Thesis for the Degree of Doctor of Philosophy

SEMANTIC SEQUENCE CONTRACTION AND EXPANSION FOR DATA INTEROPERABILITY

Fahad Ahmed Satti

Department of Computer Science and Engineering
Graduate School
Kyung Hee University
Seoul, Korea

February 2023

SEMANTIC SEQUENCE CONTRACTION AND EXPANSION FOR DATA INTEROPERABILITY

Fahad Ahmed Satti

Department of Computer Science and Engineering
Graduate School
Kyung Hee University
Seoul, Korea

February 2023

SEMANTIC SEQUENCE CONTRACTION AND EXPANSION FOR DATA INTEROPERABILITY

by

Fahad Ahmed Satti

Supervised by

Prof. TaeChoong Chung

Prof. Sungyoung Lee

Submitted to the Department of Computer Science and Engineering and the Faculty of Graduate School of Kyung Hee University in partial fulfilment of the requirements of the degree of Doctor of Philosophy

Dissertation Committee:

Prof. Seong-Bae Park* _____

Prof. Eui-Nam Huh _____

Prof. Tae-Seong Kim _____

Prof. Sung-Ki Kim _____

Prof. TaeChoong Chung _____

Prof. Sungyoung Lee _____

This thesis is dedicated to my beloved parents, family, and respected teachers, who play a pivotal role by encouraging and supporting in all hard times to keep me in the position to complete my PhD degree.

Abstract

Digitization of Healthcare services have revolutionized the availability and applicability of life enhancing medical and wellness interventions, which have greatly improved the quality of life, and increased the average life-span in most parts of the world. In particular the last fifty years have witnessed the advent of data, information, and knowledge curation solutions, improving the accessibility to Patient data and Medical Knowledge, for the various stakeholders, such as Governments, Healthcare providers, Patients, and others. One of the caveats of this rapid advancement is the creation of a technological and a technical gap between the digital services, tools, and technologies utilized by healthcare providers in the developed versus the developing world.

While, a plethora of software, medical devices, and standards have been developed to cater for various aspects of the healthcare delivery systems, their availability and usage in the developing world is greatly restricted owing to the lack of resources, operational complexity, and many others. This introduces the Technological gap. On the other hand, a Technical gap has arisen due to the rapid growth in quantity and quality of medical systems and data, which has compounded the problem of heterogeneity among the platforms and conformance guidelines (commonly known as Standards); thereby decreasing the effectiveness and increasing the cost of diagnostics, treatment and follow-up.

Healthcare service delivery in the developing world is geared towards increasing the patient turn-over rate by compromising on effective digitization tools and technologies. Physician's and other healthcare providers often feel over-burdened by the additional workload required to maintain the patient history. As a result, many clinical encounters, especially in the out-patient department are not captured at all and for in-patient encounters errors and omissions are often made, which can negatively effect the treatment plans.

Thus, in order to capture the clinical encounters automatically and to create a bridge between heterogeneous medical systems, while taking into account the limited resource environment, an automated AI based methodology supported by a state-of-the-art Big Data platform is required. In this dissertation, a solution to that effect is proposed, which has three parts. Firstly, a novel Sequence Contraction method is proposed, which identifies and extracts relevant medically aligned data from unstructured text, representing clinical conversations. Secondly a novel Sequence Expansion method is proposed, which is used to convert attribute names into sequences, subsequently used for semantically aligning formally defined and/or adhoc data schema. Thirdly, the design of a Semantic Reconciliation-on-Read based Big Data engine is proposed which, archives the original medical data in semi-structured form and reconciles the same into a form compliant with a consumer.

For Sequence Contraction, in order to identify and extract the medically aligned attributes and their values from unstructured medical conversations, a three step process is used, where raw text is first converted into sequences, followed by a classification step, where unseen sequences are filtered based on their semantic similarity with existing Medically Aligned Sequence Set (MASS). This semantic similarity is based on cosine similarity between embedding vectors obtained by encoding, test sequences and MASS sequences, using a fine-tuned DistilBERT-base-uncased model (originally prepared for semantic textual similarity in the general domain). Finally, in the third step, reduction methodologies are applied on the classified probable medically aligned sequences. Using regular expressions or conceptual semantics, the classified sequences are reduced into attribute-value form, which represents a part of the semi-structured output of the sequence contraction process.

The Sequence Expansion methodology is used to create the text based artifacts, which can be used for achieving Schema Alignment between various formal and informal schema. In this dissertation a novel methodology is presented to bridge the gap between various healthcare data management solutions by leveraging the strength of transformer-based machine learning models, to create mappings between the data elements. Sequences here, are obtained by first splitting the attribute names (which can be a combination of one or more words), into suffixes using Suffix Array generation with forward pass, backward pass, and regular expression based method. The

suffixes are then filtered, if they exist in a conceptual dictionary, which holds semantic information for various elements of the healthcare domain. Next the suffixes are enriched with their associated concepts, obtained from the same dictionary, producing a machine-understandable sequence. These sequences are then matching in an unsupervised manner to align the disjoint pair of attributes, across various schema. The output of this process is a Schema Map, which represents one-to-one alignment between heterogeneous schema.

Finally, leveraging the data extraction and schema alignment methodologies, an efficient and effective platform based on a multi-dimensional data storage engine is designed, which provides data archiving and semantic reconciliation services from a single point of service delivery, thus overcoming the limited resource problem. This platform is designed to hold a large amount of medical data and schema maps, obtained from various hospitals and clinics. This collection happens in a bursty manner, supporting the eventual consistency of medical records. For semantic reconciliation, once a physicians requests for a patient's data, the existing records from the data engine is queried using SPARQL queries. The result set thus obtained is then converted into a target schema format, by collecting all Schema Maps, between the various source data schema and the target schema. Finally, the data is converted into the target schema, and made available to the consumer.

The Sequence Contraction methodology is able to achieve an accuracy of 52.96%, in terms of correctly identify a relevant medical attribute and its value. It provides an improvement of ~ 8 percentage points, when the sentence encoder is replaced with the state-of-the-art pre-trained model. Sequence Expansion results indicate, that for biased, dependent multi-class text classification, transformer-based models provide better results than linguistic and other classical models. In particular, the sentence similarity based pre-trained model, all-mpnet-base-v2, provides the best schema matching performance by achieving a Cohen's kappa score of 0.36 and Matthews Correlation Coefficient (MCC) score of 0.43, with human-annotated data. Utilizing the previous results, the Semantic Reconciliation-on-Read platform is able to archive a large amount of data, providing scalability, while maintaining timeliness and accuracy of data retrieval. The archive contains generated data with over 115.7 million serialized medical fragments for 390,101 patients. Retrieval and transformation of data into a target data schema completes the data interoperability

loop. These results provide a proof of concept for the correctness and effectiveness of the proposed methodology.

Acknowledgement

Alhumdolillah, By grace of ALLAH Almighty, who is the most beneficent and merciful. Who gave me the strength, courage, patience during my doctoral study and showering HIS blessings upon me and my family.

I am highly grateful to my advisors Prof. Sungyoung Lee and Prof. Tae Choong Chung for their boundless moral and technical supervision, guidance, and courage in coping with the difficult challenges throughout the education period of my doctoral studies. They trained me in multi-directions to face the challenges of practical life in a professional manner. Their lively natures, clear assistance, and direction enabled me able to complete my thesis. They have refined the key ingredients for high quality research, namely my skills of creativity, thinking, and technical understanding. Moreover, I would like to acknowledge their valuable guidance and support to refine the problem statement as well as that to streamline my research direction.

I appreciate my dissertation evaluation committee for their valuable observations and insight recommendations during the dissertation defense. These comments enhanced the presentation and contents of the dissertation.

I am extremely grateful to all of my current and former Ubiquitous Computing Lab fellows and colleagues who have always provided me time, expertise, and encouragement. They were always present to guide me in various situations throughout my PhD journey. I would like to thank Dr. Wajahat Ali Khan, Dr. Taqdir Ali, Dr. Jamil Hussain, Dr. Bilal Ali, Dr. Syed Imran Ali, Dr. Musarrat Hussain, Dr. Tae Ho Hur, Dr. Hua-cam Hao, Dr. Nguyen Dao Tan Tri, Dr. Bilal Amin, Dr. Jae Hun Bang, Dr. Maqbool Hussain, Dr. Muhammad Afzal, Dr. Shujaat Hussain, Dr. Muhammad Asif Razzaq, Dr. Usman Akhtar, and Mrs. Seoungae Kim. They have contributed enormously in successfully performing various academic and personal tasks that confronted me

during my stay at South Korea.

I am very thankful to all of my colleagues for their kind support to my personal and academic life at Kyung Hee University. I am highly obliged to brilliant researchers Ubaid Ur Rehman, Abdul Muqeet, Salman Ali, Sheikh Salman Hassan, Muhammad Sadiq, Anees ul Hassan, Asim Abbas, Muhammad Zaki Ansaar, Dr. Waseem Hassan, Dr. Saeed Ullah, and Ahsan Raza. This journey would not have been possible without their support. They contributed a lot in to my personal and academic life to polish myself. Also, I appreciate all my Korean and international friends who worked as a team with me and developed my team work skills and provided me wonderful memories during my stay in South Korea. I would like to extend my sincere thanks to my kind and respected teachers at primary, high school, college and university level, who support and encourage me to pursue my higher study.

Last but not the least, I would like to express my sincere gratitude to my parents, wife, kids, and brothers for their endless love, support, prayers, and encouragement. Their support and encouragement has made this dissertation possible. I would like to extend my thanks to my friends in Pakistan, and other relatives who provide their kind support and encouragement to follow my dreams

Fahad Ahmed Satti

February, 2023

Table of Contents

Abstract i

Table of Contents vii

List of Figures ix

List of Tables xiii

Chapter 1 Introduction 1

1.1 Motivation 2

1.2 Problem Statement 5

1.3 Proposed Methodology Overview 6

1.4 Key Contributions 9

1.4.1 Sequence Contraction Algorithm 10

1.4.2 Sequence Expansion Algorithm 11

1.4.3 Semantic Reconciliation-on-Read 12

1.5 Thesis Organization 13

Chapter 2 Related Work 16

2.1 Clinical Text Mining 16

2.2 Sequence Contraction 17

2.3 Sequence Expansion 19

2.4 Semantic Reconciliation-on-Read 21

2.4.1 Big Data in Healthcare 21

2.4.2 Healthcare Interoperability 24

Chapter 3 Proposed Methodology 31

Chapter 4 Sequence Contraction 35

4.1 Research Objectives 36

4.2 Methodology 37

4.2.1 Pre-Processing 40

4.2.2 Concept Extraction 43

4.2.3 Threshold Selection for Sequence Classification 45

4.2.4 Concept to Schema Mapping 47

4.2.5 Expert Verification 48

Chapter 5 Sequence Expansion 49

5.1 Research Objectives 51

5.2 Methodology 51

5.2.1 Schema acquisition 53

5.2.2 Attribute to Sentence Transformation 55

5.2.3 Schema Map generation 56

Chapter 6 Semantic Reconciliation-on-Read 59

6.1 Research Objectives 61

6.1.1 Theoretical Representation 64

6.1.2 Implementation 68

Chapter 7 Results and Evaluation 74

7.1 Experimental Setup 74

7.1.1 Sequence Contraction 74

7.1.2 Sequence Expansion 76

7.1.3 Semantic Reconciliation-on-Read (SRoR) 78

7.2 Results 81

7.2.1 Sequence Contraction 81

7.2.2 Sequence Expansion 86
7.2.3 Semantic Reconciliation-on-Read 92

Chapter 8 Discussion 97
8.1 Sequence Classification its Implications 97
8.2 Value Extraction via Patterns vs UMLS 98
8.3 Validation on labeled data vs Verification on unlabeled data 99
8.4 Clinical perspective on formalizing the encounters 99
8.5 Schema Alignment 100
8.6 Patient Identification 101
8.7 Data Verification 102
8.8 Security and Privacy 103

Chapter 9 Conclusion and Future Direction 104
9.1 Conclusion 104
9.2 Future Direction 107

Bibliography 108
Appendix A List of Acronyms 128
Appendix B List of Publications 130
B.1 International Journal Papers [6] 130
B.2 Domestic Journal Papers [1] 131
B.3 Patents [1] 131

List of Figures

1.1	Technical motivation for achieving Data Interoperability	3
1.2	Operational motivation for achieving Data Interoperability	5
1.3	An overview of the proposed methodology	7
1.4	Interaction between the three novel solutions	8
1.5	A system's perspective of the proposed methodology	9
1.6	An abstract view of the Sequence Contraction Algorithm	10
1.7	An abstract view of the Sequence Expansion Algorithm	12
1.8	An abstract view of the Semantic Reconciliation-on-Read methodology for Data Interoperability	13
3.1	The proposed methodology workflow	32
3.2	The sequence contraction methodology	33
3.3	An abstract view of the sequence expansion method	34
3.4	Semantic Reconciliation using the proposed solution	34
4.1	An overview of the proposed methodology	39
4.2	Preprocessing methodology	41
4.3	Workflow for classifying the sequences as Medically aligned or not	43
5.1	Methodology for creating schema maps.	52
5.2	The five medical schemas used for achieving data interoperability.	53
5.3	An example of suffix arrays produced for the attribute "dateOfAdmission".	56
6.1	Semantic Reconciliation using the Ubiquitous Health Platform	60

6.2 Novelty of the proposed approach in terms of the semantic reconciliation pipeline 64
6.3 SRoR data representation 67
6.4 Schemas for Medical Fragments participating in SRoR 69
6.5 Modeling SRoR Maps 70
6.6 Class diagram, representing the SRoR model building application 72
6.7 SRoR Results for selected user 73

7.1 Plot between AuROC and threshold values between 0.0 and 1.0. 76
7.2 Iteration 1-5 results for C7 and C8 after executing Q1 and Q2 79
7.3 . Iteration 6 (a) results for C7 and C8 after executing Q3 and Q5 and (b) for C7 and C8 after executing Q4 and Q6 81
7.4 . Iteration 7 results for C7 and C8 after executing Q4 and Q6 82
7.5 Statistical information for MASS instances (a) showing the ratio of e with single vs multiple attributes, and (b) showing the unique labels and their extraction methodology. 84
7.6 Squared Pearson Correlation (r^2) of semantic sequence similarity between the annotated similarity and similarity computed by (a) by pretrained all-mpnet-base-v2 model, (b) pretrained DistilBERT-base-uncased model, and (c) fine-tuned DistilBERT-base-uncased model. 84
7.7 Cohen's Kappa (κ) score among the four annotators 86
7.8 Thresholds selection using MCC scores where t1 indicates the similarity threshold between $class_0$ and $class_{0.5}$, and t2 indicate the similarity threshold between $class_{0.5}$ and $class_1$. 88
7.9 A Heat Map showing the semantic similarity between the attributes, as indicated by the (mode of) annotated values. 89
7.10 Performance evaluation of various models using MCC and kappa (κ) scores. 92
7.11 A Heat Map showing the semantic similarity between the attributes, computed using all-mpnet-base-v2 pretrained model. 93

7.12 (a)Timeliness of recording medical fragments and their metadata in HDFS, (b)Time taken by Hive to create tables [C4, C5, C6], (c) Timeliness of retrieving medical fragments 94
7.13 Scalability in SRoR 96

List of Tables

4.1	Notations used in the manuscript	39
5.1	Sentence created from the attribute name “DateOfAdmission”	57
6.1	Comparison with existing platforms	63
6.2	Hive Queries	71
7.1	Evaluation criteria for each iteration	78
7.2	Dataset division in terms of its usage	82
7.3	Performance evaluation of the proposed methodology on labeled test instances and its comparison with the baseline methodology	85
7.4	Annotations performed by the four annotators on five medical schema	87
7.5	Performance matrix for individual classes using one vs all binarization technique	90
7.6	HDFS file size comparison for the medical fragments produced in Iteration 1-7	95

Chapter 1

Introduction

Recent advancements in information and communication technologies have led to the rapid expansion in development, deployment and usage of policies, software and devices towards better management of healthcare services [1]. Technologies, such as whole-exome and whole-genome sequencing [2], and precision medicine [3], along with smartphone based ECG, weight and activity monitors, and continuous glucose monitors [4, 5], besides others have made the traditional physician centric healthcare systems, financially unsustainable. This has also increased the number of available alternatives and caused an improvement in the quality of healthcare support systems and by extension the healthcare services, leading to an improved patient-centric diagnostic, treatment and follow-up process [6, 7]. However, this boom, has also led to a lack of interoperability between the participating software and devices [8], increased the disparity in the quality of healthcare data [9] and created communication and coordination gaps between the medical service providers and consumers [10]. Mitigating these problems, is of utmost importance for achieving ubiquitous healthcare.

This dissertation investigates an end-to-end methodology for resolving the challenges pertaining to Data Interoperability between heterogeneous medical systems in a schema agnostic manner. This resolution entails, design and implementation of a methodology to acquire structured data from clinical conversations via semantic Sequence Classification, creation of maps between formal and informal data schema, and a platform which can archive medical data, in a form closer to its acquisition format and the application of Semantic Reconciliation-on-Read to transform data into the latest version of schema compliant with the consuming healthcare platform.

Consequently, this novel solution will bridge the gap between healthcare providers and significantly reduce the redundancy in acquiring patient data and creating a comprehensive medical

history for them from heterogeneous medical data sources. It will also jumpstart the standard compliance process for organizations in low-income countries, without requiring significant, last mile, interventions.

1.1 Motivation

Technological advancements in Information and Communication Technology (ICT) has provided a boost to the quality and quantity of healthcare services. A plethora of policies, software, and devices have been developed to introduce new and extend the reach of existing best practices [11]. While the benefits of these initiatives are well documented and acknowledged in the literature, in more practical terms, access to effective digital healthcare services in the developed world versus the developing world is skewed. Similar to many other domains, application systems in healthcare also leverage the ease of implementation and deployment, enabled by relational database management systems and rule based systems. However, these traditional systems and the underlying healthcare processes are unable to deal with the large volume of patients. This problem is more prominent in the developing world, where healthcare resources are constantly under stress. Figure 1.1, illustrates the technical gap between healthcare systems whereby sharing of patient data beyond organizational boundaries in an effective manner still remains a distant reality.

The amount of time available for the physician to diagnose and prepare a treatment plan for a patient is very limited. With fewer healthcare providers and a lack of resources, this problem greatly affects the developing world. Thus in order to help mitigate this problem, it is necessary to create a mechanism to automate the acquisition of patient data and its integration with existing medical record of the correct patient, in a safe and privacy enforced manner. One of the key sources of medical data, which currently remains untapped, is the conversation between the physician and the patient. Here the physician collects information from the patient, about various aspects of the medical diagnosis procedure and provides information to the patient about prospective treatment plans. If a digital healthcare environment is present at the clinic or hospital, where this conversation takes place, the physician then has to enter the same data into an HMIS, using some kind of a form. With a high patient load, this redundancy is abhorred by the medical practitioners and the effectiveness of a digital healthcare solution, loses its apparent effectiveness. Hence the collection

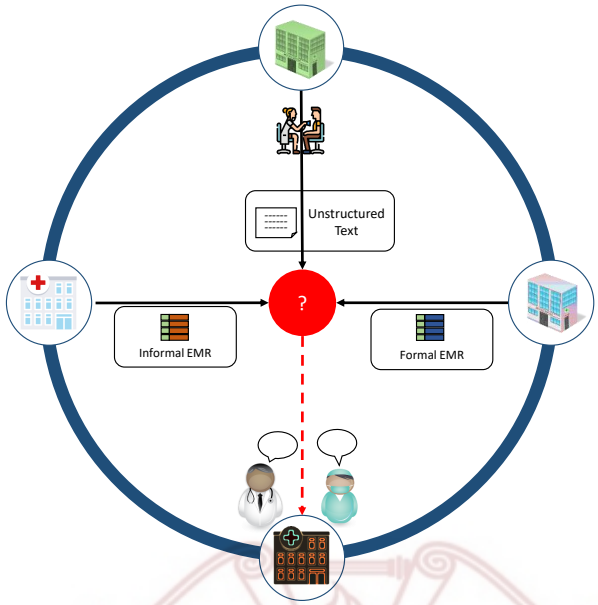

Figure 1.1: Technical motivation for achieving Data Interoperability

of data in an automated manner from unstrcutured text and informal Electronic Medical Records (EMR) is an absolute necessity for ubiquitous healthcare and to reduce the stress on healthcare resources.

