinical terms in text and where required fill the required values and corresponding clinical section. We also provided them a list of synonyms and abbreviations. Following is the table that was asked to fill.

Record No.	Hypertension (Y/N)	Congestive Heart Failure (Y/N)	Hypercholesterolemia (Y/N)	Diabetes (Y/N)	Current Smoker (Y/N)	Aspirin (section)	Unstable Angina (Y/N)	Hematocrit (Value)	ST segment change (Value)	Blood Pressure (Value)	Stenosis (Value)	Abnormal ECG / EKG (Y/N)	Shortness of breath (Y/N)	Family History of CAD (Y/N)	Age (value)	Troponin test (Y/N)	Coronary Artery Disease (Y/N)	TIMI SCORE (DON'T FILL)	SYNCOPE SCORE (DON'T FILL)

Table 5.24 Metric to be filled by Computer Science Graduates

Following instruction sheet was provided and some minimal explanation of terms, purpose of study was given to the annotators.

- Y/ N corresponded to whether the clinical condition was present or absent in the patient
- Value required the evaluators to fill the value for the corresponding clinical term. If multiple values are mentioned, they were asked to report all the values.
- Current Smoker – Whether the patient is an active smoker. For instance if smoking is mentioned in history of past illness or if the patient has quit smoking then he is not an active smoker.
- The field 'section' corresponds to the clinical section in which the term occurred such as Admission on Medications or Discharge Medications.
- Angina – whether the patient experienced unstable angina recently or had multiple episodes of angina in past few days
- Blood Pressure is usually stated as x/y where x correspond to the systolic blood pressure and y corresponds to the diastolic blood pressure.
- ST segment changes are usually expressed in mm and stenosis is usually reported in percentage
- The abnormal EKG terms include sinus bradycardia, atrial flutter, left atrial hypertrophy, right atrial hypertrophy, left bundle branch block, right bundle branch block, arrhythmia, abnormal heart rhythm, supraventricular tachycardia, ventricular fibrallation,atrial fibrillation, ventricular tachycardia, irregular rhythm, premature atrial contraction, myocardial infarction, ischemia, premature ventricular contraction, ST elevation, ST depression
- Family History of CAD means whether any of the patient's immediate family member had coronary artery disease
- Troponin is true if the patient has taken troponin test
- The list of synonyms and abbreviations are as follows:
  Congestive heart failure = CHF = chronic heart failure
  Hypercholesterolemia = hyperlipidemia
  Aspirin = ASA
  Diabetes = diabetes mellitus
  Hematocrit = HCT
  ECG = electrocardiogram=EKG
  Shortness of breath = dyspnea = breathlessness = difficulty breathing
  CAD = coronary artery disease

We found that the following parameters were difficult to annotate:

- Abnormal EKG: It is difficult to do an exhaustive search of all abnormal EKG terms, their synonyms and abbreviations
- Blood Pressure and Hematocrit Values: It was observed that annotators reported only one value for these clinical terms though multiple values were mentioned in the record.
- Family History of CAD: In case of this parameter if the family history was mentioned as a separate section then annotators could easily identify whether this coronary risk factor was present. However if it was mentioned in document, that some family member of patient had history of coronary artery disease then it was not reported since annotators mostly didn't end up reading the entire document.
- In many cases we found the negation was not identified by the annotators. The presence of a term in medical record was considered as presence of medical fact.

- Unstable Angina: We found that annotators got confused between exertional, angina at rest and unstable and multiple episodes of angina.

5.5.2 Demographics of Subjects

5.5.2.1 Demographics of Non Medical Professional Annotators

S. No	Gender	Profession	Years of Experience	Specialization	Age
1	Male	Attending Physician		Internal Medicine	52
2	Female	Medical Resident	3		33
3	Male	Medical Resident	3		26

Table 5.25 Demographics of Medical Professionals

5.5.2.2 Demographics of Medical Professional Annotators

S No.	Gender	Masters / PhD	Year of Program	Specialization	Age
1	Male	PhD	1	CS	28
2	Male	PhD	5	Semantic Web	27
3	Female	PhD	1	CS	25
4	Male	Masters	2	CS	23
5	Male	Masters	2	CS	24
6	Female	Masters	2	CS	24
7	Female	PhD	3	CS	
8	Male	Masters	2	CS	27
9	Male	PhD	2	CS	25