In the developed world, healthcare providers have started to solve the integration problem by moving away from the traditional physician and hospital/clinic centric approach to a more patient oriented one. The EMR, which were being used for capturing the relevant patient information, and were bound to the collecting organization only, are now being enriched with Patient Health Records (PHR), Medical Images, and many other sources to produce the Electronic Health Record (EHR). Proprietary solutions (such as Essentia Health , Omni MD , and BlueEHR), and open source ones (such as openMRS and openEMR[1]) are able to create a complete digital persona of a patient, by taking into account both direct data sources (e.g. the physician) and indirect data sources (e.g. insurance records). Additionally with 63% of the world population, now connected with the internet [12] the necessity to connect with the global village and provide state-of-the-art medical interventions across borders, has grown rapidly.

[1]OpenEMR: https://www.open-emr.org/

In order to bridge the gap between formal EMR, numerous endeavors have been undertaken to create standards for storing and exchanging data, such as Health Level 7 (HL7) [13] based Clinical Document Architecture (CDA) and Fast Health Interoperability Resources (FHIR), openEHR, Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT) [14], Logical Observation Identifiers Names and Codes (LOINC) [15], and others. Based on these standards, expert driven initiatives such as the Clinical Information Modeling Initiative (CIMI) [16] have shown great promise. CIMI aims to integrate the best features of HL7v3 and openEHR. Similarly, the Yosemite group [17] is utilizing cloud sourcing to create the mappings between various standards. However, the heterogeneity in healthcare standards and data remains a major challenge, which prevents integration, exchange and effective utilization of medical data, across system boundaries (as defined by IEEE 610.12) [18]. The key to solving this problem lies in identification of relationships between the participating schemas, which can be achieved by using schema matching and schema mapping approaches [1]. Data Interoperability between healthcare data and information management systems, allows the participating organizations to directly share their respective data, so that it can be used by consumers in a seamless manner. Such an implementation can not only benefit the physician and the patient by reducing overhead and redundant costs and saving time, but can also prevent operational waste, and support policy makers in improving accountability and privacy [19].

However, the motivation behind an interoperable healthcare system far exceeds the current hurdles. As shown in Figure 1.2, an interoperable healthcare environment, removes the data acquisition redundancies, provides rich clinical histories by integrating multi-modal patient data, jump starts standard compliance for the small and mid scale hospitals, especially in the developing world, there by reduces stress on hospitals and clinics and provides safe and secure archiving of patient medical data.

This thesis provides an end-to-end solution for automating one aspect of the data acquisition problem by automating the collection of medical data from unstructured medical text based on the Sequence Contraction methodology, aligns the resulting semi-structured data schema, with formal and informal schema based on the Sequence Expansion methodology, and provides a framework to integrate these two solutions into a Semantic Reconciliation-on-Read data curation engine.

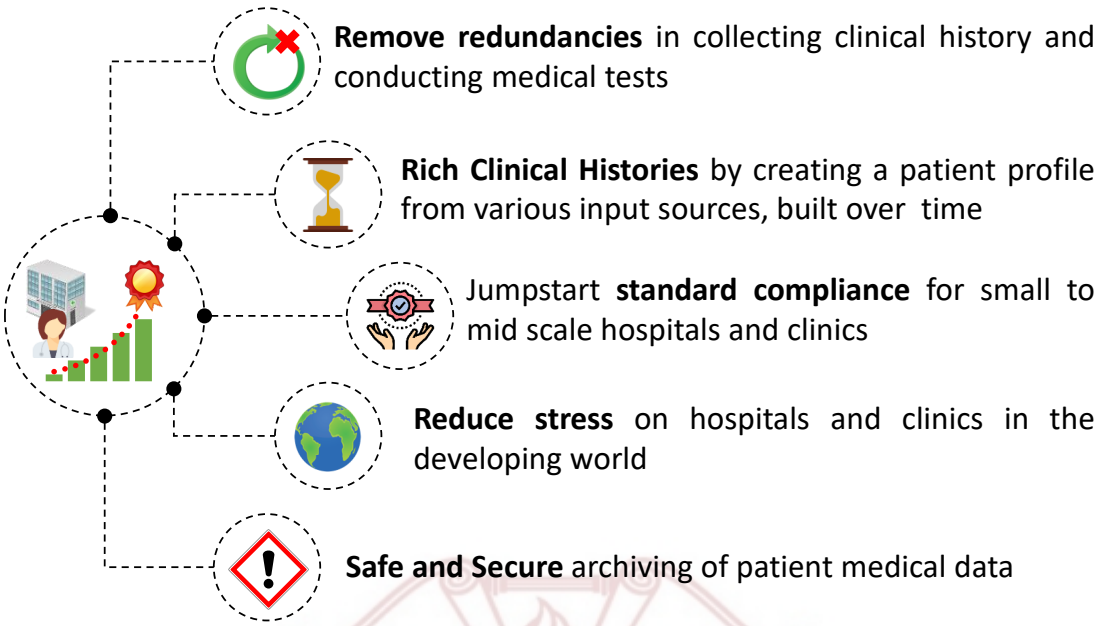

Figure 1.2: Operational motivation for achieving Data Interoperability

1.2 Problem Statement

[20] has highlighted the most prominent operational factors influencing the differences between healthcare services, which includes the availability of specialists and equipment, for adapting, developing, and using global standards and technology. In high-income countries, spurred by the effectiveness of technology, variety of sources, and complexity of domain requirements, many novel platforms have been proposed and are being utilized to improve the clinical interactions [21–23]. On the contrary, in low-income countries, financial limitations, increased patient load, and the availability and access to healthcare experts, clinical facilities (public and private setups), in-patient care, internet, and electricity, have a very large impact on the healthcare services [24,25]] Many commercial solutions are financially not feasible for low-income countries, and many open-source solutions, such as OpenEMR or GNU Health[2], are difficult to adapt, without substantial intervention by ICT experts.

In particular, existing health interoperability solutions are expert-driven and standard dependent, which require a lot of resources for converting legacy systems into a globally connected

[2]GNU Health: https://gnuhealth.org/

one. Additionally, this conversion can also result in data loss, due to the gap between legacy and informal schema compliant medical systems and a standard one.

A comprehensive solution to this problem should enable identification and extraction of clinical data from a formal or informal producer and transform it into a form usable by the consuming medical system. In particular, the design and implementation of an effective data interoperability methodology is based on the resolution of the following three research questions.

1. How to Identify & Extract clinical attributes and their values from unstructured text?

   aim: Find attributes and values from clinical sequences.

2. How to Align attribute-value pairs with structured schema?

   aim: Align Attributes with heterogeneous schema for data format transformation

3. How to Design a scalable infrastructure, automating data interoperability?

   aim: Design a practical platform which supports mapping evolution and low resource usage.

1.3 Proposed Methodology Overview

Digitization of clinical encounters necessitates the creation of a data interoperability framework which can operate on structured as well as unstructured data, and can create a bridge between heterogeneous medical data curation platforms. In order to create this bridge between medical data obtained from formally structured sources, such as HMIS and informally structured sources, such as medical reports, clinical conversations, a specialized methodology is presented in this dissertation. It is comprised of three process; Sequence Classification, Schema Alignment, and Semantic Reconciliation-on-Read support engine, which work in tandem to achieve the aims of standard-agnostic data interoperability, as shown in Figure 1.3.

Sequence Contraction methodology, is used to identify and extract semi-structured data elements from unstructured text representing the clinical conversations between physicians and patients. On the other hand, structured data can be converted into semi-structured format using naive serialization techniques. The output thus produced, provides the medical data representing the patients and its semi-structured schema. In order to align the semi-structured schema, Sequence

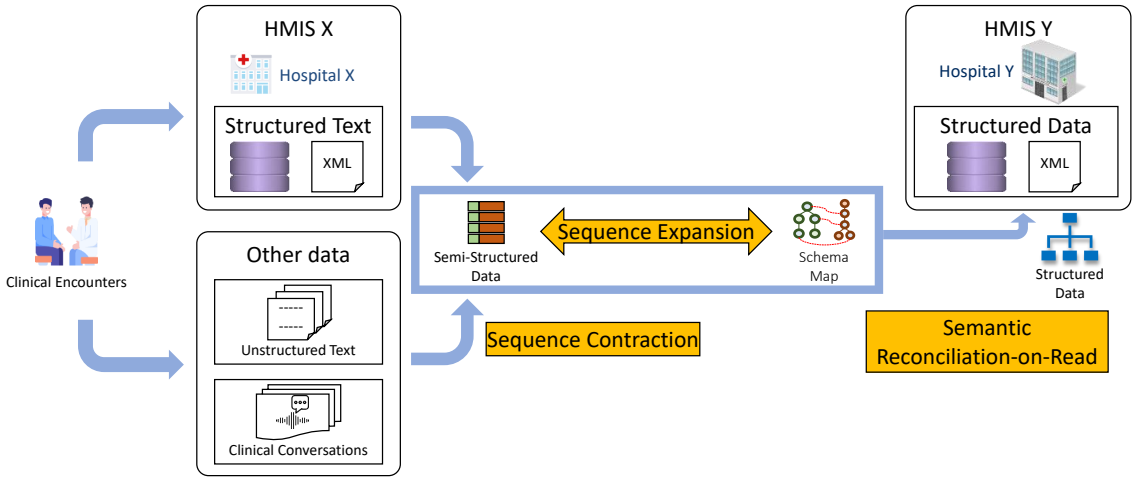

Figure 1.3: An overview of the proposed methodology

Expansion methodology is used, where attribute names are converted into sequences which are used to identify similarity between the source attributes in an unsupervised manner. Finally, the objective of Semantic Reconciliation-on-Read methodology is to provide a scalable and usable engine to archive semi-structured medical data and by using the schema map from process two, apply semantic reconciliation to transform it into a target schema. The interaction between these three methods is shown in 1.4.

Once implemented, the novel algorithms presented in this dissertation are used in a data interoperability engine, which is shown in 1.5. In particular, the Sequence Contraction Algorithm is used in the Data Acquisition phase, where it operates on unstructured text and converts it into semi-structured form. This pathway, utilizes the NLTK library to create sentences and apply preprocessing on the unstructured textual data. UMLS is used to collect the semantic concepts associated with n-gram words in the text sequences. The other part of this phase provides a conversion of structured data into a semi-structured form, which can be achieved by a naive serialization based method, which converts relational data or structured medical reports into attribute-value form. The Sequence Expansion algorithm is used in the Schema Alignment phase, and provides a mechanism to identify the mapping criteria between heterogeneous attributes. This criteria is used to build a schema map between any two, standard-agnostic schema. These schema are obtained from the Structured Data or as a consequence of the Sequence Contraction process in the Data

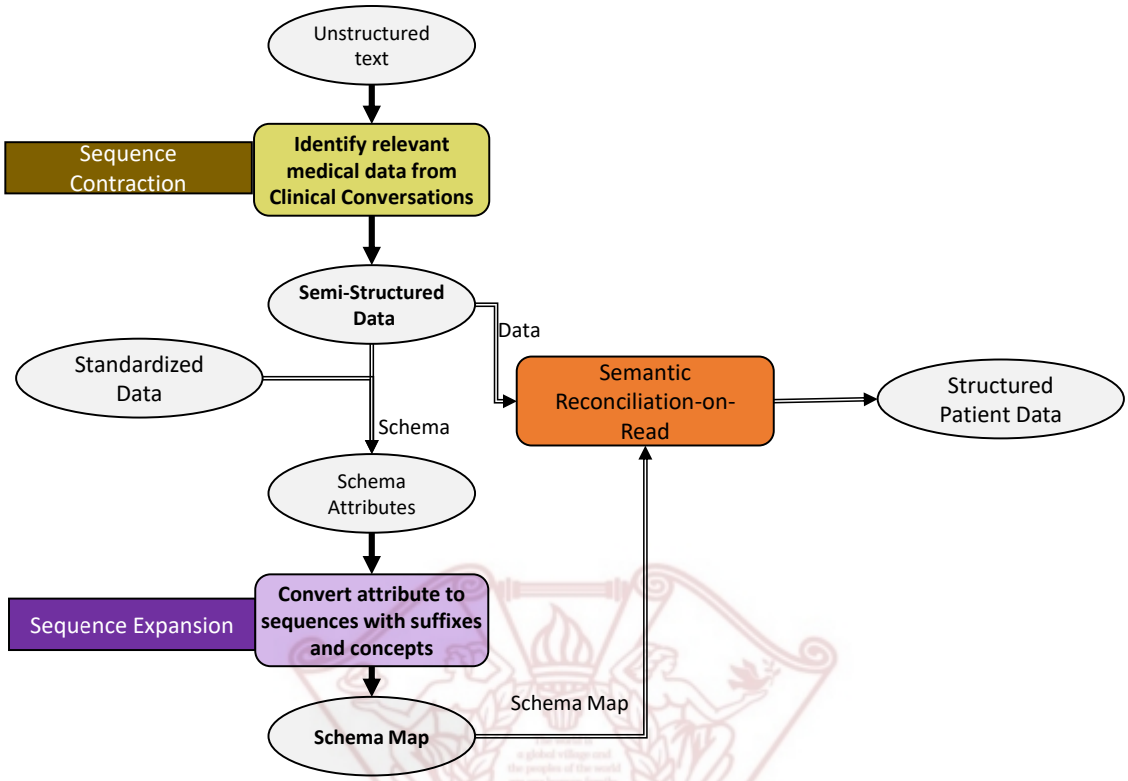

Figure 1.4: Interaction between the three novel solutions

Acquisition phase. Here UMLS is used to filter the suffix array to medically aligned concepts only and to semantically enrich the sequences, before they can be used for positional semantic similarity matching. Finally the Semantic Reconciliation-on-Read method provides an archiving and processing framework, for transforming semi-structured data into a target structured data. This utilizes the Schema Map and the data produced by the Schema Alignment module and Data Acquisition module, respectively.

The process to collect data and build the Schema-Map are seldom occuring ones, while the output of the Semantic Reconciliation-on-Read is triggered by a medical expert looking up some data. The architecture is supported by Hadoop providing the NoSQL data engine, which persists the semi-structured medical records of the patient collected from remote HMIS and after the application of Sequence Contraction on unstructured text. It also holds the Schema-Maps produced by the Schema Alignment module. In order to retrieve the medical records Hive is used to temporarily build a SQL compliant interface first and then Semantic Reconciliation is applied on the

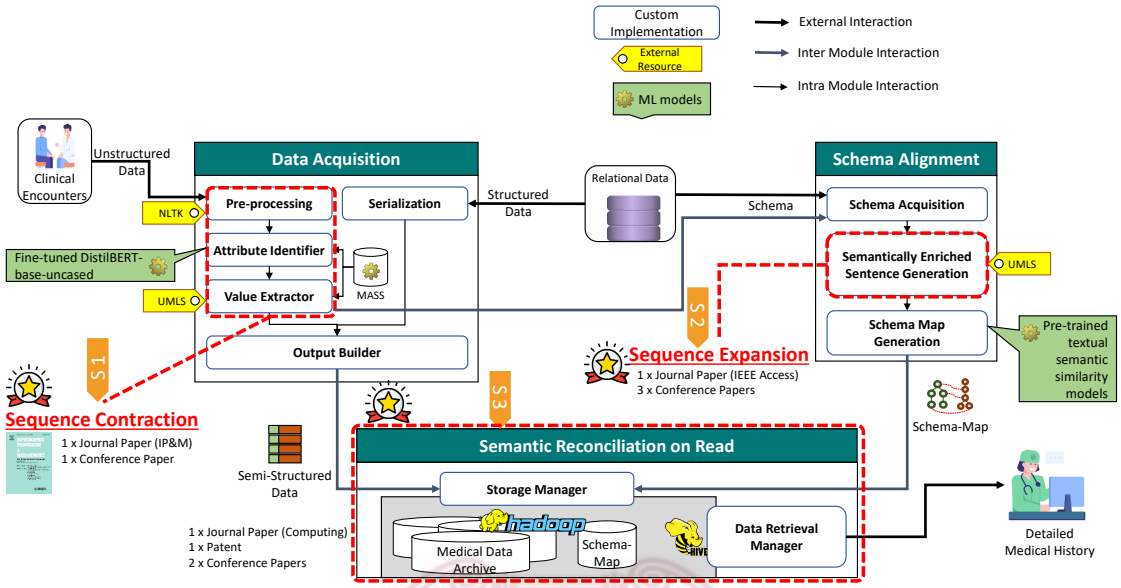

Figure 1.5: A system's perspective of the proposed methodology

medical data with its corresponding Schema Map. In order to support the positional semantics based similarity matching, various pre-trained ML models are used. For the Sequence Contraction process, we have fine-tune the DistilBERT-base-uncased pretrained model, with the conversational sequences obtained from real world data. While for the Schema Alignment process, we have used various pre-trained semantic textual similarity models.

1.4 Key Contributions

The goal of this research work is to provide an end-to-end methodology for achieving Data Interoperability between heterogeneous medical data, in Structured and Unstructured form, acquired from HMIS and clinical conversations, respectively. While the overall end-to-end architecture for Data Interoperability is shown in 1.5, the key contributions, as presented in this dissertation include the Sequence Contraction algorithm, Sequence Expansion algorithm, and Semantic Reconciliation-on-Read framework. These are further explained in the following sub-sections.

1.4.1 Sequence Contraction Algorithm

Formally, the Sequence Contraction algorithm is defined by Equation 1.1. For an unstructured corpus C, sequence contraction process pertains to the creation of the η function that can take a portion of C as input and produce a probable medical artifact p. This artifact in turn contains both a name of the attribute and its value. In order to build η, Transfer Learning is used to utilize a sentence encoder for classifying sequences, followed by the application of syntactic and semantic extractors for creating p.

$$\underbrace{\exists C}_{\text{Unstructured corpus C}} \wedge \exists \eta | \forall c \in C . \eta(c) \rightarrow p | p \in C \qquad (1.1)$$
$$p =< p_a, p_v > | p_a \models p_v$$

Figure 1.6 illustrates, in an abstract manner, inner workings of the novel η function.

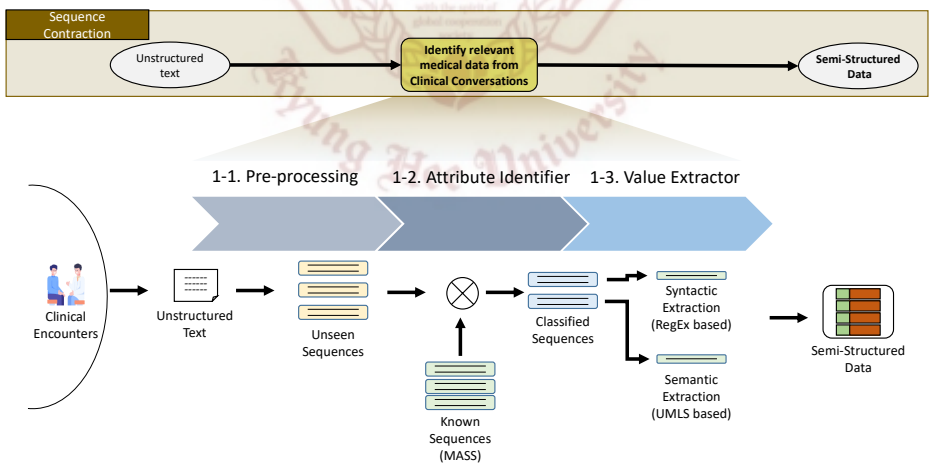

Figure 1.6: An abstract view of the Sequence Contraction Algorithm

Here the unstructured text is first converted into sequences, which represent the statements and question/answers text sentences. Each sequence is encoded to form an embedding vector,

which is compared against all the instances in a pre-built set of medically aligned sequences. Using an optimal threshold value, and cosine similarity between the embedding vectors, the unseen sequences are classified to produce the probably medical sequences. These probable sequences are marked with attribute names, obtained from the classification instance of known sequences. Using syntactic or conceptually semantic extraction methodology, a value corresponding to the attribute name is extracted. This attribute and its value for the pair, which is used as an atomic element of the semi-structured data.

1.4.2 Sequence Expansion Algorithm

The Sequence Expansion Algorithm is represented by Equation 1.2, where S_1 and S_2 are two semi-structured schema with heterogeneous attributes. The attributes of these schema do not follow any preset pattern or a standard and are defined in an adhoc manner by the storage designer. In order to align the elements p and q belonging to S_1 and S_2, respectively, the χ function is defined. χ is a total function which returns 1 when the attribute p_a is semantically similar to q_b. It returns some similarity between 0 and 1, if they are nearly identical and 0, if they are not. Thus, in order to define this function p and q must first be encoded into a format, which allows a semantic comparison to be made between the two elements.

$$\begin{gathered}\text{Semi-Structured Schema} S_1, S_2 \\ \forall p \in S_1 \wedge q \in S_2 \\ \chi(p,q) = \left\{ \begin{array}{l} 1 \;\; if \;\; (p_a = q_b) \\ \sim \;\; if \;\; (p_a \cong q_a) \\ 0 \;\; otherwise \end{array} \right\}\end{gathered} \tag{1.2}$$

Using an unsupervised approach, each attribute p or q is converted into a sequence using Suffix Array generation, UMLS based filtering, and conceptual semantics inclusion. These sequences are then encoded into an embedding vector form, which is used to identify the positional semantic similarity between the source attributes. An abstract view of this algorithm's workflow is shown

in Figure 1.7.

1.4.3 Semantic Reconciliation-on-Read

While the main novelty of the Sequence Contraction Algorithm and the Sequence Expansion Algorithm is the formulation of an automatic mechanism for extracting medical data and schema alignment, respectively, their application in the real-world necessitates the existence of a framework for integrating these solutions and providing a holistic solution for Data Interoperability.

Using the best engineering process from Big Data Curation engines and relying on the schema-on-read property, the corresponding Semantic Reconciliation-on-Read methodology collects medical data in semi-structured format and Schema Maps, produced by the Sequence Contraction and Sequence Expansion implementations. The overall process is illustrated in Figure 1.8.

Here, Patient Medical Records, conforming to EHR Y from Hospital Y, is first stored into the Medical Data Archive, after being converted into Semi-Structured form. This conversion from Structured data into Semi-Structured data is performed via naive data serialization. Subsequently, unstructured data is converted into Semi-Structured form A, conforming to the schema X, as utilized by the Sequence Contraction methodology. This data is also saved into the Medical Data Archive. Additionally, a Schema Map is created between X and Y by the Sequence Expansion methodology, which is stored in the Schema Map store. Eventually, when a medical expert for Hospital X, wants to read the patient's data, then the Data Retrieval Manager, collects the pa-

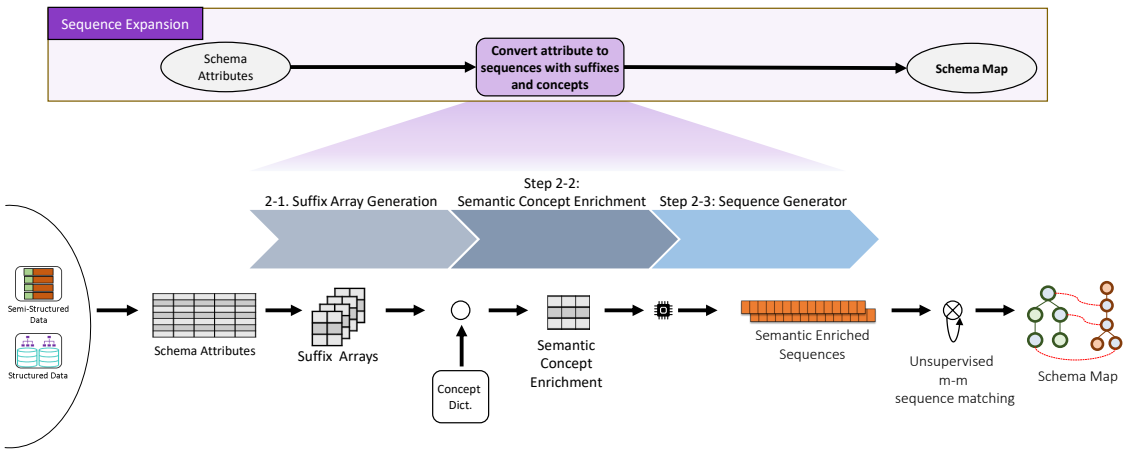

Figure 1.7: An abstract view of the Sequence Expansion Algorithm

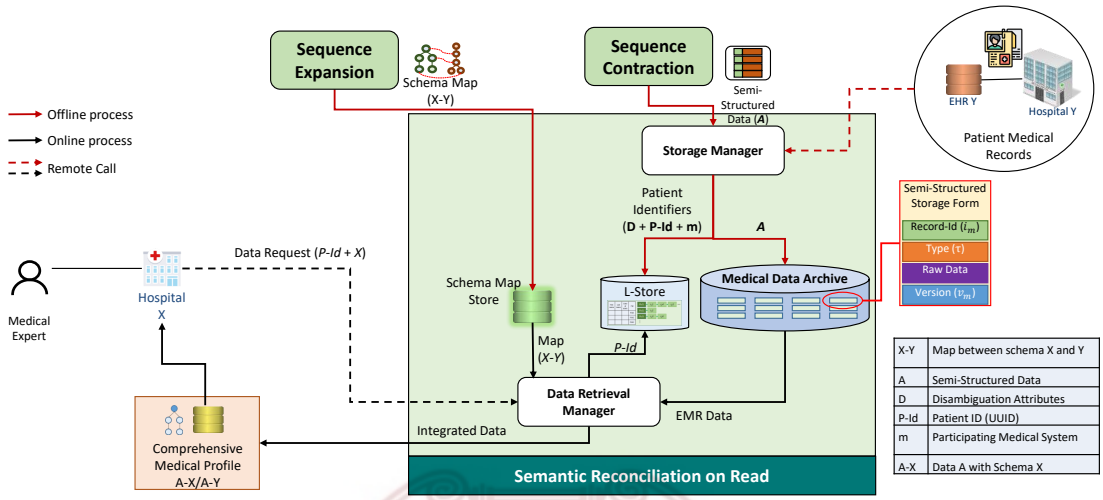

Figure 1.8: An abstract view of the Semantic Reconciliation-on-Read methodology for Data Interoperability

tient's medical records (A) using a using patient id (P-id). It then collects the Schema Map (X-Y), between the source schema Y and the requested Schema X. Using these components, a comprehensive medical profile of the patient is built into the form of A-X (where data A conforms to the schema X). This response in finally sent to the medical expert, completing the Semantic Reconciliation-on-Read process. It is pertinent to note here that this transformation is only temporarily applied, using the latest version of the data and the schema map, thus enabling the usage of version control for managing data and schema evolution and ensuring that the original data only loses the explicit relational relationships, if defined in the EHR-Y. This ensures, that the actual data of the patient remains free from error or unintentional modifications.

1.5 Thesis Organization

This dissertation is organized into chapters as follows.

- **Chapter 1: Introduction**. Chapter 1 provides an overview of the standard-agnostic data interoperability methodology. In this regard, the motivation for utilizing an automated method

and sequence similarity for this text processing task is emphasized. Furthermore, problem statement and overview of the proposed methodology is also put forward in this chapter. In the end, the key contributions of the dissertation are discussed.

- **Chapter 2: Related Work**. Chapter 2 focuses on the literature review for similar approaches for clinical text classification, sequence contraction, sequence expansion and semantic reconciliation-on-read. The key limitations of the existing approaches are also identified and enlisted here. Finally, this chapter summarizes how the identified limitations are mitigated via the proposed solutions.

- **Chapter 3: Proposed Methodology**. In Chapter 3, we present the proposed end-to-end methodology for standard-agnostic data interoperability. This chapter deals with the three building blocks of the methodology, namely, sequence contraction, sequence expansion, and semantic reconciliation-on-read.

- **Chapter 4: Sequence Contraction**. Chapter 4 provides the details of various efforts made for robust classification of clinical conversations into medically aligned data elements. It elaborates the proposed ML assisted attribute identification and value extraction process, based on semantic similarity.

- **Chapter 5: Sequence Expansion**. The novel methodology for identifying the semantic similarity between standard-agnostic attributes is presented in Chapter 5. In particular, this chapter delves into the detail of how string tokens are converted into syntactically and semantically enriched sequences, which are used for unsupervised Schema-Map generation.

- **Chapter 6: Semantic Reconciliation-on-Read**. Chapter 6 details the integration of semi-structured data and schema maps to resolve the Data Interoperability problem. Based on a NoSQL data storage and processing engine, the transformation of medical data into a target schema is provided in response to a demand, ensuring the usage of latest schema maps.

- **Chapter 7: Results and Evaluation**. The results and evaluation of various portions of the proposed methodology are explained in Chapter 7. Firstly, it explains the text classification results for the real world Test dataset. Secondly, the sequence expansion approach and its

effects on the Schema Alignment are presented in this section. Finally, the scalability and timeliness of the proposed Semantic Reconciliation-on-Read platform is presented. Additionally the effects of the complete methodology application on individual conversations to produce structured data is presented in this chapter.

Chapter 8: Discussion. This Chapter 8 introduces the limitations of the proposed approaches, implications of the achieved results and various other caveats effecting the proposed methods and the platform.

- **Chapter 9: Conclusion and Future Direction**. This Chapter 9 concludes the thesis and also provides future directions in this research area. The main contribution of the thesis is also highlighted in this chapter.

Chapter 2

Related Work

Patient clinical histories, such as previously diagnosed diseases, provided treatments, allergies, and others, play a pivotal role in healthcare decisions. The requirement of clinical record tracking and timely access initiate the idea of clinical record storage and maintenance systems such as EHRs. The history of clinical records can be linked back to the fifty century B.C. when Hippocrates specified two of its aims, including correctly reflecting the course and potential cause of a disease [26]. The modern EHRs started to appear in the 1960s, supplementing the prescribed goals with additional functionalities. The clinical records contain both structured and unstructured information however, about 80% of clinical observations are not directly machine-understandable due to its unstructured format [27]. The unstructured clinical text is one of the most significant barriers of EHRs and clinical data in quality improvement, operations, and clinical research [28]. Starting from the 1940s till to date, Natural Language Processing (NLP) has made tremendous advancements in processing narrative text for various tasks such as Machine Translation, Automatic Summarization, Co-Reference Resolution, Discourse Analysis, Named Entity Recognition, information extraction, etc. [29, 30].

2.1 Clinical Text Mining

Clinical text withholds valuable information, including symptoms, diagnosis, treatment, medication details, and follow-up plans that can help in improving healthcare service provision. Clinical text mining refers to the automatic processing of a clinical text for understanding and interpretation of the content [31]. Plenty of research has been conducted to extract valuable information out of this text. The field of clinical text mining has advanced rapidly transitioning from hand crafted rule based methods to machine learning and recently more advance approaches such as

deep learning for information extraction and modeling [28].

2.2 Sequence Contraction

Medical information extraction is a challenging task of automatically deriving high-quality structured information from text. Several research initiatives have aimed to solve related problems and achieved very good results. Some of these will be explained in this section.

Automatic keyword extraction has gained a lot of traction in the research community [32] as it pertains to extracting potential information from raw textual data with minimum human intervention. The Named Entity Recognition (NER) task in NLP is also related to this problem, whereby techniques are devised to build models for identifying attributes of interest, according to some preset features (such as identifying all persons, cities, and others), in text [[33,34]]. Similarly, the task of identifying sequences in text, pertaining to some input attribute names (similar to preset features) is also related to NER [[35,36]].

These tasks are specially popular in the domain of e-commerce, where product attribute extraction is used to identify the implicitly defined characteristics of a product, from its description. Thus, retailers can continue describing their products in a poignant manner, while businesses can use state-of-the-art tools and techniques to identify the key features, necessary to forecast demand, optimize search, and provide contextual recommendations to the buyers [[37–40].]

In the past, rule-based approaches, such as [41], [42], and [43] have been proposed, which typically used regular expressions obtained from domain knowledge. However, these and other rule-based techniques [44], suffer from the generality problem and are unable to replicate their performance, for any text, syntactically and semantically different from the source, as shown by [41] and [45]. Ontologies and semantic web-based solutions can help resolve the semantic matching problem, however, these solutions require a large amount of human effort to build the semantic knowledge graphs. For shortened text, such as that provided by classified ads, identification of attributes and their values is an enhanced challenge. Due to restrictions on the number of words used in these ads, a specialized NER methodology is needed. [33], presented a supervised NER methodology and evaluated the performance of the supervised Hidden Markov Model (HMM), Support Vector Machines (SVM), Maximum Entropy (MaxEnt), and Conditional Random Fields

(CRF). The authors found that SVM or MaxEnt when used with the Viterbi algorithm to smoothen their prediction, provides the best classification results. The main drawback of these techniques is the use of initial seed lists and concept dictionaries, which have to remain ever-evolving.

[46] proposed a system that converts a medical text into a table structure. The proposed system is based on a medical event recognition module and an SVM-based negative event identification module. CRFs are also used in the extraction method. This study specifically examines patient discharge summaries generated by medical personnel. [47], introduced a generic model with a feed-forward neural network and word embedding to attain high performance in various NLP tasks, such as part-of-speech tagging, chunking, NER, and semantic role labeling. [48], used an unsupervised Fine-Grained Entity Recognition (FIGER) model, which can provide an automatic tagger for text, using a trained CRF model for text segmentation, followed by an adapted perceptron algorithm for multi-class, multi-label classification.

[49] explored the application of neural network-based models to produce word embeddings for biomedical text. It is demonstrated that these embedding approaches generate vector representations that capture useful semantic properties and linguistic relationships between words. Unstructured text often consists of typographical errors and abbreviations, which act as an impediment to improving the performance of the word embedding-based approaches. [50] presented an approach based on Bidirectional LSTM (Bi-LSTM) with character level embeddings to avoid this problem and achieve better performance. Similarly, [51], used a Bi-LSTM and CRF-based architecture to identify drug names, using both word-based and character-based representations of each word.

A transformer-based sequence labeling architecture, called AdaTag, is proposed by [52] for multiple attribute value extract task. AdaTag uses adaptive decoding in which the decoder is parameterized with pre-trained attribute embeddings through a hyper network and a Mixture-of-experts module. This allows for separate, but semantically correlated, decoders to be generated on the fly for different attributes. [37] proposed a methodology to extract missing attribute values from a free text input such as product profiles. The methodology can leverage open-world assumptions in which case the possible set of values are not known beforehand. [53] proposed a high precision and scalable framework for extracting numeric attributes from product description text. A distant

supervision approach is used for training data generation and removing dependency on manual labels. Moreover, a multi-task learning architecture is proposed to deal with missing labels.

[54] demonstrated that neural network-based representations e.g. word2vec, Glove, fastText, dramatically improve the performance of natural language processing tasks such as concept extraction. Recently, as demonstrated by [55] and [35], more advanced neural embedding methods and representations (such as Elmo and BERT) have further pushed the state of the art in NLP. [56] proposed a deep learning-based approach to extract medically relevant attributes from electronic medical records, using ALBERT model, which provides much better results than the traditional LSTM-CRF model. [57] have formulated a biomedical entity recognition task as a machine reading comprehension problem, which achieves good performance on six BioNER datasets. The proposed formulation can introduce more prior knowledge through well-defined queries.

In summary, research trends have moved from rule-based syntactic matches to supervised learning, then to unsupervised learning, and to most recently hybrid learning methods leveraging syntactic and semantic matching.

2.3 Sequence Expansion

Althubait et al. [58] proposed an ontology expansion methodology that identifies and extracts new class from text articles using word embedding and machine learning techniques. The authors identified the similarity of tokens and phrases of the text articles with the exiting classes of the ontology. The target ontology is expanded with classes from text articles having greater similarity with that of already added classes. A similar word embedding technique was also used by Nozaki et al. [59], where the authors used instance based schema matching technique to identify the semantic similarity between two instances. The results of the study showed the possibility of detecting similar string attributes of different schemas. Yousfi et al. [60] also utilized semantic base techniques and proposed xMatcher XML schemas matching approach. xMatcher transforms schemas into a set of words, followed by measuring words context, and relatedness score using WordNet. The terms from different schemas having similarities greater or equal to 0.8 are considered similar. Bylygin et al. [61] devised an ontology and schema matching approach by combining lexical and semantic similarity with machine learning approaches. The authors used lexical and

semantic measures as features and trained various machine learning algorithms including Naive Bayes, logistic regression, and gradient boosted tree. The result achieved showed that the combination of algorithms outperformed the single modal.

Martono et al. [62] provided an overview of a linguistic approach for schema matching. The authors presented various linguistic methods to identify token strings in element names, followed by similarity evaluation between various schema. The process starts by normalizing the strings, which can be achieved via tokenization, generalization, elimination or semantic tagging. The normalized strings are then categorized based on their information relatedness. Elements belonging to the same categories are then compared with each other using two similarity measures, which include Lavenstein distance between words and Jaro-distance between 3-grams character substrings. Alwan et al. [63] has summarized the techniques used in literature for schema matching based on database schemas and instances. These techniques can be categorized based on the type of information used for schema matching which includes schema level, instance level, hybrid (schema and instances) and auxiliary (which can include information from external sources). Accordingly, most research is focused towards schema level and instance level approaches which can utilize syntactic techniques (such as n-gram, and/or regular expressions) and/or semantic techniques (such as Latent Semantic Analysis, WordNet/Thesaurus, and Google Similarity), to achieve data/information interoperability. Kersloot et al. [64] performed a comprehensive systematic review to evaluate Natural Language Processing (NLP) algorithms used for clinical text mapping onto ontological concepts. The authors categorized the findings of various studies based on the use of NLP algorithms, data, validation and evaluation techniques, result presentation, and generalization of results. The authors revealed that over one-fourth of the NLP algorithms used were not evaluated and have no validation. The systems that claimed generalization, were self evaluated and having no external validation.