Table 5.26 Demographics of Non Medical Professional Evaluators

5.5.3 Parameters those were difficult to predict

5.5.3.1 Age

If age of patient is mentioned in the section Age then it was easier to capture. When the age of patient was mentioned in the history of present illness, we had to identify set of regular expressions that could capture the age of patient. We observed that there was very little consistency in the way age was reported. It was difficult to identify a generic regular expression that could capture all the ways in which age of patient was mentioned. Further there also instances when the age of patient was mentioned in sections such as brief history of hospital stay or brief resume of treatment. Some of these sections are not considered valid clinical sections by SecTag. Hence when age of patient was reported in such sections the system failed to identify it.

We identified 23 different ways in the age of patient could be stated and identified 3 different regular expressions that could capture them:

Following are few of the ways in which age of patient was reported:

75 YOF	65 YO
67-year-old	108-year-old
Seventy eight year old	seventy-two-year-old
57 years old	sixty nine-year-old
111y year old	twenty-six year old
47 yo	32 y / o

Age: 81
58 y.o.
64 y / o w
79yo
35 yoF
40 yo F
34y
77F
80 yM
68y / o
80 female

The regular expressions to capture above were:

- ([0-1]?[0-9][0-9])[.-]?(Y|y)[.-/]?(O|o)
- (.*)[-]?year[-]?old
- [0-1]?[0-9][0-9][y]?[-]?year[-]?old

5.5.3.2 Aspirin Dosage and Multiple episodes of angina

We discuss in Chapter 7 how queries such as whether aspirin was administered in past 7 days or whether patient had multiple episodes of angina in past 24 hrs require deeper semantic analysis. At the same time they also require understanding of context in which clinical term such as chest pain occurs. Sometimes even if the patient is experiencing exertional angina but if it is occurring frequently then the physician would assess it as multiple episodes of angina. We had to make some simplifying assumptions while asserting these facts such as if the patient's admission diagnosis suggest aspirin or whether patient has been on aspirin dosage as indicated by clinical sections such Admission Medication or History of Present Illness then the patient has been administered aspirin recently. However certain clinical sections such as history of past illness or discharge medications clearly indicate that patient has not been on aspirin dosage recently. Similarly in case of multiple episodes on angina we assumed that unstable angina could indicate that patient had multiple episodes recently. A similar approach was followed for the clinical parameter abnormal EKG. Ideally if the discharge summaries included laboratory test results then estimating abnormal EKG would have been easier.

5.5.3.3 ST segment changes and Stenosis greater than 50 %

It was observed that the exact value of ST segment change was less frequently mentioned. Usually there were statements such as ST segment elevations or ST segment depression which makes it difficult for a computer system to predict the value of ST segment change. A physician who assesses the entire medical record might predict only ST segment elevations without any mention of value as significant ST segment change. Similarly problem was identified for stenosis. For such parameters we need a domain expert or an intelligent inference system that could deduce whether the ST segment changes or stenosis was significant based on other terms found in the context.

5.5.4 Computation Time Analysis

The size of the largest discharge summary was 17 KB and the smallest discharge summary was 1 KB

Parameter	Time Require in seconds
cTAKES Aggregate UMLS Document Processor and UIMA CPE to process 889 medical records	19214.448
Extracting information and building knowledge base to predict TIMI and Syncope scores from the largest record	6.78

Compute TIMI and Syncope score from the knowledge base	1.58
Extracting Information and building knowledge base of all cardiovascular facts	7.13
Infer additional cardiovascular facts using HF Ontology, Jena OWL Reasoner and one simple clinical rule	7.998

Table 5.24 Computation Time Analysis

The Performance Report for cTAKES Aggregate UMLS Document Processor is as follows:

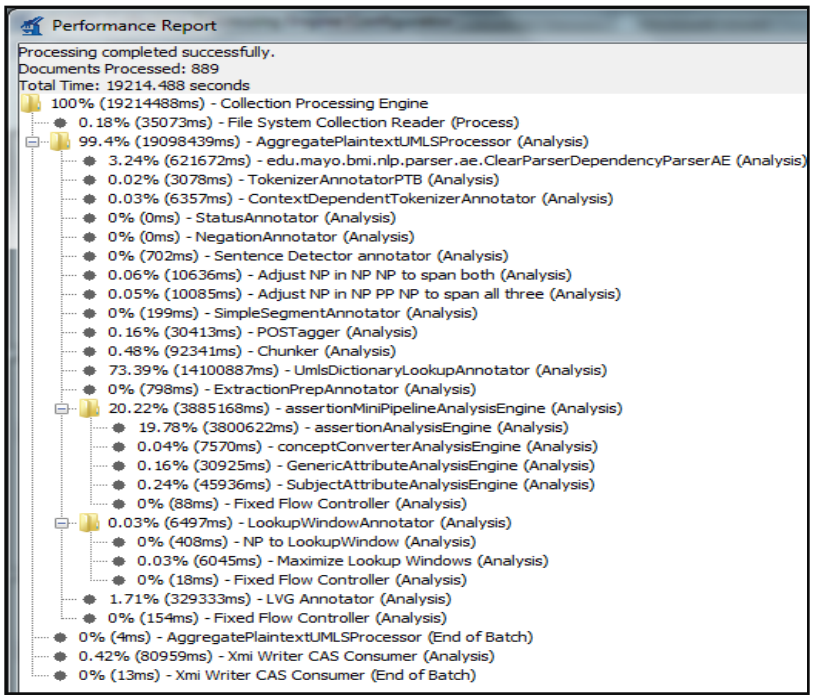

Fig 5.1 Performance Report generated by UIMA CPE after processing 889 discharge summaries

This shows that the UMLS Dictionary Lookup i.e. identifying a clinical term from a document consumes maximum time whereas identifying polarity of a term or detecting end of line markers require negligible amount of time. Depending upon the size of data set the XMI Writer CAS Consumer and the File System Collection Reader would require moderate to significant amount of time to create the annotations files.

Chapter VI

LIMITATIONS AND CHALLENGES

6.1 Challenges

6.1.1 Missing values

The computations of clinical scores require estimation of values associated with the clinical term. For instance while deciding whether a patient has coronary artery disease the percentage of stenosis is checked. However rarely clinicians report the value and rather use phrases that indicate clinically significant stenosis. Similarly we found the electrocardiogram test results were missing from most of the records. Hence we had to identify terms that indicate abnormal EKG. The missing values from records, limited the estimation of clinical parameters based on these values. The issue of missing values was handled to a certain extent by using additional context surrounding the term.

6.1.2 Multiple values associated with same concept

A clinical term is mentioned multiple times throughout the document and it is difficult to associate a unique value with the term. Mapping multiple values to a single clinical term would complicate the decision support tasks. Hence depending upon the target clinical score, section tagging, dependency parsing, lookup window of term was used to associate a unique value with the term. For instance hematology test results would be mentioned in history of present as well as past illness. Further this data would also be available in laboratory results section. Hence while computing CHESS score we focus on hematocrit values from laboratory results section and ignore hematocrit values from other section.

6.1.3 Heterogeneous nature of records

There is no standard terminology or style of writing or structure that is followed throughout the i2b2 records which makes it difficult to develop a generic system that would understand the underlying text. For instance the clinical parameter of family history of coronary artery disease might be reported in family history section or in history of past illness section as a single statement that the ‘patient's brother had coronary artery disease’ or may be in section cardiac risk factors. Similarly there is no consistency in the clinical section headings used across the text. Some admission records have a medication section that includes discharge, current, admission medications while in some of the records each of this is a separate section.

6.1.4 Different concept unique identifier associated with canonical forms of clinical terms

cTAKES does limited identification of canonical forms of clinical terms. One of ways we thought of addressing this issue is to compare the UMLS identifiers of canonical forms of a clinical term. For instance we assumed that atrial fibrillation and atrial fibrillations would be having same CUI and hence it could be used for extracting canonical forms of a word. cTAKES UMLS annotator recognizes atrial fibrillations as medical concept. However the ontology does not recognize the canonical form atrial fibrillations as instance. Only atrial fibrillation is considered a valid semantic type. Hence there was a glitch in mapping extracted clinical term to an ontology term since the canonical forms have distinct concept identifier. Hence the text to ontology mapping has to be done using string comparison and not concept unique identifiers.