Xu et al. [65] presented a framework for discovering indirect links besides direct links among schema elements. The indirect matches were detected for relations such as union, composition, decomposition, selection, and boolean. The indirect links are useful to handle concepts merge, split, generalization, and specialization. The matching techniques utilized in the study considered terminological relationships (word synonym and hypernym), structural characteristics, data-value

characteristics, and expected data values. The experimental results revealed framework effectiveness by achieving more than 90% precision and recall for direct and indirect link matching.

2.4 Semantic Reconciliation-on-Read

2.4.1 Big Data in Healthcare

One of the consequences of the changing healthcare environment is the production of heterogeneous, voluminous, medical data which necessitates the creation of comprehensive medical profile of the patient to improve healthcare service delivery. In particular Clinical Decision Support Systems(CDSS) require the combination of several data sources, such as diagnostic tests, patient's clinical history, CPG, vital signs, symptoms and others, to aid the decision making process [66].

Traditional healthcare systems have focused on using relational databases for persisting EHRs. Based on the idea of a well-structured storage solution, with the ability to uniquely store and identify tuples and their inter-relations, relational databases are beneficial for small to medium scaled medical systems, with little to no interoperability. Other research led initiatives are now turning towards NoSQL technologies [67] [68] [69] such as cloud based Column Oriented data store for storing healthcare data in HL7 v3 form by Celesti et al. [70], which provides very low query(with aggregation and filter operations over column data) execution times on very large amount of data, and Graph DB utilized by Balaur et al. [71] to integrate statistical data on molecular interdependencies from a manually curated and annotated relational database. The usecase, of retrieving related medical records for a patient, necessitates the use of a document oriented data store, which can hold each EHR record as a document. While a lot of effort has been put into developing proprietary solutions (like Essentia Health[1] , Omni MD[2] , and BlueEHR[3]), and some open source ones (openMRS[4] and openEMR[5]) which can capture heterogeneous data and create an EHR, there is a general lack of Big Data solutions for the healthcare market [72]. While there is no formal definition of the term "Big Data", any data will require a specialized storage and pro-

[1]Essentia Health: http://www.essentiahealth.org
[2]https://www.omnimd.com/
[3]https://blueehr.com/our-services/electronic-health-records/
[4]https://openmrs.org/
[5]https://www.open-emr.org/

cessing engine, if it has the following 5 properties (also known as the 5 Vs of Big Data), Volume, Velocity, Variety, Veracity, and Value [73].

Volume Medical data can be classified into two types, primary data sources and secondary data sources [74]. Primary data sources require direct interaction with the patient for data creation. On the other hand, Secondary sources, represent the knowledge management systems, clinical research systems, Biobanks and other tools used by epidemiologists and medical experts, which provide supplementary diagnosis, treatment, and follow-up plans, based on indirect observations (e.g. environment and general living habits). Compounded by the number of patients (e.g. 500,000 participants in UK Biobank [75], 100 million for mendelian disorder risk [76], EHR4CR project with 45 partners in EU [77]) and medical IoT(producing streaming data using body sensors) the storage requirement for a comprehensive digital health persona has already grown beyond the scalability, and speed of traditional relational databases.

Velocity Healthcare data producers, emit data at different rates, pertaining to the use of information systems or medical devices. While medical information and knowledge systems, produce non-streaming data, which is seldom updated (relatively). Medical IoT can produce streaming data, which is continuously produced and has to be shared in real-time [78, 79]. E.g. a heartrate monitor on a smart watch produces many instances of very shallow data, while the EHR is longitudinal and deeper, with infrequent instantiation. This requires the use of specialized hardware with low latency, high reliability, and rapid access to the data.

Variety Variety or Heterogeneity in healthcare data, stems from the existence of a large number of formal standards [1] and non-formal/custom standards [80]. This has led to the creation of several semantic reconciliation techniques and platforms which can resolve interoperability among the EHRs [81]. Medical systems also suffer from a variety of purpose, whereby they are created and used to serve the patient (e.g. smart watches), the medical experts, organizations (hospital and/or insurance companies), or environment (e.g. government, consortium) [82]. Consequently, the data produced by these systems only conforms to their own abstraction level. This means, if an HMIS has to be used for running a small clinic, in a developing country like Pakistan, it

would only work at the medical expert's level, leading to the usage of a cheap solution, creating non-standard, EMR.

Veracity Due to the heterogeneous nature of medical systems, EHRs suffer from a lack of universal quality. Universal quality is a made-up term, which is used to identify a golden set of features that an ideal EHR storage and processing system should have. In the real world, EHRs do usually conform to some (standard) schema, making them accurate, true and valid in a given context. However, as the (standard) schema is changed, the existing data becomes stale and often loses its usefulness as well. Additionally, the mere presence of schema would not enhance the quality of data. Additional enrichment information in the form of linked medical records and supplementary knowledge bases are necessary for achieving this aim. LinkedEHR has presented a good approach to partially resolve the data veracity problem, by identifying and building a common platform for primary and secondary data [83], leading to actionable insights into diagnosis, risk stratification and treatment [84]. Yet another key factor to consider here is the fact, that high volume does not always translate to veracity. While it is possible to dilute the gaps in data, when doing quantitative research, the same is not really possible in qualitative research [85]. One way of verifying the truthfulness or veracity of medical data is to measure the data quality in terms of its timeliness (e.g. When did it happen?), completeness (e.g. Did we capture/record everything?), uniqueness (e.g. Is this a duplicate entry?), validity (e.g. Does the data correspond to its schema?), consistency (e.g. Is there any conflicting data?), and accuracy (e.g. Was the medical data recorded accurately, mirroring the real world events?) [86, 87].

Value The main driving force behind the creation of semi-structured data is to ease the process of converting high volumes of diverse healthcare data, being produced at ever increasing velocity and of varying quality into information and knowledge. Due to its nature as an integrated healthcare record, the semi-structured data is able to provide value, to the patient, the medical experts, organizations, and the environment. The semi-structured data complements the benefits from traditional healthcare systems [88] by enriching each patient record with supplementary data from secondary sources and medical IoT.

2.4.2 Healthcare Interoperability

As defined by IEEE 610.12, interoperability is the ability with which, two or more participating information systems or components can not only exchange information but also use it [89]. Building on this basic definition, Health Level Seven International(HL7), a healthcare standard management body, divides interoperability into functional and semantic types; where the former relates to reliable exchange of information, while the later allows the receiver to interpret and use the information. Additionally, CEN ISO/IEEE 11073, is a multi-part standard, developed in collaboration with other standards development organization, that defines the communication standards, enabling real-time, efficient exchange of data produced by (plug-and-play supported) medical care devices [90]. HIMSS, provides a more comprehensive definition of healthcare interoperability by defining it as the ability to exchange data, at foundational (only relates to exchanging data, without the need to interpret it), structural (an intermediate level, that takes the schema of the data into account as well), and semantic (takes, schema and meaning of the information into account) levels, within and across organizational boundaries [91].

Ubiquitous healthcare can be formalized using these definitions. However, achieving interoperability, in the presence of voluminous, heterogeneous, low quality healthcare data, produced at different rates [72,73], is an uphill task. This is compounded due to the development of a plethora of messaging, terminological, decision support and other standards [1,81]. Besides the well-defined and developed standards, practical healthcare informatics also suffers due to the existence of non-formal standards, which are used to build specialized small-to-medium scaled systems. Healthcare organizations tend to move towards standards that are easy to use and cost effective [92]. While, this is usually not a problem when medical components have to be made interoperable within the organizational boundary, interoperability between different, often competing, healthcare organizations is a major challenge [93].

Data Interoperability, is a part of the general interoperability problem, which represents the set of policies and guidelines, and their application towards building systems and services that can help create, exchange and consume data while maintaining its contents, context and meaning. These tasks require the use of schema matching/mapping approaches, to map (transform) source data into a consumable form [94]. The main approaches to data interoperability can be categorized as stan-

dard based and mediation based approaches. Whereby the former, is focused towards creating and using agree-able standards, which all participating organizations must conform to, while the later, more autonomous approach, creates data translations from descriptions of the data in participating schemas [95]. Linked Data is a well-known example of standardization based data interoperability approach [96], while Semantic Information Layer (SIL) [97] is an ontology mediation approach for data interoperability among Enterprise Information Systems (EIS).

In healthcare, data interoperability can greatly enhance the financial and administrative aspects by reducing overhead and redundant costs, saving time at both the patient and physicians end, preventing operational waste, and allow policy makers to employ the best accountability and privacy services across the board [19].

Overall, healthcare Interoperability (when achieved), will additionally enable the healthcare organizations to increase the data and service delivery quality [7] and remove gaps between healthcare providers and patients [10].

In order to resolve the heterogeneity problem in healthcare, we have to look at the use cases, where an interoperability service can be utilized. In the case of Ubiquitous Health Platform, as shown in Fig. 6.1, input medical fragments can either be transformed from a source schema to a target schema, or it can be amalgamated into a comprehensive model for the patient's medical history. The former, challenge can be resolved using semantic matching algorithms while the later requires semantic amalgamation. These techniques are further discussed in the following sub sections.

Semantic Matching

While there can be many ways to cater for bridging the ever growing gap between heterogeneous medical systems and bringing them on the same connected platform, two primary strategies are standard based and mediation based semantic reconciliation [95]. Here, the former aims to develop a central standard, which all medical systems can comply with [77], while the later, uses mediating ontologies, which can semantically transform data from one format to another [97].

In coming up with a central standard, Clinical Information Modeling Initiative (CIMI) [16] has shown great promise, by integrating the best features of Health Level Seven Version 3 (HL7v3) [13] and openEHR. This endeavor is especially important, given the fact that both HL7v3 and

openEHR provide structurally distinct templates (and archetypes) for medical data representation and exchange [98]. In the same way Systematized Nomenclature of Medicine - Clinical Terms (SNOMED CT [14]) , is a terminological standard representing systematically codified clinical nomenclature, while and Logical Observation Identifiers Names and Codes (LOINC) [15] is a terminological standard for laboratory tests and other measurements. Until 2013, both of these standards had some overlapping, leading to problems in using them together. However, efforts are now underway to link LOINC and SNOMED CT, removing any overlapping, leading to healthcare interoperability at the terminological level. In terms of achieving some automation for this semantic reconciliation process, a lot of state-of-the-art ontology matching tools have been presented using the Ontology Alignment Evaluation Initiative) OAEI [99] platform. However, apart from few matching tools, most have limited extendibility, reusability, and expressive mapping representations, leading to their low adoption rates. Semantic reconciliation, using mediation based approaches, require the usage of similar ontology matching and transformation techniques, which can bridge the gap between heterogeneous systems. Over the years, several methods have been proposed and implemented for achieving the objective of interoperability. These methods, include but are not limited to, the use of standards, mediation via third parties, specification-based interaction, and mobile functionality [100]. Semantic Mediation Systems, represent a formal transformation process, which can provide coupling and cohesion between different data sources [95, 101], using Model-Driven Engineering [102].

A plethora of medical platforms have achieved some form of interoperability by mediating between healthcare standards, and extending the benefits of formalization and systematic definitions. One of the most prominent and active semantic transformation tools is the LinkEHR [103], which provides transformation between between HL7 Clinical Document Architecture(CDA) [104], openEHR, CEN/ISO 13606, CIMI reference model, and others [105]. LinkEHR, uses archetypes which contain definitions of clinical information models and a mapping specification generated by the knowledge engineer which is then used for converting legacy data into one of the supported standard types, finally producing a normalized XML file. This conversion is based on a common ontology which provides both syntactic and semantic relationships between the two participating schemas [106, 107]. The knowledge engineer, with ample knowledge on informatics can use

a purpose built UI for matching the schemas. Application of LinkEHR have also proven effective to achieve interoperability between CDSS and EHR, which correspond to different levels of abstraction in terms of patient information (usually CDSS is a more abstract representation than EHR) [108]. The LinkEHR platform doesn't provide native data storage services but can be integrated with other similar implementations(including other LinkEHR deployments) and can also act as a semantic transformation engine for other healthcare interoperability platforms.

Semantic Integration

Traditionally, healthcare solutions have focused on the use of well-structured storage for resolving interoperability. However, with a variety of medical platforms becoming widely available the interoperability problem now requires the use of ontologies and semantic maps which can identify and create relationships between various data elements from various sources [109]. Semantic Integration, provides a solution to the interoperability problem, by utilizing standardized models, in the form of Resource Description Framework (RDF) [110] and Web Ontology Language (OWL) [111]. Three main methodologies to achieve semantic integration are discussed as follows. Ontology-Based Data Access (OBDA) framework represents such a solution that is dependent on well-defined domain ontologies, which can map concepts from several data sources. The OBDA model consists of data elements and their semantic relationships build using a terminological service. When a user queries for some selected variables associated with the patient data, it is converted into SPARQL [112] which identifies the semantic relations between participating systems and creates native subqueires, which are executed in a federated manner. The results from these queries are finally integrated using unique identifiers from their data tuples.

The usefulness, of this framework to semantically integrate medical data for cancer patients is proved in [113]. The authors used a top-down approach to first construct an Ontology for Cancer Research Variables(OCRV), which contains the semantic relationships between the concepts in virtual RDF graph forms, from four different relational data sources. This ontology contains well-defined terminologies which are based on the National Cancer Institute(NCI) Thesaurus [114]. For converting SPARQL queries into native SQL queries the Ontop OWL API [115] is used which relies on the Ontop model, containing both semantic axioms and the data configuration necessary

for connecting with the data sources. The results from each data source is then integrated using the unique identifiers for all records, and presented to the client.

Some multi domain semantic integration strategies, have focused on the development and/or usage of Enterprise Service Bus(ESB), which provides a loosely-coupled, highly distributed, communication channel for software applications and modules in a service-oriented architecture (SOA). In general, several services can connect with this shared communication channel as a consumer or a producer. Each producer converts the messages into an internal format understood by all services, especially consumers. Using a publish/subscribe model, the services are able to communicate with each other using event driven paradigm. In healthcare IBM provided an early implementation of the ESB to create the IBM Healthcare Service Bus [116] which enables the integration of multiple services by using Web Services Description Language (WSDL) [117], Simple Object Access Protocol(SOAP) [118], and HL7 Standards. The service has now been upgraded [119] to become completely deployable on the cloud and to provide support for many healthcare standards suchs as HL7v2.X, Fast Healthcare Interoperability Resources (FHIR) [120], Digital Imaging and Communications in Medicine (DICOM), and others. It also now supports several types of message flows including eXtensible Stylesheet Language Transformations (XSLT), Extended Structured Query Language (ESQL), File Transfer Protocol (FTP), Java Message Services (JMS), and others. Health Service BUS(HSB) [121] is an implementation of the Mule ESB [122], which uses a native XML-database and XSLT to provide semantic translation services from HL7v3 to HL7v2 [123] and openEHR. The patient EHRs are stored using OpenEHR database, while HL7v2 and HL7v3 are used for sharing messages belonging to a particular patient. The HSB uses SNOMED CT for providing terminological services, which are also embedded into the ESB as XML messages and used with a custom ontology mapping tool called OWLmt to provide semantic interoperability between patient records. In [116] an event-based HSB based on the JBossESB is presented which converts heterogeneous data into RDF quads, before utilizing the Health and Lifelogging Data (HLD) Ontology for building a semantically linked graph of health and lifelog data. The authors have used LOINC as the terminological handler, which is used to provide semantically annotated versions of input sensory data from wearable devices, before creating the RDF quads and applying semantic integration using the HLD ontology. Internally, the bus is able to provide point-to-point

communication between any two services, and a publish/subscribe broadcast model using JMS queues. The overall platform can be used to push notifications to the users, using event-driven paradigm and can also provide query services for executing SPARQL queries.

Yet another interesting initiative is the Yosemite Project [17], which aims to bridge the gap between healthcare standards and the data. The main driving force behind this initiative is the conversion of messaging standards like HL7v2 and FHIR into RDF graph for semantic representation. It is also concerned with resolving the ambiguity in the human language by using Natural Language Processing technologies for processing unstructured medical data(such as Clinical Notes, Clinical Practice Guidelines, and others). Their methodology consists of two related process, standardize the healthcare standards and using crowdsourcing for translations. Here the former task has been undertaken to find and create semantic links between 30 most used vocabularies amongst over a 100 listed by Unified Medical Language System(UMLS) [124].

Using a custom tool, iCat, which currently only support International Classification of Diseases (ICD)-11 [125], the yosemite group provides an easy to use interface to the medical experts. Over 45,000 concepts with 17,000 links to external terminologies have been defined by the medical experts, which are converted into RDF form for creating computable data resources. The latter task of translation using crowd sourcing resolves the problem of standard complexity, evolution of technologies and methodologies in computing and healthcare, and finally change in the standards themselves. This translation process is an extension of the inference process which can identify implicit relations, RDF assertions and localization between languages to enrich the existing semantic maps. This semantic integration is language/tool agnostic and can be used with any other platform.

While many paradigms have been introduced to resolve the semantic matching and integration problem, it is clear that the difficulty in creating ontologies and semantic bridges between various standards and terminologies is greatly hampering any functional interoperability solution. Additionally, the initiatives for standardizing the standards are still about a decade away from becoming implementable. Meanwhile, the healthcare data is growing beyond the management abilities of traditional data curation engines. Moreover, the top-down approach necessitates the use of medical experts for initially creating a rule base and/or ontology, which is not always possible. It is

therefore necessary that a novel methodology is used to archive the existing medical data and keep it available to create and test semantic integration methodologies. Additionally, due to the various methodologies involved in this semantic reconciliation process, it is important to store this data, while maintaining most of its original schema. Conversion to RDF quads, XML, relational or other methodologies can lose the original schematic information.

Chapter 3

Proposed Methodology

Digital healthcare interventions have led to an explosive growth in the quantity and quality of medical data. HMIS are able to capture structured data from clinical encounters, while some of the unstructured data in the form of clinical conversations goes to waste. Additionally, the medical experts, often have to bypass the restrictions of traditional interfaces and record medical reports in free text. In these circumstances, ubiquitous healthcare becomes a distant reality, owning the gaps between heterogeneous data. In order to bridge this gap, this dissertation presents novel solutions for converting unstructured text into semi-structured data, schema alignment, and semantic reconciliation-on-read. These novel solutions, when working in tandem, are able to provide Standard-Agnostic Data Interoperability. The underlying technique, used to achieve these goals is the semantic similarity between text sequences. However, before calculating the similarity it is important to extract relevant artifacts from existing text, and generate sequences from token strings. The methodology for producing the former is Sequence Contraction, while the methodology for the later is Sequence Expansion. The integration of these three techniques is shown in Figure 3.1.

In the Sequence Contraction phase, unstructured text is used to identify the relevant medical data from clinical conversations, by first pre-processing it to produce text sequences. These sequences, correspond to the question-answers, statements, and pharasal text, which act as an atomic unit for further processing. Next, we classify the produced sequences, by using a set of known medically aligned sequences. Beyond simple classification, this process also identifies the medical attribute which may exist in the classified instance. Next using an extraction methodology, corresponding to the identified attribute and identified by the classification process, is used to extract the value pertaining to the attribute. For syntactic extraction, regular expressions are used, however these can be replaced by any ML based methodology, as well. The semantic extraction

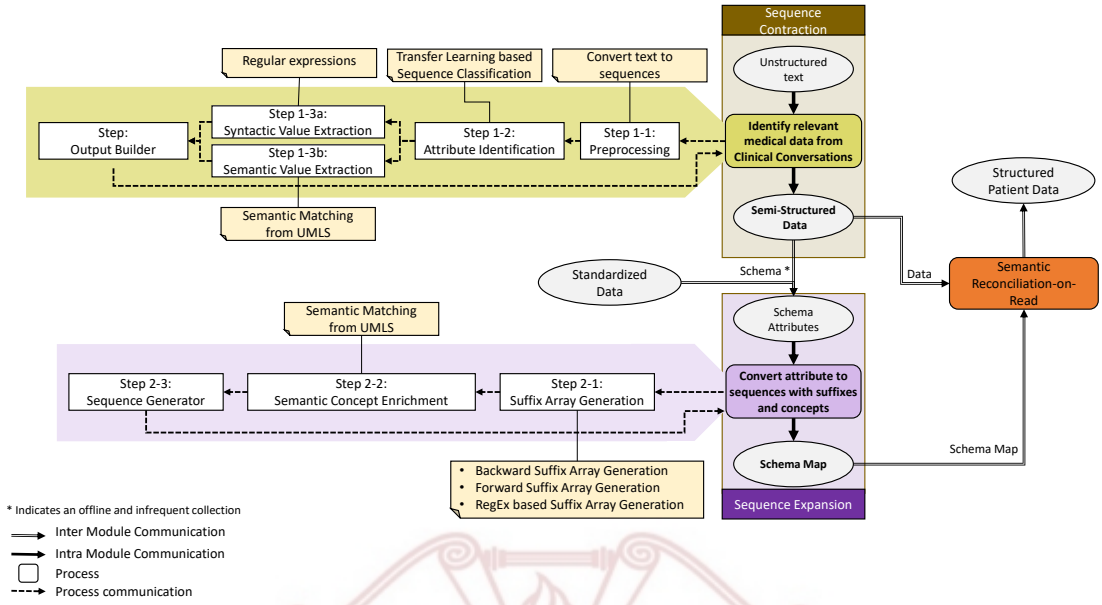

Figure 3.1: The proposed methodology workflow

uses UMLS concept dictionary to identify unigram and bigram tokens semantically similar to the identified attribute. Finally, an amalgamation of the identified attributes and their corresponding values, are used to generate the semi-structured data necessary for archiving and schema alignment. This process, is shown in Figure 3.2.

Sequence Expansion, on the other hand, is used to convert token strings into text sequences. The process is shown in Figure 3.3. Here the token strings can contain a combination of words with out separation between them. Typical database design follows this paradigm, where spaces are not allowed in the names of the attributes. In order to identify the terms within these tokens, first we apply the suffix array generation methodology, using forward and backward suffix generation, as well as regular expression based suffix generation. The resultant set of suffixes are then checked in the UMLS concept dictionary to filter out the non-medical words. The remaining suffixes are then sent as a query to the UMLS again, this time picking out all the concepts associated with the term. A sorted combination of all sequences and their concepts, thus creates a target sequence. In order to test the correctness of these sequences, attribute from various schema were compared and marked by human annotators, providing a true set. Then using various pre-trained

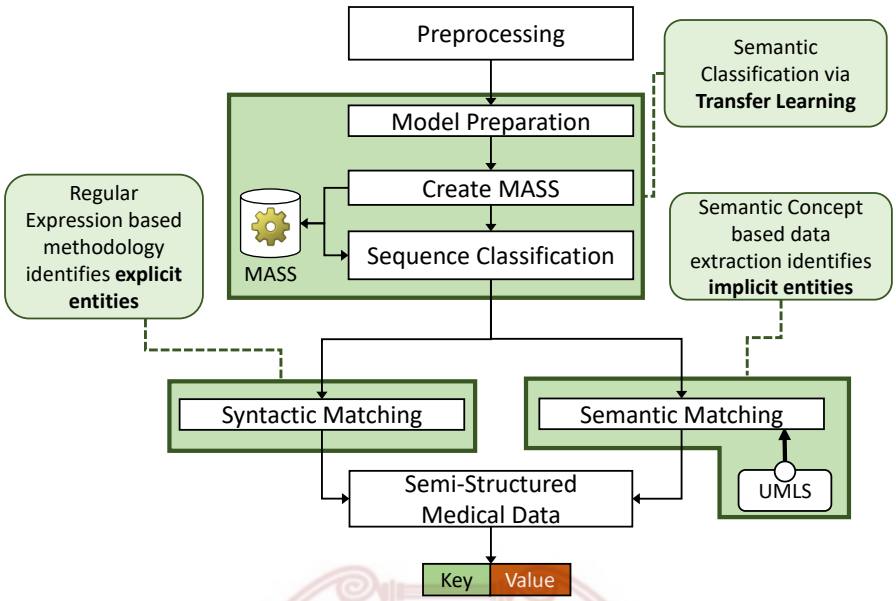

Figure 3.2: The sequence contraction methodology

models for semantic similarity the same attributes were compared. The resultant set is then compared to the true set. The final outcome of this process is the agreement between the computed method and the annotators using various correlation metrics. Due to the unsupervised nature of this similarity matching technique, it is well suited to enable automatic Schema Alignment, which provides many-to-many Schema-Maps between all disjoint attributes of a pair of medical schema.

The Semantic Reconciliation-on-Read solutions provides a framework for archiving and processing medical data to enable Data Interoperability. This is achieved through the use of two services, data archiving and data processing. Here, the former utilizes a Big Data storage platform to provide persistence for semi-structured form of the medical data. This semi-structured data is obtained from unstructured text, using the Sequence Contraction algorithm, while for structured data, naive serialization is utilized. Additionally, the Schema-Map produced as a result of the Sequence Expansion process are also stored in the archive. The archive provides version control which is necessary to enable evolution of the data and the Schema-Map, ensuring transparency and consistency of the Semantic Reconciliation process. An overview of this process is shown in Figure 3.4. Data processing is initiated on request by a medical expert, which triggers a search of the necessary patient data in the archive. Additionally, for each data element a corresponding

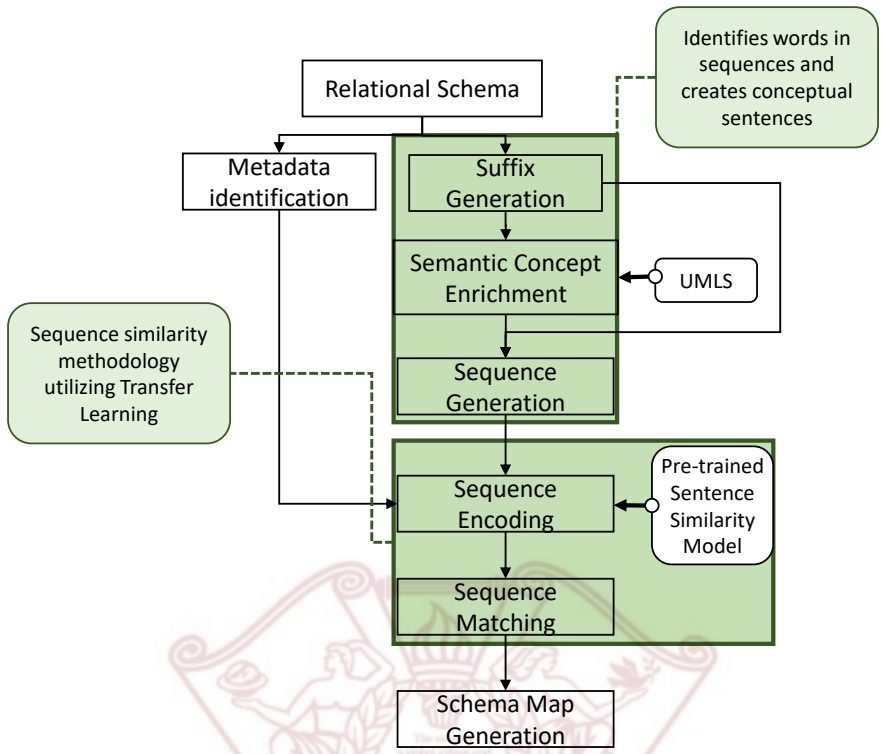

Figure 3.3: An abstract view of the sequence expansion method

map for its source data schema and the target schema is picked up. Then using the Schema Map, each instance is converted into the structured form, consumable by the requesting medical system. This is finally returned to the medical expert. An implementation of the proposed solution on a hadoop datastore was created, with hive providing an interface to query the data in a fast manner. This platform was tested in terms of its timeliness to store and query medical data, across various iterations and under heavy data load.

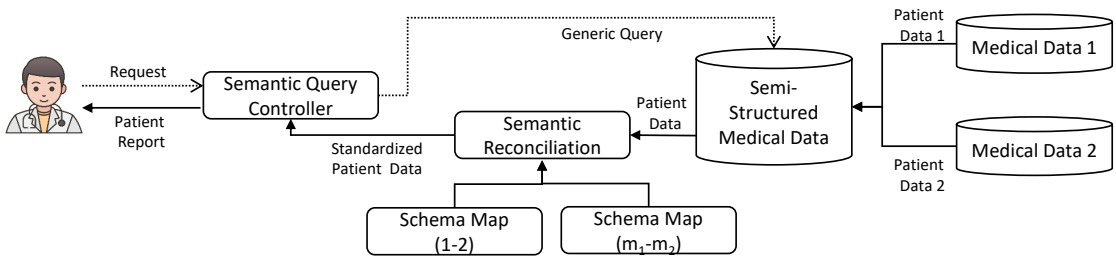

Figure 3.4: Semantic Reconciliation using the proposed solution

Chapter 4

Sequence Contraction

An important requirement of any Healthcare Information Management Systems (HIMS) is the ability to create, store, and share Electronic Medical Record (EMR) for enabling the long-term management of patients and diseases [126]. However, due to the challenges faced by healthcare services in the developing world, the encounters are partially recorded via offline methods such as registers and printed forms. The EMRs are collected via structured data acquisition and processing engines, which, ensure a controlled input, well-suited for the Scheme-on-write methodology. This methodology is well suited for quickly querying structured data (such as relational databases or constant-sized data blocks on disc). Schema-on-read provides an alternate solution, where the structure of the collected data is of little concern during the acquisition and storage time, however, when the data is queried, a temporary schema is utilized to create a view. Due to the unpredictable nature of the acquired data in an unstructured format, the Schema-on-read methodology is useful for descriptive and predictive analysis. A good middle ground between these two methodologies relies on the use of semi-structured data, whereby validation is performed on some critical structures and elements of the data, meanwhile, further processing is utilized to extract the value, during read [[127]].

Due to recent advances in technologies, particularly, Machine Learning, Optical Character Recognition (OCR) and Natural Language Processing (NLP) [[128]], it is now possible to develop a bridge between the existing HIMS and the challenging requirements of healthcare service delivery in the developing world. Physical interactions in the real world can now be digitized and converted into information and knowledge, which can remove the redundancies, associated with the traditional data collection ways of the HIMS [[129]].

4.1 Research Objectives

The main objective of this research work is to convert unstructured data from natural language used in clinical conversations into a structured format, which can be verified in a timely manner.

Theoretically, our proposed methodology is similar to other methodologies, such as the one presented by [130], which focuses on extraction of metadata, classification and clustering of data, and mapping the data onto the target schema. However, in creating a practical solution, we had to face many challenges, which required novel interventions and assumptions to simplify the problem space. In particular, we are interested in resolving three challenges, which are defined as follows.

- *Challenge 1*; The first challenge is to identify the key domain elements which can relate to an attribtue's name. This name can either be a textual identifier from the consuming schema (such as a database or web service) or a generic identifier (such as from a domain adapted concept dictionary), which can be mapped onto the storage schema.
- *Challenge 2*; We should also identify the portion of the input text which corresponds to an identified attribute's name and holds its value.
- *Challenge 3*; Lastly these identifications should take into account, the time required to verify the contents. Essentially, while it would be possible to classify each word and group of words as a valid attribute's name and its value, the resultant dataset would be too large, contain many incorrect results and would greatly increase the verification time for the physician.

In order to fulfill our objective, focusing in particular on the above mentioned challenges, we have developed a sequential pipeline, which applies semantic matching and transformative functions on a specialized dataset, to transform unstructured text into preset schema specific key-value pairs. The novelty of this applied solution lies in developing an end-to-end methodology for solving a real world problem, while utilizing and re-purposing various state-of-the-art tools and technologies. While many research initiatives have proposed novel methodologies for extracting relevant information and knowledge from unstructured text, to the best of our knowledge, and within the bounds of the real world challenge, our methodology is a unique solution.

The main contribution presented in this research work is as follows.

- *Data Acquisition*; Firstly, our primary source of data is the interaction between physicians and patients or their guardians. For this we recorded short conversations between physicians and patient/guardians from two hospitals in Pakistan, specializing in pediatric care. This included District Headquarters Hospital, Kotli, Azad Jammu and Kashmir, Pakistan (DHQ-Kotli) and Care+ Medical Center, Islamabad, Pakistan (Care+ MC-Islamabad). Since these conversations were held in the national language of Pakistan(Urdu), the audio files were transcribed and the contents were translated into English, using human experts.

- *Data Pre-Processing*; Secondly, we converted the translated conversations into sequences, which represent a unit of conversation, in the form of a question and its answer, or a statement.

- *Model Development*; Thirdly, using transfer learning methodology and using real data we have created the Medically Aligned Sequence Set (MASS), which contains only 190 instances. Each instance holds enough data to classify unseen sequences, identify an appropriate attribute's name, and an extraction methodology to obtain its corresponding value.

Once an appropriate attribute's name and value have been identified we can then transform the structured contents of each conversation into a relational schema, designed for an HMIS. The particular data interoperability methodology, for matching the attribute names is based on our previous work presented in [131]. The final key-value, compliant with a consuming platform (such as a database system or a form) can then be presented to a human expert for validation, before it is stored.

4.2 Methodology

In order to convert the unstructured input text into structured schema elements, we have developed a pipeline, comprising of various transformation, matching, and filter processes. Throughout this manuscript, we have used many terms and notations to simplify the explanation. Some of the most important terms are briefly explained here.

- **Sentence**;

  - (within manuscript write-up) Based on the definition by merriam-webster[1], a sentence is a collection of one or more words, forming a syntactic unit, which can be used to ask a question, provide an answer, and present an assertion or an instruction. In written form, a sentence should end with punctuation (such as question mark, period, semi-colon) or an indicator (such as "Doctor:", "Patient's father:", "Patient's mother:") used while transcribing the conversations.
  - (in the context of sentence-similarity) used to describe a famous NLP task of determining the similarity between two texts. While our task is similar to sentence similarity, in order to avoid confusion, we shall call it sequence-similarity, where required.

- **Sequence**; A sequence is a collection of one or more sentences, with at least two words (to support the lookup of key-value pairs). In particular, a sequence can contain, a question and its answer, a question followed by another question, an assertion or instruction, or phrases from the sentence, split on "and" or ","(comma).

- **Medical Sequences**; A sequence containing at least one key-value pair, where the key is a medical concept such as "Finding", "Disease" or others. A probable medical sequence contains computed key-value pairs, while a valid medical sequence is validated by a human expert.

A brief overview of the notations used in this text are presented in Table 4.1. As shown in Figure 4.1, conversations between physicians and patients, are first converted into a textual form by human intervention, which is then pre-processed to build manageable sequences (S). These sequences are then used to fine-tune the DistilBERT base (uncased) model[2], creation of a MASS, threshold selection, and test set creation. MASS is composed of enriched sequences (E), which contain an embedding vector obtained by encoding valid medical sequences, a label representing the attributes in the sequence, and an extraction methodology, to extract the value corresponding to the label. By classifying Test Sequences (T), based on their similarity with the set E and the ability to extract value from the sequence, we can produce a smaller set of probable medical sequences

[1]https://www.merriam-webster.com/dictionary/sentence
[2]https://huggingface.co/distilbert-base-uncased

Table 4.1: Notations used in the manuscript

Term	Definition
S	Set of Sequences
E	Set of Enriched Sequences
H	Set of Sequences used for threshold selection
T	Set of Test Sequences
P	Set of Probable Medical Sequences
A	Set of Key-Value pairs extracted from Sequences
M	The medical schema used to identify the target attribute names
$\vec{V}$	The embedding vector produced by encoding a sequence.
l	A label indicating the textual key and the expected value
x	A methodology to extract values from a sequence
α	Threshold used for sequence classification
β	Threshold used for filtering similar attributes

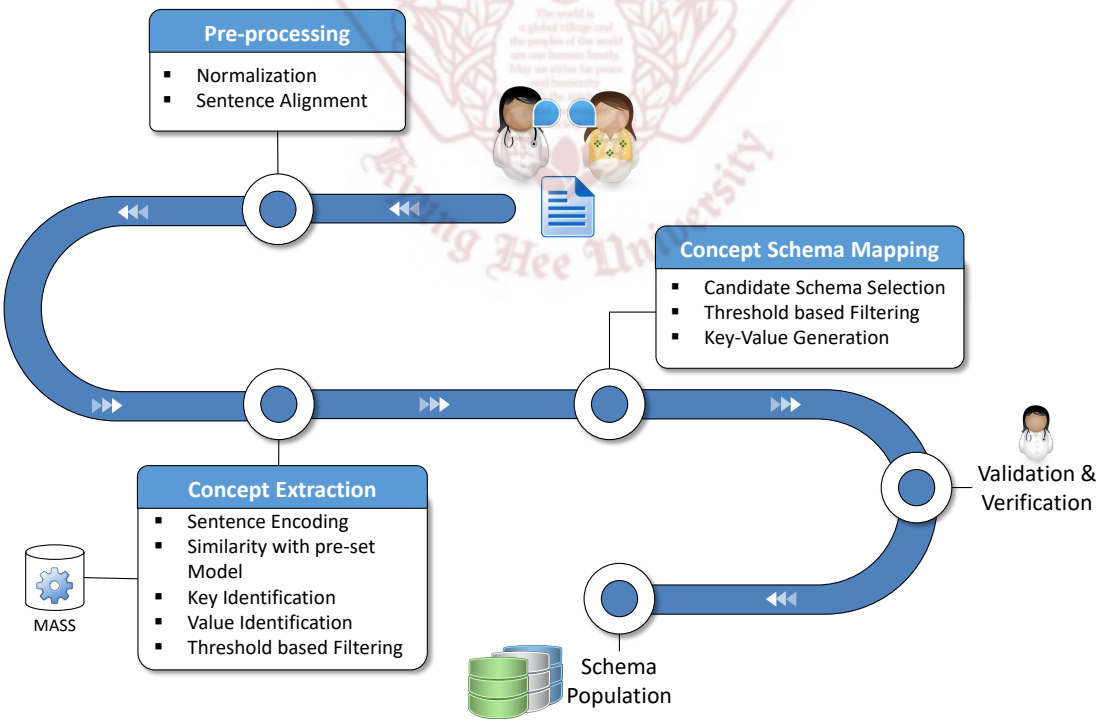

Figure 4.1: An overview of the proposed methodology

(P), which are labeled with the attribute name and extracted value, predicted from a member of MASS and contained within the input sequence. Using the identified keys and values from P, we then produce the set A, In this way, not only can we identify more than one key-value pair present in each instance but we also discard the extra tokens in the input text. Each instance of A is then used to identify similar elements from the participating medical schema (M). We then obtain an alternate representation of A, where each key is replaced by one or many attribute names from M while keeping the value part of this pair intact. This new set of elements in A is verified by the medical expert before it can be split into schema-wise sets of key-value pairs and passed on to a data store for storage. Thus some sentences from the conversation can end up in key-value pair form, which is syntactically closer to the database schema, now, than it originally was. We shall now delve into the details of each of these steps.

4.2.1 Pre-Processing

In the first phase, conversations between the physicians and patients or guardians(for young patients) are converted into a set of sentences that can contain the keys and values of the final structured form. Here, we made two assumptions which are described as follows.