6.2 Limitations

6.2.1 Queries that require deeper semantic analysis

For instance 'whether the patient has taken aspirin in past seven days'

With section tagging and limited text processing one can ascertain whether the patient has been administered aspirin recently. For instance if aspirin occurs in the section - discharge medication

we conclude that the suggested medication is aspirin and if it occurs in the section - medication on admission we conclude that the patient has taken medication in recent time. However pinpointing the fact whether it has been taken in past 7 days is a deeper NLP problem. Even if we consider the lookup window of terms or sentence in which aspirin is mentioned it is hard to find evidence that could conclude whether the aspirin dosage was taken in past one week. Further the current data set is a set of discharge summaries and not admission notes. Hence they don't have a detailed account of the hospital course which makes it even more difficult to do temporal analysis of the occurrence of an event.

6.2.2 Medical Jargon, Abbreviations and Usage of Informal Language

Consider the following sentence:

48 yo f, fx CAD, admitted CP, r/o'ed MI[27]

Only a clinician can understand that the above sentence meant that a 48 year old female with coronary artery disease was admitted to ER following chest pain; however myocardial infarction has been ruled out.

Analyzing such statements require deeper investigation of medical jargon and informal terms. Further this highlights another issue i.e. of use of abbreviations. Commonly used abbreviations such as MI - myocardial infarction, CAD - coronary artery disease are usually listed as synonyms in the ontology. However CP for chest pain is a less frequently used abbreviation and r/o'ed for ruled out is an example of informal language. In our current work we don't intend to identify all alternate usage of terms for a medical concept. Another point to be noted in the given example is the manner in which age is stated. We have identified 28 different ways in which age of patient is reported. However as we analyze more records we realized that it is very difficult to build a regex that covers all i2b2 records.

Further in the current work we do a limited resolution of abbreviations essentially those which are explicitly stated as synonyms by the ontology. Ambiguous and relatively unknown abbreviations considerably limit the performance of information extraction phase.

6.2.3 Inconsistency in cTAKES annotations

The term ‘blood pressure’ was not annotated as a medical term in few records. Hence the systolic blood pressure value was missing from the RDF graph constructed for those records. Similarly the term coronary artery disease was not annotated as a medical term in few records. Hence we identified following three classes of terms that are currently not being extracted:

Not annotated by cTAKES but a valid semantic type in ontology – For example: post infarction which is a class in the ontology but cTAKES UMLS Annotator does not recognizes it as a clinical term.

Annotated by cTAKES but not a valid cardiac term according to ontology – For instance: stenosis. Stenosis is a known medical term, however the ontology does not define any super class. The ontology does have instances such as Bilateral Renal Artery Stenosis or Valve Stenosis.

Neither annotated by cTAKES nor a valid semantic type according to ontology but a well known cardiac term and a valid UMLS concept – For instance: bypass surgery. cTAKES UMLS Dictionary Lookup Annotator does not recognize ‘bypass surgery’ as a clinical term. Further since we are using a heart failure and not a cardiovascular domain specific ontology this term is missing from there as well.

Hence we intend to use UMLS RRF browser to create a subset of cardiovascular terms and accordingly extend the ontology. However for the terms that are not recognized by cTAKES we are not able to do text analysis using cTAKES Annotation results such as identifying the polarity of the entity.

6.2.4 Extraction of values for limited terms:

We intend to build a generic decision support system for cardiovascular domain. This necessitates asserting as many cardiac facts about patients as possible. The numeric values of the clinical terms play a significant role in any decision support task. We have identified the following cardiovascular terms that can be quantified: blood pressure, stenosis, ejection fraction, and pulse per minute, ST-T wave abnormality, hematocrit, blood cells, heart and respiratory rate, hemoglobin, white blood cell count and aspirin dosage. This is not an exhaustive list of cardiac terms that have an associated value.

6.2.5 Usage of generic term for a specific purpose:

For instance consider the following statement:

His angiograms from Morehegron Valley Hospital from his four prior catheterizations were reviewed and the patient was gradually tapered off all of his vasoactive medications .