- *Assumption 1*; This assumption, dealt with compound sentences. Natural conversations between the participants included many compound sentences, where multiple questions were following each other, such as “How old is she? And what happened to her?”. Some answers were completely skipped, and in some cases, isolated answers were contained in sentences, such as the case where a guardian said, “t**** age 3 months and she has some pain dont know where”. Finally, some sequences contained multiple keys and values, such as “whats his name and age? U**** hes 3 month sold”. In these cases even if we are able to partially predict the answers as correct, they are considered completely correct.

- *Assumption 2*; The conversations either contain medical sequences of the form Question-Answer or Statements (Instructional or Assertive), containing both key and its corresponding value. These two types are explained below.

  – Firstly, the question and answer type sequences are made up of two consecutive sen-

tences, where the key lies in the question part, while the value lies in the answer part. As an example, consider one of the most common sentences, “What is his name?”; here the key is “name”, while its value is found in the statement by the patient/-guardian.

– Secondly, statements given by physicians and patients/guardians may also contain both keys and values, such as the sequence, “Fever is a little bit”, where the key can be “Finding” with its value of “fever”.

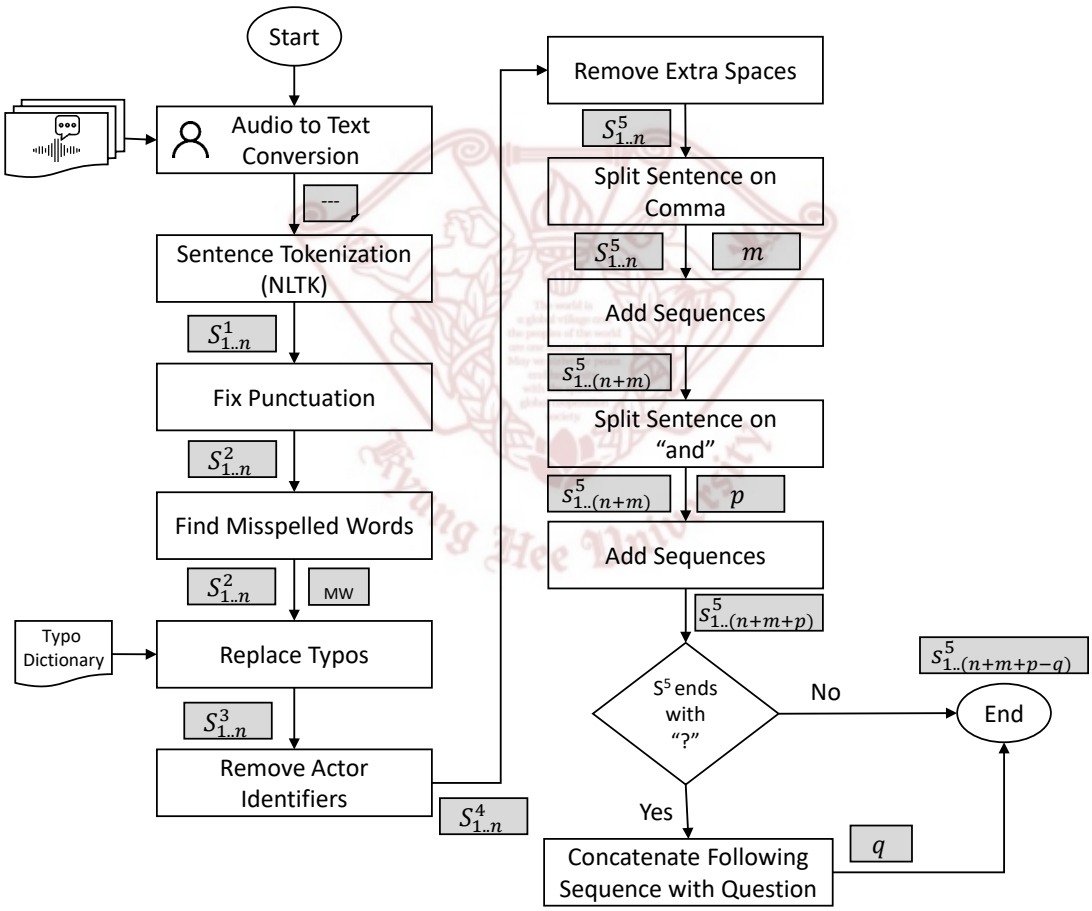

Figure 4.2: Preprocessing methodology

The detailed flow chart for pre-processing is shown in Figure 4.2. The eventual output of this phase is the set of sequences S. As shown in Figure 4.2, the various transformations applied to

each sentence of this set are represented by the superscript symbol, while the subscript represents the number of sequences in S.

The process starts, by transcribing the conversations into English text by human experts (further details of this process are described in 7.2.1). The transcribed text is first split into sentences using the Natural Language Toolkit (NLTK)[3] library in python's "sent_tokenize" function. The output of this subprocess is $S^1_{1..n}$, where the first transformation has been applied, producing n sentences. Next, we fixed the punctuation to fix some human errors in placing the punctuation marks (such as adding a space before punctuation, no space after punctuation, various quote type usage, and others) to align the syntax of the sentences. Here the second transformation is applied, while the number of sentences remains the same, producing $S^2_{1..n}$. Typographical errors (Typos) and incorrect spellings by the transcribers were searched by using a spell checker (based on a blog post by Peter Norvig [4]). This process produced an additional set of misspelled words MW. Using a custom typo dictionary, we fixed some of the common errors in the set MW, eventually producing $S^3_{1..n}$.

As a part of the transcription process, each sequence was marked by the actor, speaking it. These identifiers were then removed to apply the fourth transformation, producing $S^4_{1..n}$. We then removed multiple spaces in each sequence, to produce $S^5_{1..n}$.

Acting on the *Assumption 1* of fixing some problems with compound sentences, we split the sentences on ",", producing additional sub-sequences. If the length of the sub-sequence was greater than one, we added it to the set of sequences. This would add an additional m sentences into the set of sequences, producing, $S^5_{1..(n+m)}$. We then repeated the same process for the keyword, "and", producing $S^5_{1..(n+m+p)}$.

To resolve the *Assumption 2*, we selected the sequences ending with "?", and concatenated the next sequence with this sequence. This additional sequence does not have an independent existence. Thus, the q answers to the sequences ending with "?", are removed from the set, producing $S^5_{1..(n+m+p-q)}$, which we shall simply call S, henceforth.

[3] https://www.nltk.org/
[4] https://norvig.com/spell-correct.html

4.2.2 Concept Extraction

The sequences identified from the conversations are used to fine-tune the DistilBERT base uncased model, preparing MASS, threshold selection, and to filter the test sequences containing medical concepts and their values. The methodology, for this phase, is shown in Figure 4.3. First, a portion of the sequences in S are separated to be used for developing the model MASS, which contains the medically aligned sequences, predetermined as interesting by human annotators.

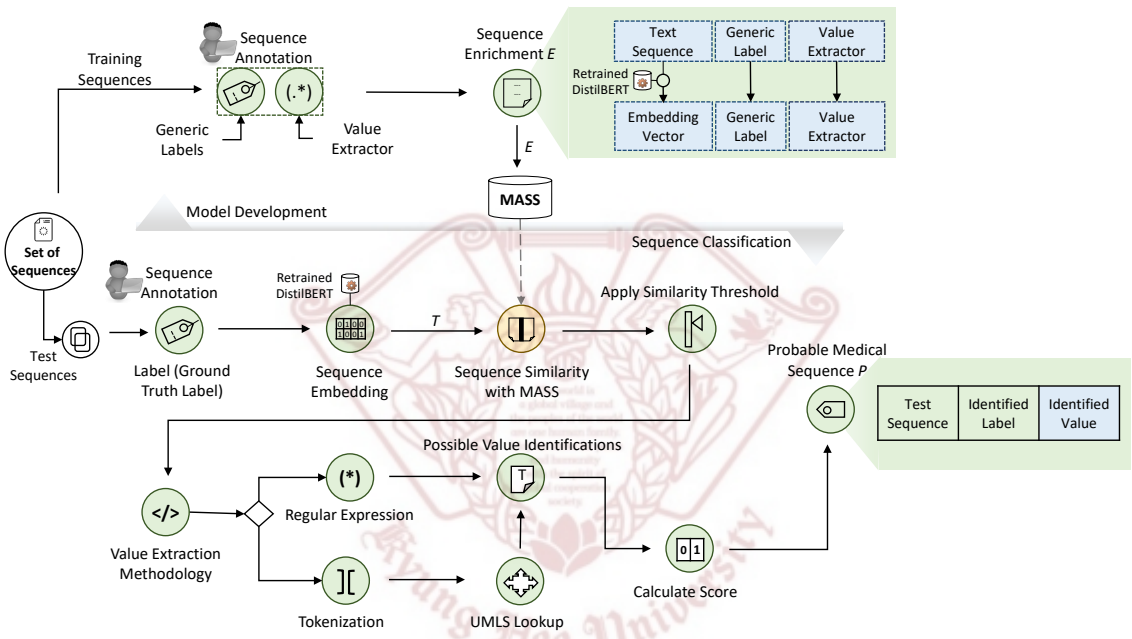

Figure 4.3: Workflow for classifying the sequences as Medically aligned or not

The annotators identified these sequences d by identifying the medical concepts and their values within them. Then each sequence was converted into a pattern with three items. Firstly it includes, a generic form of the sequence with special tags ("[CLS]", "[SEP]", and "[MASK]") which is used by our Fine-Tuned DistilBERT model (further explained in Section 7.1.1) to encode and produce the embedding vector $\vec{V}$. Secondly, the medical concepts associated with words and phrases within the sequence are attached to the pattern as label l. These labels are based on the semantic types, included in UMLS (such as Diagnostic Procedure, Disease or Syndrome, Finding, Sign or Symptoms, and others) and other generic tags (such as name, age, date of birth, and others). Thirdly, a value extraction method x is attached to this pattern, which can be used to

extract the value from a target, unseen, data instance through the use of UMLS lookup or the application of a regular expression.

This labeled data is then processed and converted into instances for MASS, which is of the form shown in Equation 4.1. This includes the embedding vector, produced by encoding the sequence ($encode(sequence) \rightarrow \vec{V}^E$), the label l, and the extraction methodology x. The actual text of the sequence d is not used again.

$$e = \left\langle \vec{V}^E, l, x \right\rangle | e \in E \tag{4.1}$$

Then using a threshold selection process, further explained in Section 4.2.3, the optimal threshold α for semantic similarity classification of the two sequences was obtained. With MASS and a semantic similarity threshold α, we then process the unseen test dataset, as shown by the sequence classification pathway in Figure 4.3.

The test dataset is of two types, labeled and unlabeled. The Labeled Test dataset $T_{labeled}$ is the bigger of the two and is first annotated by human experts, to attach an expected final label l with the sequence. This is useful to evaluate the sequence classification methodology without expert intervention. On the other hand, the unLabeled Test dataset $T_{unlabeled}$ is verified by the expert to evaluate the accuracy of attribute key-value extraction from the sequences.

Both types include the sequence text (d) which is encoded using the ine-tuned DistilBERT model to produce the embedding vector ($encode(d) \rightarrow \vec{V}^T$).

For classification, the embedding vector from set t ($\vec{V^T}$) is compared with all the embedding vectors from MASS ($\vec{V^E}$). This comparison is performed using cosine similarity, as shown in Equation 4.2, which assigns a score between 0 and 1 to the pair.

$$sim_t = \frac{\vec{V}^T \cdot \vec{V}^E}{\sqrt{\vec{V}^T \cdot \vec{V}^T}\sqrt{\vec{V}^E \cdot \vec{V}^E}} \tag{4.2}$$

For all pairs of $e \in E$ from MASS and $t \in T$ from the test set, with a similarity score above α, we then apply the value extraction methodology, predicted from the training set, against each

test sequence. This methodology can be one of "UMLS Lookup" or "Regular Expression". The "Regular Expression" methodology is used to identify the value for concepts such as "name", "age", "duration", and "frequency" from the test sequence. These patterns are manually built using the sequences found in MASS, during the annotation process. The intuition behind using these patterns is to allow value extraction from sequences that do not contain medical concepts, which act as a value to the corresponding key. On the other hand, many sequences contain medical concepts, such as "fever", "cough", "flu" and others, which can be found using the UMLS API. The textual part of the test sequence d is split into unigram and bigram tokens, which are sent to the approximate search API of UMLS. The API returns a list of semantic concepts which may be associated with the search term. By caching both positive results and negative results, we can avoid recurring queries, which is essential to bypass the search limits enforced by UMLS. By matching the semantic type returned by UMLS and the predicted attribute name from the associated label in e, we can determine the labels associated with the test sequence.

Finally, we calculate the match score, based on the sentence similarity score sim_t and the ability to extract a value (irrespective of the value being correct or not) from the text. If the value can be extracted then a score of 1 is assigned to the pair, otherwise, it is set as 0. Instances with a score less than $1 + \alpha$ are filtered out, leaving behind the set P, which represents the Probable Medical Sequences. The structure of instances in P is shown in Equation 4.3. Here d is the text of the test sequence, l^E is the predicted label and η is the extracted value obtained from the predicted extractor function x^E.

$$p = \langle d, l^E, \eta \rangle \, | p \in P \wedge \eta \leftarrow x^E \, (d) \tag{4.3}$$

P is then used to extract concepts and produce the key-value pairs, which are used in the next steps.

4.2.3 Threshold Selection for Sequence Classification

For threshold selection, each instance contains the sequence text d, its embedding vector $\vec{V}$, and the associated label l. While the keys for this label are the same as in MASS, they also additionally

contain a correct value, for each key. This set H is represented as shown in Equation 4.4.

$$h = \left\langle d, \vec{V}, l \right\rangle | h \in H \wedge encode(d) \rightarrow \vec{V} \tag{4.4}$$

In the threshold selection process, we processed each instance in H by calculating the cosine similarity (shown in Equation 4.5) between its "embedding vectors" ($\vec{V}^H$) and the "embedding vector" ($\vec{V}^E$) of all instances in E. We then dropped all matches with similarity scores under 0.1, to reduce the number of comparisons in the next stage. In this way, we obtain a pair (ρ) of enriched sequences and their similarity score, represented in Equation 4.6.

$$sim = \frac{\vec{V}^H \cdot \vec{V}^E}{\sqrt{\vec{V}^H \cdot \vec{V}^H}\sqrt{\vec{V}^E \cdot \vec{V}^E}} \tag{4.5}$$

$$\rho = \langle e_i, h_j, sim \rangle | e_i \in E \wedge h_j \in H \wedge 0.1 \leq sim \leq 1.0 \tag{4.6}$$

We then compare the labels of each pair, to validate the similarity ρ in terms of the keys and values obtained from e_i and h_j. The total function, representing the computed match between the keys and values of the labels from the threshold set and its corresponding match with the MASS instance is shown in Equation 4.7. This process assigns one of three values to ρ, including "0", " ", and "1". If the two keys from any of the labels in the pair ρ are not equal, a value of 0 is assigned to it. On the other hand, if the two labels are equal, but the value annotated with the threshold selection set ($l^H.value$) and the value extracted from the application of regular expression or through the use of UMLS, as identified by the corresponding instance from MASS (x^E) on the text sequence from threshold selection set (d^H) are not equal, "$\sim$" is assigned to ρ. Otherwise, if the labels and the value extracted match the annotated value, "1" is assigned to ρ. The "$\sim$" matches were then manually verified and updated to either "0" or "1".

For all ρ, if $\chi(\rho)$ is zero, this indicates that while there is some cosine similarity (> 0.0)

between the sequence in MASS and in the threshold selection set, their label keys and the expected values do not match with what can be achieved by the current matched instance. The value for $\chi(\rho)$ defines the computed actual class label("0", "∼","1"), which is converted into verified actual class label by expert intervention ("0" or "1"). This final value is used as the actual class score while the semantic similarity score provided by the fine-tuned DistilBert model, is used to calculate the predicted class label.

$$\chi(\rho) = \left\{ \begin{array}{l} 1 \;\; if \;\; \left(l_i^E \cdot key = l_i^H \cdot key \;\wedge\; x_i^E(d^H) \in l_i^E \cdot key\right) \\ \sim \;\; if \;\; \left(l_i^E \cdot key = l_i^H \cdot key\right) \\ 0 \;\; otherwise \end{array} \right\} \tag{4.7}$$

In order to define the predicted class label as similar or dissimilar, we move a threshold iterator from 0.0 to 1.0, with a step of 0.01. At each iteration, if the value of sim inside ρ is below the threshold iterator, the predicted class is assigned as dissimilar, and if it is equal to or above the threshold iterator the predicted class becomes similar.

Thus for 100 iterations, we have one set of actual class labels and 100 sets of predicted class labels based on the value of H. At each step, we calculate the area under ROC (AuROC) which provides a numeric value representing the ratio between the True Positive rate and False Positive rate. The maximum value of AuROC across all iterations then provides the semantic similarity threshold (α) between the two expressions, whereby the pair is considered, actually similar.

4.2.4 Concept to Schema Mapping

Each instance of P, produced in the previous phase, can produce zero or more key-value pairs which, become a part of the set A. This set is a structured representation of the conversations between physicians and patients/guardians. However, this structured representation is very different from the database or API structure, making it difficult to connect the methodology up to the previous step, with any real application. Additionally, the schema underlying the structured data in A is a naive representation, which can capture only some parts of the recorded conversations, however, by increasing the labels used for annotation and restructuring the conversations to follow

some concise protocol, this problem can be resolved.

In our prior work, we have discussed the issues underlying healthcare data interoperability in [132] and introduced our novel semantic reconciliation methodology to map heterogeneous data schemas using BERT-based sequence encoding and semantic similarity measurement in [131]. This process is further elaborated in Chapter 5.

4.2.5 Expert Verification

The set A and its schema mapping is then used to create a transformation from $A \rightarrow m_i | m_i \in M$. This transformation is then presented to the medical expert to verify the contents before they can be passed on to a database for storage.

Chapter 5

Sequence Expansion

Data and Information modeling in the healthcare domain have witnessed significant improvements in the last decade owing to advances in the development of state-of-the-art Information and Communication Technologies (ICT) and formalization of storage and messaging standards. Subsequently, the scope of Healthcare Management Information Systems (HMIS), medical ontologies, and Clinical Decision Support Systems (CDSS) has broadened, beyond the operational capabilities of traditional rule based systems. One of the major reasons behind this limitation is due to the numerous heterogeneities in healthcare at data, knowledge, and process level. Thus, healthcare interoperability which aims to provide a solution to this problem, can be compartmentalized into data interoperability, process interoperability, and knowledge interoperability.

Data interoperability resolves the heterogeneity between data artificats to enable seamless and interpretable communcation among source and target organizations, while preserving the data's original intention during storage, communication, and usage (as defined by IEEE 610.12 [18], Health Level Seven International (HL7), and Healthcare Information and Management Systems Society HIMSS [133]). On the other hand, process interoperability regulates the communication among organizational processes to provide compatability between process artifacts within and seamless transformations across different organizations [8]. Lastly, knowledge interoperability provides a sharing mechanism for reusing interpretable medical knowledge, acquired through expert intervention and other mechanisms, across decision support systems [134].