The clinician uses a generic term catheterization which is not a valid semantic type in the heart failure ontology. The ontology contains a class right heart catheterization. Further neither right heart catheterization is a canonical form or a subtype of catheterization. Rather they have different UMLS identifiers. Since discharge summary would be reviewed by a ER physcician, the clinician did not feel the need to explicitly mention the term cardiac catheterization. However since the prefix cardiac or heart was absent the ontology did not assert the fact that patient had been diagnosed with right heart catheterization.

Another instance is usage of term ‘bypass surgery’, The term bypass surgery was not getting asserted since the ontology defined an instance coronary artery bypass surgery which had no super class bypass surgery. Since we using ontology to assert facts, terms having partial instance or class names were not getting asserted. This further strengthens the need to extend the ontology. According to UMLS coronary bypass surgery, bypass surgery and coronary artery bypass are all valid UMLS terms with different concept unique identifier.

Chapter VII

FUTURE WORK

This work can be extended in multiple directions. Following are few of them:

7.1Adding ontologies from other domains:

The knowledge base can be made richer by using ontologies from multiple domains. For instance using NCI thesaurus or Renal Gene ontology would result in adding Cancer or Kidney related facts to knowledge base. With such diverse data about patient we could use more generic clinical rules to infer valuable information about patients.

7.2 Scalability:

Currently processing a single medical record to infer additional facts using patient facts as RDF triples, heart failure ontology, single clinical rule and JENA[46] OWL reasoner requires 7.998 seconds. However in future we can develop a faster automated system using a better reasoner such as JESS to process more number of records in less time.

7.3 Extending the gazetteer list of cardio vascular terms:

Using UMLS Metamorphsys RRF browser we could identify additional cardiovascular terms which are not present in heart failure ontology. These terms could be mapped to the semantic types defined in UMLS Semantic Network. There is slight overlap between the HF super classes and UMLS SN which allows using properties from HF ontology to assert facts involving these terms.

7.4 Untapped Clinical Sections: There is lot of hidden information in laboratory data section which couldn't be captured using simple text processing techniques. Further due to limitations of existing ontology, facts related to cardiovascular medical devices or anatomies of heart are not asserted.

7.5 UMLS Definitions as SWRL rules: [2]

UMLS defines every clinical term in the following format: X is characterized by Y1, Y2 and results in Z1, Z2 etc. These definitions could be transformed into clinical rules to enrich the knowledge base.

7.6 Canonical forms:

Currently cTAKES handles a limited number of canonical forms of a clinical concept. For instance the terms such as anemicon, ischemic, thrombal, atrial fibrillations are not detected as medical terms. One way to address this issue is to build a subset of canonical forms of every clinical term using UMLS RRF browser and expand the gazetteer list.

7.7 Word sense disambiguation[28]

An example of this is whether the clinical term 'cold' is mentioned as a disease or as a symptom. This requires identifying additional context surrounding the term as well as using the correct semantic type restrictor while building the gazetteer list.

7.8 Resolving misspelled terms [28]

It was observed that in few records the clinical terms were not being detected since they were misspelled. This requires using Levenshtein scores or cleaners such as Metaphone or Tolentina to resolve such ambiguities. Clinical reports also contain typographical error which would result in losing significant information about patient. For instance 'hyptension' It is difficult to resolve this ambiguity since the clinician might be referring hypertension or hypotension. Hence effectively we lose one observation about the patient.

VIII
CONCLUSION

Thus we propose a novel and hybrid system that provides diagnostic warnings using text and semantic web analytics. We introduce how combination of annotators can be used to extract clinical relevant information and how ontologies could be exploited to establish relation between various clinical parameters. We present a design for decision support system using patient facts, medical ontologies and clinical rules to identify patients presenting at ER with adverse outcomes. Our experiment demonstrates that such a system worked for computing clinical scores from cardiovascular domain.

We defined the preprocessing steps needed to extract clinical entities from unstructured discharge summaries. We identified how the structure of ontology and medical thesaurus can be exploited to assert medical facts about patient. We also evaluated the system for computing TIMI, San Francisco Syncope scores. We put forth certain challenges and limitations in the current work and suggest how this work could be extended in future. In the current work we emphasized only on heart failure ontology however similar system could be used in some other domain by changing ontology. Due to time constraints we focused on using existing text processing tools and ontologies and enhancing them as needed.