In more tangible terms, healthcare interoperability at data, process, and knowledge level can be exemplified within the healthcare constraints experienced due to the emergence of Covid 19. The operational capabilities of the current healthcare service delivery infrastructure has gone under tremendous stress due to Covid 19. World over, large primary healthcare units have managed

to create separate units for managing patients, suffering from extreme cases of the novel coronavirus. For secondary and tertiary care units, government involvement has become necessary to filter coronavirus patients and adhering to a national pandemic response policy. These complex circumstances have enhanced the need for sharing patient data and state-of-the-art medical knowledge in real-time, to provide the medical experts with a tool to make accurate and timely decisions. Data interoperability can enable the front line medical workers to fetch, understand, and use patient data, especially comorbidities, across organizational and physical boundaries, without suffering from societal taboos that may prevent the patient from sharing their complete and accurate medical histories. Knowledge interoperability can improve the knowledge acquisition and sharing protocols to provide the medical experts such as epidieomologists and vaccinologist, with latest information on affected population trends, disease diagnosis, treatment, and followup procedures, and interpretable decisions leading to positive or negative outcomes. Process interoperability can help reduce and in some cases remove the operational redundancies between health centers. In this way, successive healthcare treatments can take benefit from earlier diagnosis, treatment, and followup procedures, thereby reducing the stress on healthcare experts and systems.

Healthcare Standards such as HL7 - Fast Healthcare Interoperability Resources (FHIR), and openEHR provide the foundations for storing and communicating medical data, through the use of well defined protocols. While Systematized Nomenclature of Medicine—Clinical Terms (Snomed-CT) [135] and Logical Observation Identifiers Names and Codes (LOINC) [15] provide a standard definition for clinical terminologies and laboratory tests, respectively. Similarly Medical Logic Module (MLM) provides a standardized way for expressing medical knowledge. Variety in these and many other healthcare standards necessitates the creation of bridging standards that can resolve the heterogeneity between the medical standards [136]. Substantial effort has gone into this endeavor with the Clinical Information Modeling Initiative (openCIMI) [16] taking the lead in bridging the gap between HL7v3 and openEHR. Similarly, SNOMED CT and LOINC are working to resolve the redundancies between the two terminological standards since 2013. This healthcare interoperability solution follows a formal, albeit long process, which is greatly dependent on the human factor. However, the current healthcare scenario requires a quick solution to create a scaffolding of an interoperable bridge between various healthcare providers. It is also

important to ensure that this scaffolding should be able to support the formal standardization processes of the future. In [137], we have presented the Ubiquitous Health Platform (UHP), which provides semantic reconciliation-on-read based data curation for resolving data interoperability between various schema. This methodology is based on the creation and management of schema maps, that can provide the framework for transforming a source schema into a target schema.

5.1 Research Objectives

The main objective of this research work is to align attributes from heterogeneous schema with each other if they are semantically similar. This alignment provides a bridge between heterogeneous data schema.

In the current study, we will present our research work to build and manage the schema map knowledge base. Overall, our methodology is based on the creation, evaluation, and application of a novel schema matching technique to identify the relationships between attributes of the participating medical data schema. Since the terms used to identify attributes in the data schemas are not defined in any standard way, it is important to first identify the component words of the attribute term and then to append semantic concepts with these to create a meaningful sentence. This process is based on word expansion using concept lookup from Unified Medical Language System (UMLS). Once the sentence has been created, it is then trivial to create its embedded vectors representation using transformer based pre-trained models. The cosine similarity of any two embedded vectors can then indicate the degree of similarity between the original attributes.

5.2 Methodology

Healthcare interoperability, with a focus on non-standard compliant medical schema is dependent on the generation and validation of schema maps as discussed above. To this end, the creation of a cohesive workflow is of utmost importance. In our earlier work [137] we used maximum sequence identification and suffix trees based matching for syntactic matching of two distinct data schemas. This was followed by semantic concept enrichment and subsequently concept matching for creating rules in the form of schema maps. The simplified mapping functions, thereby

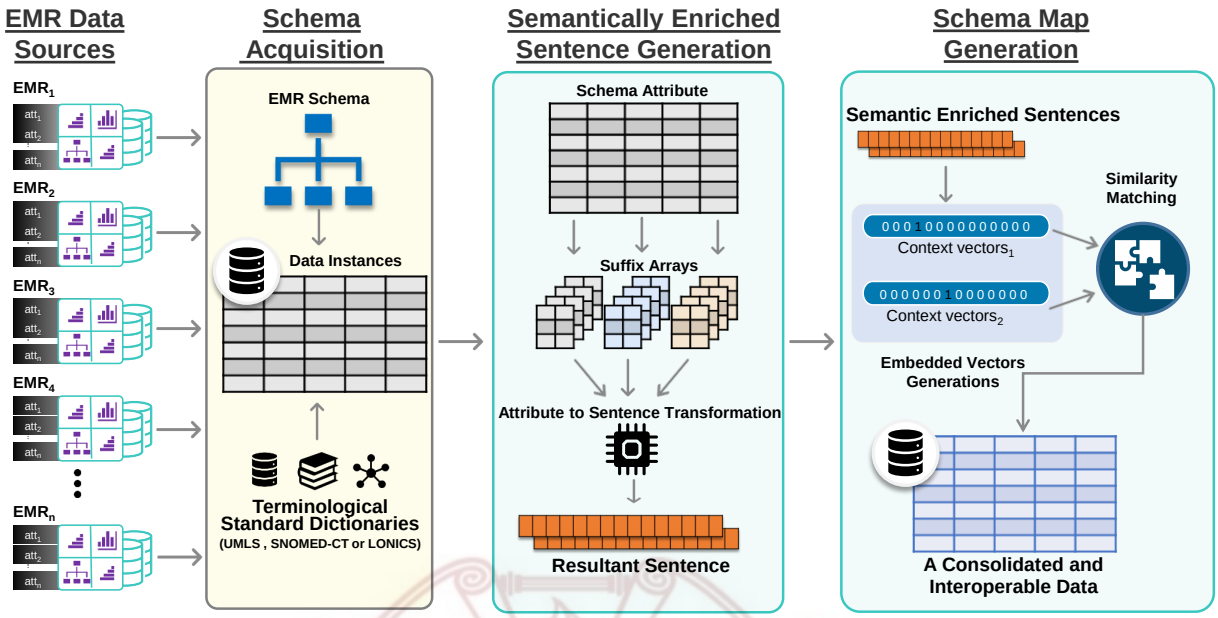

Figure 5.1: Methodology for creating schema maps.

created, provided a simple methodology for converting semi-structured medical data into an interpretable model form. In our current methodology as visualized in Figure 5.1, we have utilized state-of-the-art natural language processing (NLP) techniques to extract the schema mapping rules from semi-structured data schemas. This methodology is based on identifying similarity between vector representations of two attributes, belonging to different medical schemas. Traditional NLP techniques such as Word2Vec are able to convert a word into an embedded vector, while Bidirectional Encoder Representations from Transformers (BERT) extracts an embedded vector from a sentence [138]. Many of the attribute names within data schemas are represented by terms that are bigger than a word (combination of multiple words) and smaller than a sentence. In order to resolve this problem, we extracted the set of suffixes from the terms forming the attribute names. The bidirectional nature of BERT, allows the creation of contextual embedded vectors, where each target word is affected by its neighboring words. Hence to convert the set of suffixes into a sentence, we collected the set of concepts corresponding to each suffix, from UMLS. This operation has two effects, firstly it is used to remove any suffix, which does not have a corresponding concept and secondly the extracted concepts are used to add context to each suffix and produce a

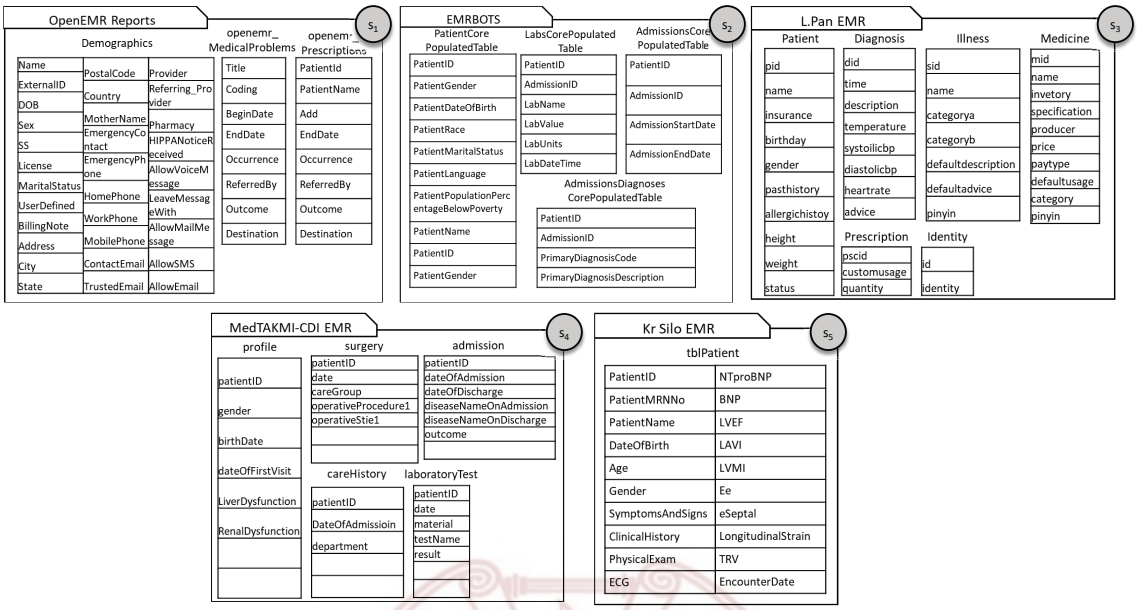

Figure 5.2: The five medical schemas used for achieving data interoperability.

contextual sentence. The following subsections provide the practical details for our methodology from schema acquisition to attribute name expansion, and finally schema map generation.

5.2.1 Schema acquisition

In the first step of our semantic reconciliation methodology, we simulate medical data acquisition from five distinct Electronic Medical Records (EMR) storage systems (S). These include patient reports from OpenEMR (s_1), 100,000 patient records from EMRBOTS (s_2) [139], custom database design by Pan et. al(s_3) for supporting regional clinics and health care centers in China [140], clinical knowledge discovery tool MedTAKMI-CDI (s_4) [141], our custom implementation (s_5) [142] and the schema from Sequence Contraction process (s_6). Each of these medical systems as shown in Figure 5.2, follows the relational database design with logical entities such as demographics, diagnosis, medicine or others, placed into tables which can be further linked to one or more tables. While the database design implemented by each of these systems, fulfills the need of their respective information processing applications, the lack of interoperability, in terms of identifying similar attributes or exchanging the medical data is very much evident here.

A similar notion of data heterogeneity, in terms of medical data schema is evident across the healthcare domain. This is caused by various factors, including the lack of one all-encompassing, and universally applicable terminological standard and different normalization level for representing attributes.

In the former case, while SNOMED-CT provides a mechanism for identifying the standard codes for clinical terms and LOINC can be used for laboratory related terms, most attribute names are created based on the gut feeling of the database designer. Additionally, while these codes can be used to represent elements in the data instances (such as when recording the disease name, a standard code is more beneficial than the text string for semantic interoperability), the elements in a data schema (such as attribute names which are used in queries) achieve no benefit from the same. Consider the terms "name" and "patientName", which refer to the same attribute of the patient entity. However, since there is no standard way to represent this attribute, both are considered correct (s_1 and s_3 use the former representation, while s_2 and s_5 use the latter).

In the later case, differences in normalization also cause semantic differences, due to which some data could be available in one schema but absent in others such as OpenEMR demographics identifying the patient's residential location using specific attributes like "Address", "City", "State", "Postal Code", "Country", and others. Similarly, "EncounterDate" from s_5 is semantically similar to "BeginDate" of "openemr_MedicalProblems" table in s_1, "AdmissionStartDate" of "LabsCorePopulatedTable" in s_2, "time" in "Diagnosis" table of s_3, and "dateOfAdmission" in "Diagnosis" and "CareHistory" tables of s_4. Finally, s_1 and s_3 have separate tables containing the medicinal prescription details, however the same details are unavailable in s_2, s_4, and s_5. Once again, this is not an incorrect behavior since this information, might not be a part of the context or the requirements for the EMR/EHR storage systems. In fact, the change in context of the medical data storage system from the initial time of development to a later stage of collaborative processing systems, is the main cause of heterogeneity. In order to provide an interoperable solution, it is therefore necessary to enhance the semantics of each data attribute by its contextually equivalent sentence.

5.2.2 Attribute to Sentence Transformation

In order to process the EMR/EHR schema set S and produce a set of corresponding semantically enriched sentences, we use the data representation s_i, generated through the process explained in sequence acquisition to collect the various medical fragments in memory. We then iterate over these fragments, building a set of attributes, distinguished by their name, schema's name, table's name, schema's version, source, and recorded data. This entails that "PatientID" from each of the four tables in s_2, and "patientID" from five tables in s_4, would result into nine attributes (assuming, as in the current case of no differences in versions of these systems). For each attribute, we then generate the suffix array, which provides all possible substring representations contained within the attribute name. In order to generate the set of suffixes, we employ three strategies, forward suffix generation, whereby for a word w of length n, $n-1$ suffixes of size 2 to $n-1$ are produced, backward suffix generation, to produce $n-1$ suffixes in reverse order with size $n-1$ to 2, and regular expression based suffix generation, which splits each word on, change of case, special characters(such as -, _, !, and others), and numbers. In this way a large list of suffixes is generated, which is combined using a "TreeSet" data structure of Java, which internally sorts this list as well. An example of this suffix generation process, using the attribute name "dateOfAdmission" as it appears in s_4 is shown in Figure 5.3.

Suffix strings for similar attributes such as "AdmissionStartDate". "diseaseNameOnAdmission", and "AdmissionEndDate" appear in s_2, produce many, synctactically similar suffixes, to the presented example. This process, is only able to generate syntactic suffixes, producing many incoherent and unrelated suffixes. In order to counter this problem, and to limit the list of suffixes within the domain, we then query UMLS with exact search strategy, looking for the existence of any concepts against each suffix. In case, no semantic concept is found for a particular suffix, it is removed from the final Suffix Array. On the other hand, if atleast one semantic concept is found against the queried suffix, it is retained. Meanwhile the process continues for the next attribute, then the next table, and finally the next system, till no further processing is possible. The set of suffixes and their corresponding concepts are then used to build the sentence, where by each concept corresponding to a suffix is appended next to the suffix. An example of the resultant sentence for the attribute "DateOfAdmission" is shown in Table 5.1.

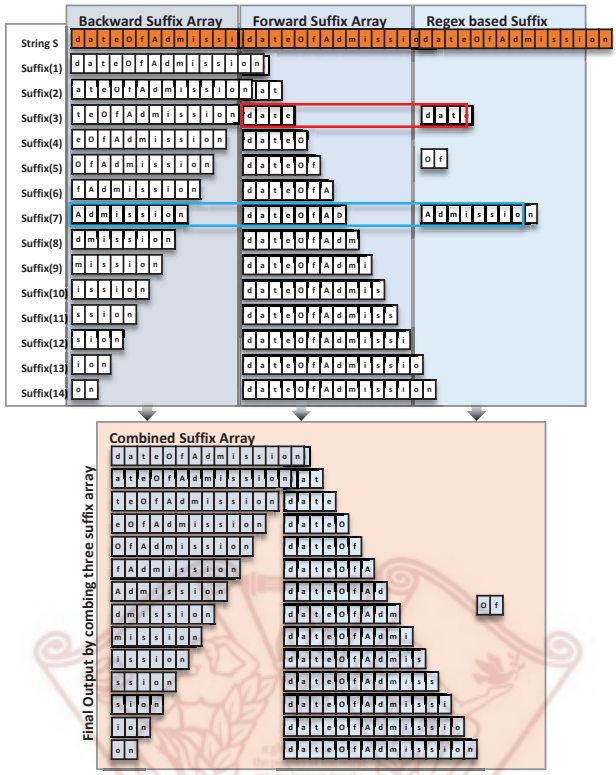

Figure 5.3: An example of suffix arrays produced for the attribute “dateOfAdmission”.

Here the various suffixes and their concepts are separated by the symbol “;”, however together they form one sentence, for which an embedded vector is generated.

5.2.3 Schema Map generation

Schema Maps provide an interoperable bridge between two medical systems ($s_i \wedge s_j$), by identifying the semantic relationship between their participating attributes. This identification is based on the similarity between the embedded vectors, of the semantically enriched sentences corresponding to each data attribute. While the embedded vectors can be generated using any methodology, we tested 11 methodologies with Word2Vec and 10 models based on BERT. Our results indicate that the large/STSb version of Robustly Optimized BERT Pretraining Approach (RoBERTa) [143], provides the best matching results. The pair of embedded vectors thus produced are then used to calculate cosine similarity, which is based on the inverse cosine distance between them. For our classification, we used the raw results (unnormalized) of cosine similarity, which produces a score

Table 5.1: Sentence created from the attribute name “DateOfAdmission”

Date Value type - Date date allergenic extract Date in time Data types - Date Date Fruit;Of SPI1 wt Allele SPI1 gene TAF1 wt Allele BRIP1 gene Within Degrees fahrenheit Oral contraception BRIP1 wt Allele;Da Displacement of abomasum dalton Anterior descending branch of left coronary artery deca units cytarabine/daunorubicin protocol Dai Chinese Asymptomatic diagnosis of Drug Accountability Domain;ion Iontophoresis Route of Drug Administration Ions;on SPARC wt Allele Osteonectin SPARC gene On (qualifier value) Upon - dosing instruction fragment;Admission Admission activity Hospital admission;Dat SLC6A3 gene SLC6A3 wt Allele dopamine transporter Direct Coombs test SLC6A3 protein, human Test Date cytarabine/daunorubicin/thioguanine Alzheimer’s Disease;mission Religious Missions;

between -1, and 1. Cosine similarity score of 0 indicates orthogonal relationship between the two vectors, which in our scenario indicates that the two sentences, and by extension their attributes are not related to each other. -1 indicates inverse relationship between the attributes, while 1 indicates the two attributes are very much the same. For producing our schema maps, we are interested in three types of relationships, “equal” (the two attributes are same), “related” (the two attributes are related to each other), and “unrelated” (no relationship between the attributes). In order to classify the similarity results into one of these three classes, we then calculated the best thresholds using Matthews Correlation coefficient (MCC) [144] for classifying each instance as “equal”, “related”, and “unrelated”.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP+FP) \times (TP+FN) \times (TN+FP) \times (TN+FN)}} \longrightarrow [-1, 1] \quad (5.1)$$

MCC provides a fair measure of the ability with which a classifier can correctly predict both positive and negative instances [144]. The formula for calculating MCC is shown in Equation 5.1, which is based on classification performance measures such as true positive (TP), true negative (TN), false positive (FP), and false negative (FN). MCC score of 0 represents random classification, however, with an increase in the number of true positives and true negatives MCC moves closer to 1. It also takes into account the false positives and false negatives, which shift the MCC score towards -1. This measurement is markedly different from accuracy that fails to account for imbalanced datasets and F_1 measure which is not affected by the true negative scores. As a result, MCC provides an acceptable alternate in our current scenario comprising of imbalanced dataset

(largely in favour of class “unrelated”) to measuring the true performance of the models, used for threshold selection and model evaluation. Finally on a test dataset we evaluated our multi-class classification approach using MCC and Cohen’s Kappa coefficient (κ) [145] to identify the relationships between each pair of attributes.