One of the strengths of our work is that we evaluated the system on discharge summaries from other domains as well. A second strength is being able to capture maximum cardiovascular related facts about patients. We also demonstrate that simple text mining without machine learning algorithms can be used to uncover medical facts about patients. A major contribution of this work is identifying the power of combination of text and semantic analytic to deduce implicit facts about patient. In chapter 3 we saw how such a system could be realized. In chapter 4 we discussed the intricacies of every intermediate step in the proposed system.

The final results showed that the root mean square error in predicting TIMI and San Francisco Syncope scores were 1.48 and 1 respectively. The difference in results indicate that we need to improve on capturing contextual information associated with a clinical term. We need to have an inference system that estimates those parameters which require a value. We identified certain shortcomings of existing work and aim to achieve them in near future.

Our final solution is a novel clinical decision support system using semantic web and clinical text processing techniques that computes cardiac risk scores from clinical narratives.

BIBLIOGRAPHY

1. US Department of Health and Human Services, National Institutes of Health, National Library of Medicine. Unified Medical Language System (UMLS). Available from: www.nlm.nih.gov/research/umls, 2012AB.
2. Jovic A, Gamberger D, Krstacic G. Heart Failure Ontology. Bio-Algorithms & Med-Systems. 2011;7(2):101-110
3. Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY
4. http://www.mdcalc.com/timi-risk-score-for-uanstemi/
5. http://www.mdcalc.com/san-francisco-syncope-rule-to-predict-serious-outcomes/
6. Kashyap V, Borgida A. Representing the UMLS semantic network using OWL: (Or "What's in a Semantic Web link?") In: Fensel D, Sycara K, Mylopoulos J, editors. The SemanticWeb–ISWC 2003.LNCS 2870. Heidelberg: Springer-Verlag; 2003. pp. 1–16.
7. Garvin JH, DuVall SL, South BR et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *J am Med Inform* Assoc 2012. Doi:10.1136/amiajnl-2011000535
8. Bruce, Lars-Erik. Processing Electronic Medical Records. Masteroppgave, University of Oslo, 2012
9. Esposito M.: Congenital Heart Disease: An Ontology--Based Approach for the Examination of the Cardiovascular System. Lecture Notes in Computer Science 2008, 5177: 509-516.
10. Chiarugi F., Colantonio S., Emmanouilidou D., Martinelli M., Moroni D., Salvetti O.: Decision support in heart failure through processing of electro- and echocardiograms. Artif. Intell. Med. 50(2), 95-104, 2010.
11. Cenan C., Sebestyen G., Saplacan G., Radulescu D., "Ontology-based distributed Health Record Management System", on ICCP International Conference, pp.307-310, 2008.
12. D. C. Wimalasuriya and D. Dou. Ontology-based information extraction: An introduction and a survey of current approaches. Journal of Information Science, 36(3):306–323, 2010.
13. Halpern Y, Horng S, Nathanson LA, Shapiro NI, Sontag DA. A Comparison of Dimensionality Reduction Techniques for Unstructured Clinical Text. ICML 2012 Workshop on Clinical Data Analysis, July 2012.
14. Halpern Y, Horng S, Nathanson LA, Shapiro NI, Sontag DA. Patient Surveillance Algorithms for the Emergency Department. NIPS Workshop. Sierra Nevada, Spain, Dec 16, 2011.
15. Nate Blaylock, James Allen, William de Beaumont, Hyuckchul Jung. Towards and OWL-based framework for extracting information from clinical texts. In *Proceedings of the First International Workshop on Semantic Applied Technologies on Biomedical Informatics (SATBI)*, Chicago, Illinois, August 1-3, 2011.
16. Jiang M, Chen Y, Liu M, Rosenbloom ST, Mani S, Denny JC. et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. J Am Med Inform Assoc. 2011;18(5):601–606.
17. Deléger L, Grouin C, Zweigenbaum P. Extracting medical information from narrative patient records: the case of medication-related information. J Am Med Inform Assoc. 2010 Sep-Oct;17(5):555–8.

18. Uzuner Ö, Solti I, Cadag Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17:514–18
19. Patrick J, Li M (2010) High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 17: 524–527.
20. Y. Li, R. Krishnamurthy, S. Raghavan, S. Vaithyanathan,and H. V. Jagadish. 2008. Regular expression learning for information extraction.
21. Meystre SM, Thibault J, Shen S, *et al*. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. J Am Med Inform Assoc 2010;17:559–62
22. Gedzelman, S.; Simonet, M.; Bernhard, D.; Diallo, G.; Palmer, P., "Building an ontology of cardio-vascular diseases for concept-based information retrieval," *Computers in Cardiology, 2005* , vol., no., pp.255,258, 25-28 Sept. 2005 doi: 10.1109/CIC.2005.1588085
23. S.B. Johnson, A semantic lexicon for medical language processing, J Am Med Inform Assoc. 6 (3) (1999) 205–218
24. Wang X, Chused A, Elhadad N, Friedman C, Markatou M: Automated knowledge acquisition from clinical reports**.** *AMIA Annual Symposium* 2008.
25. Meystre SM, Haug PJ: Comparing natural language processing tools to extract medical problems from narrative text. *AMIA Annu Symp Proc* 2005, 525-529.
26. Chen E, Hiripcsak G, Friedman C. Disseminating natural language processed clinical narratives.Proc AMIA Annu Fall Symp. 2006:126–30.
27. 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.
28. Automated Classification of the Narrative of Medical Reports Using Natural Language Processing. Lin Sun, Ira Goldstein
29. H. Xu, S. P. Stenner, S. Doan, K. B. Johnson, L. R. Waitman, J. C. Denny, MedEx: A medication information extraction system for clinical narratives. J. Am. Med. Inform. Assoc. 17, 19–24 (2010).
30. Prcela, Marin, Dragan Gamberger, and Nikola Bogunović. "Developing Factual Knowledge from Medical Data by Composing Ontology Structures." *MIPRO 2007, 30. Jubilarni međunarodni skup* (2007).
31. M. Jiang, Y. Chen, M. Liu, S. Rosenbloom, S. Mani, JC. Denny, H. Xu. A study of Machine Learning based Approaches to Extract Clinical Entities and their Assertions from Discharge Summaries, Journal of the American Medical Informatics Association (JAMIA). April 20, 2011.
32. Jovic A, Prcela M, Gamberger D. Ontologies in medical knowledge representation. In: ITI2007, Cavtat, Croatia
33. Gedzelman S, Simonet M, Bernhard D, et al. Building an ontology of Cardio-Vascular diseases for Concept-Based Information Retrieval, Computers in Cardiology, Lyon, France, Dept. 2005.
34. OWL Web Ontology Language http://www.w3.org/TR/owl-semantics/
35. Resource Description Framework http://www.w3.org/RDF/
36. Apache Clinical Text Analysis and Knowledge Extraction System https://wiki.nci.nih.gov/display/VKC/cTAKES+2.5
37. XML Grid http://xmlgrid.net/

38. RDF Validator http://www.w3.org/RDF/Validator/
39. SecTag http://knowledgemap.mc.vanderbilt.edu/research/content/sectag-tagging-clinical-note-section-headers
40. Apache Unstructured Information Management Architecture http://uima.apache.org/
41. Protégé http://protege.stanford.edu/
42. Eclipse http://www.eclipse.org/
43. http://gate.ac.uk/sale/talks/gate-course-aug10/track-3/module-9-ontologies/module-9-ontologies.pdf
44. General Architecture for Text Engineering (GATE). http://www.gate.ac.uk/.
45. Ontology Development Information Extraction http://www.bioontology.org/ODIE
46. Apache Jena http://jena.apache.org/
47. Ontology Development and Information Extraction http://www.bioontology.org/ODIE
48. Prcela M, Gamberger D, Jovic A. Semantic Web ontology utilization for heart failure expert system design. Stud Health Technol Inform. 2008;136:851–6
49. Root Mean Square Error http://en.wikipedia.org/wiki/Root-mean-square_deviation
50. Standard Deviation http://en.wikipedia.org/wiki/Standard_deviation
51. Confusion Matrix http://en.wikipedia.org/wiki/Confusion_matrix
52. Accuracy http://en.wikipedia.org/wiki/Accuracy_and_precision