Chapter 6

Semantic Reconciliation-on-Read

In the last decade, the digital healthcare space has witnessed a rapid technological expansion, which has led to the development and deployment of a plethora of policies, software and devices [1]. As a result, the quality and quantity of healthcare delivery, in terms of diagnostics, treatment, and follow-up has greatly improved [7], [6]. Additionally, supplementary healthcare sources, such as whole-genome sequencing [2], precision medicine [3], Clinical Practice Guidelines (CPGs) [146], and medical Internet of Things (IoT), and others have added new dimensions, to medical data. Today, healthcare data is characterized [73] by its large Volume (number of patients, size of patient data, additional information), Velocity (production rate, which can range from seldom produced non-streaming data to streaming data from medical IoT, like continuous glucose monitor), Veracity (different quality), Variety (formal and/or non-formal standards), and Value (insights).

Consequently, new challenges have emerged in the domain of healthcare, including lack of interoperability, globalization, collaborative capacity gap, tele-medicine, and ubiquitous healthcare [80]. The scale and scope of these challenges, has pushed beyond the scope of traditional data mining and integration techniques. Expert driven solutions are no longer feasible, while machine learning approaches are not mature enough to guarantee complete conversions, every single time. Numerous endeavors have focused on resolving different aspect of the interoperability problem. Our review indicates that most interoperability tools and techniques, work under the assumption, that some form of standards are already in use by the participating medical platforms. On the other hand, the healthcare domain has many formal and even a larger number of non-formal standards(custom data representation, and exchange formats which are at-most used at institutional or regional levels) catering to different aspect of the interoperability problem. As a result, the techni-

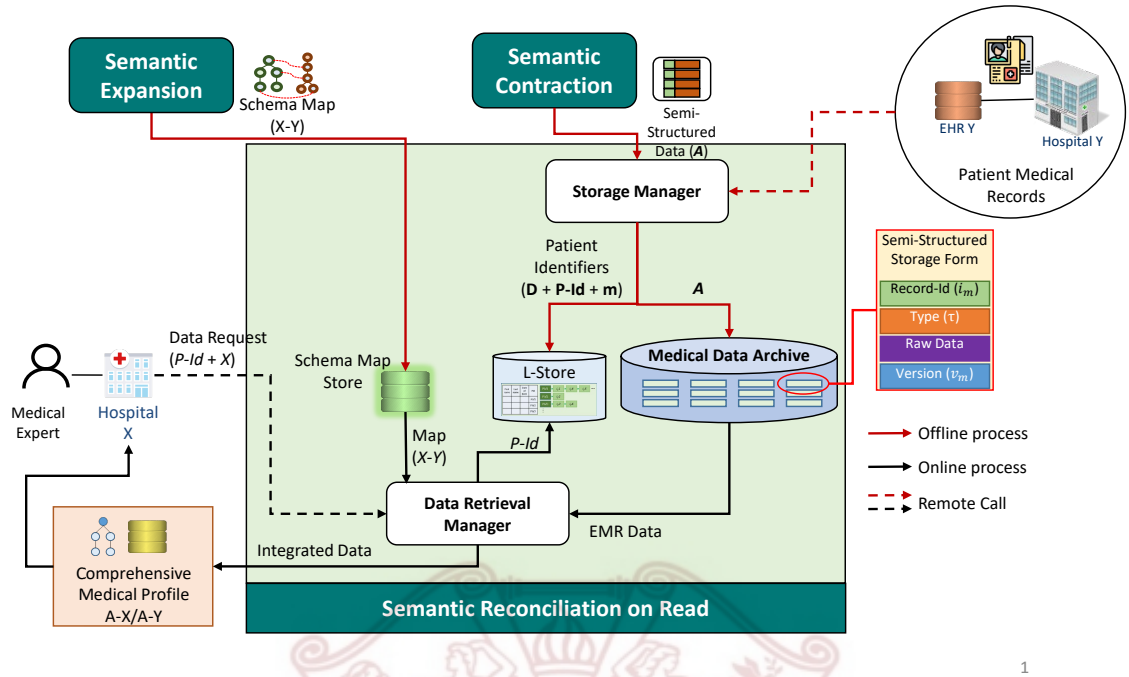

Figure 6.1: Semantic Reconciliation using the Ubiquitous Health Platform

cal aspect of the interoperability problem, can only be solved by applying semantic reconciliation at data, knowledge and process level. Current solutions are focusing on the use of two distinct approaches; a more formal and slower process of standard integration (to merge commonalities and novelties of numerous standards, producing only one universally accepted standard) and mediation based approaches (bridge the gap between all heterogeneous standards for a quick and dirty solution).

Our approach towards resolving this problem, is based on Semantic Reconciliation-on-Read (SRoR) methodology, which is shown in Fig. 6.1. A key part of this large platform is the semi-structured data which is used for storing, integrating and exchanging, multidimensional healthcare data. As shown in Fig. 6.1, the SRoR based Platform acts as a bridge between a patient and medical experts/systems. On the one side of this bridge lies the big data archive service which consumes healthcare data from various sources, extracts meta information related to each patient, serializes the input to strip away its schema and converts it into a relatively flat/denormalized data structure, which is finally stored in a semi-structure form. The other side of this bridge is occupied

by service consumers, which receive the healthcare data in a target structured form, containing either semantically linked or semantically integrated comprehensive medical profile of a patient (e.g. in Fig. 6.1, EHR A and B are transformed into EHR X, using semantic integration).

6.1 Research Objectives

As evident from the discussion above, healthcare interoperability, presents a major challenge towards achieving ubiquitous healthcare. Many factors influence this challenge, including availability of a large number of standards, evolution of standards, privacy concerns around patient data, lack of access to healthcare data, large number of healthcare information management and support systems, and others. In aiming to resolve these problems, one crucial question has been left unanswered in literature, relates to, how do we provide interoperability support to the large number of small and medium scaled HMIS and other healthcare platforms, which are not currently complying with any formal standard?

The Ubiquitous Health Platform, aims to provide a solution to this problem by providing a large medical archive and transformation platform, which can evolve and apply the semantic reconciliation process with changing organizational needs. It is also imperative to mention here, that while initiatives to standardize the standards, like CIMI and Yosemite project are slow in their development, they are necessary for any healthcare interoperability solution to evolve and generalize in future. On the other hand, technologies and platforms such as LinkEHR, OBDA, and HSB provide an alternate to the proposed approach, which have been discussed above and briefly compared in the Table 6.1. The platforms have been compared in terms of their features during data acquisition and data retrieval. This comparison has been based on the available literature only and what has been achieved so far and not in terms of their capabilities which is beyond our scope. In particular, LinkEHR has focused on using well defined archetypes to provide a semantic and syntactic transformation engine, with large input from the knowledge engineer, leading to high dependency on well-defined standards such as HL7 CDA, OpenEHR, and others. Once the mapping has been provided, there is little to no chance of data loss during data acquisition since LinkEHR does not natively store the data, rather it provides transformation on the fly. However, based on how complete the metadata and bridging ontology is, LinkEHR may lose data while data retrieval. Based

on the user query, the internal XML representation may require an additional conversion to the requested form, with the help of two way semantic relations defined by the knowledge engineer. Finally, from the current literature, there is no evidence to suggest that LinkEHR manages the traceability of healthcare records, beyond what may be present in the standard itself.

On the other hand, OBDA is dependent on a well formed ontology, to which all participating systems must comply. As a result adding a new source can become problematic if it does not comply with the structure in the current ontology. Similar to LinkEHR, OBDA does not lose healthcare data owing to its use of only remote data source connections. During data retrieval, data loss by OBDA is dependent on the accuracy of the reasoner and how well the custom API is able to transform the SPARQL queries into native SQL queries. This may lead to some data loss, in terms of the number of retrieved healthcare records. There is no data conversion during acquisition or retrieval by OBDA, however there is very limited traceability in terms of unique identifiers from various healthcare sources, participating in the result set.

Both OBDA and LinkEHR utilize the federated query model to resolve interoperability during data retrieval and are based on well-defined semantic bridges between participating healthcare sources. While LinkEHR uses a one-to-one model, where each pair of systems have a supporting archetype and metadata, OBDA uses a central ontology and thesaurus to bridge many systems together. HSB also uses a semantic interoperability paradigm similar to OBDA, however in HSB, the various healthcare systems, as producers and consumers are only loosely coupled with each other and require transformation services from well-defined standard form to an an internal format for exchanging data. Additionally intermediate conversion at both data acquisition and retrieval phases is required to convert from one standard form into another. Data Loss in HSB is mitigated through the use of buffering queues. Finally, the current implementations do not show any traceability at the data source level.

Finally, the proposed data storage engine, is not reliant on any well-defined healthcare standard but requires serialization of the data and its conversion into a semi-structured format before storage. This process is explained, in some detail, in the next sections. Adding a new data source to the proposed platform is relatively a trivial process and is dependent on writing a simple java class which can read the data, extract meta patient information(name and date of birth), serialize

Table 6.1: Comparison with existing platforms

Method	Data Acquisition			+	Data Retrieval		
	{well-defined standards}	intermediate conversion	Effort to add a new data source	Data Loss	well-defined standards	intermediate conversion	Traceability
LinkEHR	✓	✓	✓	✓	✓	✓	X
OBDA	✓	X	✓	X	✓	X	✓
HSB	✓	✓	✓	X	✓	✓	X
Proposed (SRoR)	X	✓	X	X	X	✓	✓

the data as a single string. The proposed platform on retrieval requires extensive conversion to convert the semi-structured data string into a structured form. Since the methodology is based on archiving of the medical data and semantic maps for bridging schema it does not suffer from data loss. Additionally it also provides traceability for identifying the patient and the source medical system.

Fig. 6.2 shows the comparison between various interoperability paradigms. Fig. 6.2 (a) shows the standards based approach, whereby the semantic reconciliation process is used to transform non-compliant data sources. After this process, the interoperable medical data is in one standard form, which enables the user to execute one query and get the results. An alternate to this approach is shown in Fig. 6.2 (b) and as an example the pipeline typically followed by the federated query approach is shown. In this approach a controller is used to generate separate queries for each of the data sources, these are executed on the corresponding Medical data, the results of which are then combined and shown to the user. Fig. 6.2(c) shows the pipeline of our proposed approach which archives all medical data after integration (conversion to semi-structured form) and then uses Schema Maps to apply semantic reconciliation on subsets based on their individual schema and relationship to the inquired schema. The proposed approach is able to efficiently deal with data volume (using well established Big Data tools and technologies), variety (unlike the other two approaches, requiring less intervention for each integrating new data source), and velocity (by separating the data acquisition and semantic reconciliation process like other mediation based approaches). This platform then provides the foundation for identifying new values from the integrated medical data and enhance its veracity. On the other hand, most interoperability initiatives are tightly bound with existing standards and data exchange interfaces [147]. The novelty of our approach, towards solving the Interoperability problem, lies in delaying the semantic reconciliation process, and thereby moving it away from the data and closer to the user. As a result, semi-structured data has been optimized for acquisition, storage and minimal

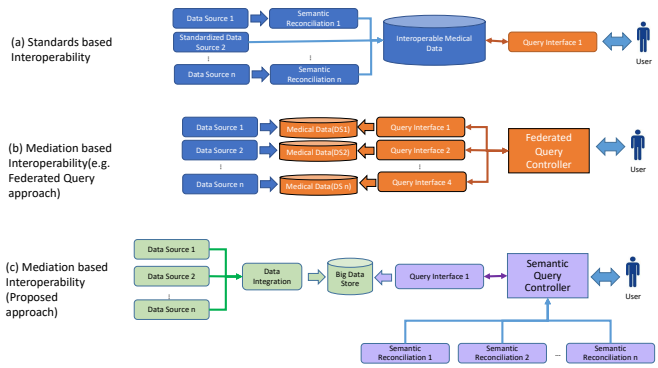

Figure 6.2: Novelty of the proposed approach in terms of the semantic reconciliation pipeline

processing of the medical data.

The use of heterogeneous data models in hospital management and information system (HMIS) obstructs the communication and integration of the systems in clinical workflows. The diverse medical concepts diminish the systems' interoperability. The aforementioned barrier is overcome using semantic reconciliation model, which is proposed in our previous work [148]. The proposed mapping model maps diverse localized concepts, called domain clinical model (DCM), with standard and non-standard medical terminologies. The existing mapping algorithms only focus on the internal semantics of terminologies such as parents, childs, and siblings similarity matching within the source and target terminologies. While in our proposed model, we include the external semantics of the source and target concepts in the form of concepts and relationships provided by the semantic libraries such as UMLS [124] and ConceptNet5 [149].

6.1.1 Theoretical Representation

The semi-structured data storage form, as shown in Fig. 6.3 (a) represents the medical fragments acquired from a variety of healthcare sources and stored in semi-structured form, characterized by a data and a metadata component. The data part of this storage form, also called the Medical Data Archive, is represented by a 4-ary Cartesian product of the set of Identifiers(I), Types(τ), Serialized Fragments(F), and Versions(V). Where set of Identifiers, as defined in Eq. 6.1, is used to uniquely identify an individual record in the storage engine. The identifier is dynamically

generated using any fixed or dynamic length technique [150]. It is not dependent on any features related to the patient or the source medical system and is used for linking the data components with their respective patient's metadata component, which are in-turn identified by their own unique identifiers. Universally Unique Identifier(UUID) as defined in RFC 4122 [151] version 4(pseudo-random) with its 128 bit encoding can be used to uniquely identify upto 5.3×10^{36} objects and is well qualified for use as identifier for both data and metadata components.

$$I = \{i_f | f \epsilon F\} \tag{6.1}$$

The set of Types, defined as in Eq. 6.2, holds a unique identifier for the participating medical fragment schema that a particular medical fragment corresponds to. For practical purposes, the name of the medical fragment schema(such as OpenEMR, HL7 CDA, KrsiloEMR, or other) is sufficient to be used as an identifier. In case of collisions, the name can be augmented by other differentiating features, such as the organization name, country code and so on. This meta information is used to select the appropriate Schema-Map for semantic linking or transformation, during retrieval. Consequently, a medical fragment is considered unique, and becomes a candidate entry in the sent τ, if it has a different schema than the ones already participating. Essentially if a two medical fragments, coming from two different organizations, but following the same schema τ_1, would result in one unique entry in the set τ.

$$Type(\tau) = \{\tau_1, \tau_2, \tau_3, ...\} \tag{6.2}$$

The non-empty set F, defined in Eq. 6.5, represents the serialized form of the medical fragment, provided by a connected medical system M and identified by a type τ. This serialization, de-normalizes the data into key:value form, where each key belongs to and is unique within the schema τ. For disambiguation, keys can be prepended with the database name and table name, if they come from a relational data source. This is used to provide disambiguation between the keys,

which enables correct semantic matching and transformation, at retrieval.

$$F = \{f_m | f_m : \tau \& m \epsilon M\} \tag{6.3}$$

The set of versions V, in Eq. 6.4, represents a ternary of author, timestamp, and the changes to a previous version of the fragment f_m. Version control is provided only for handling minor errors in existing medical fragment data. Any change to the metadata(such as patient's name or date of birth) should be managed by creating a new medical fragment and handling this corner case at the consumer's end. In line with the Big Data architecture, the curation engine discourages any update or deletion of records, which would require a deletion of the entire archive fragment containing many records and reinsertion of the same(a very expensive operation in terms of data consistency and availability).

$$V = \bigcup \{(t, a, v_f) | v_f \subset f_m \& t = timestamp \& a = author\} \tag{6.4}$$

Metadata Storage, also known as the L-Store (Location Store), contains meta elements of the semi-structured data. This store, as shown in Eq. 6.5, provides a logical indexing service, by storing references to the global identifier i_{SRoR}. These references, in turn refer to the medical fragment identifier from Ipatient's meta information and the medical system sourcing the health record.

$$L = \{i_{SRoR} \Rightarrow (i_f, d, m) | i_f \epsilon F \& d \epsilon D \& m \epsilon M)\} \tag{6.5}$$

In addition to the medical schema type τ, some information is also required to uniquely identify the source medical system. This information is available in the metadata component of the SRoR storage form, and is defined in Eq. 6.6. This information is also kept as a single string to keep the overall data structure largely denormalized. Since this information is a part of the metadata, it can become a part of the SRoR engine, only if there is a medical fragment in the archive, sourced from

the medical system (m).

$$M = m|hasFragment(m) \tag{6.6}$$

Additionally, some disambiguation attributes(D) are necessary to keep the global identifiers unique. The selection of appropriate attributes to uniquely and universally identify a patient, across medical systems is a big challenge, which is discussed briefly in section 8.6.A naïve implementation can use the patient's name and date of birth for this purpose. This is shown in Eq. 6.7.

$$D = \{d|\exists d_x \epsilon D : (\forall d_y \epsilon D \rightarrow d_x = d_y\} \tag{6.7}$$

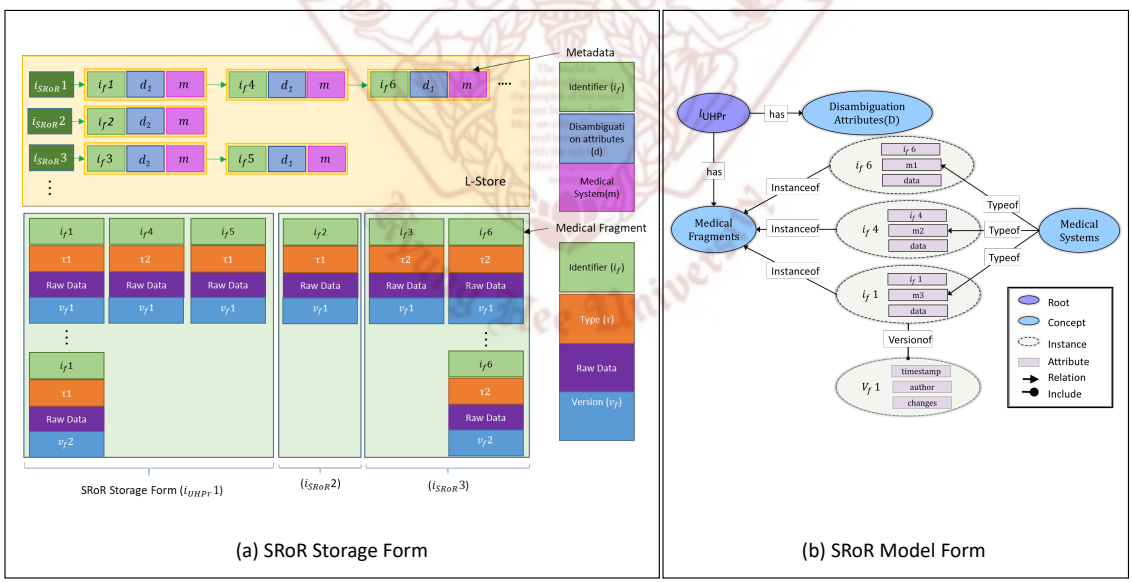

Figure 6.3: SRoR data representation

SRoR model represents the structurally integrated output of the SRoR storage engine. This particular data representation is used to zip together the most important aspects of the user's record and to provide an iterate-able data structure to the consumer. The resulting data structure can be in the form of a well-defined standard(such as HL7 V2, HL7 V3, HL7 FHIR, CIMI archetypes,

or others), or in a graph data structure, shown in Fig. 6.3 (b). This data structure is obtained by structurally transforming the SRoR storage form. SRoR storage conversion to a well-defined standard form requires a supporting schema map, however out of box support for the SRoR graph form is provided by the SRoR engine.

This graph data structure contains the i_{SRoR} as the root node. The root node is linked to patient's disambiguation attributes (D), which can be used by the consuming agent to identify the patient. Additionally, it is linked to the set of all the medical fragments instances belongs to the patient. Each instance is identified by its unique identifier i_f. It also contains the medical system m (from Eq. 6.6) and the data element, which unlike SRoR is semantically enriched to contain semantic relations or transformed into a target schema, based on the retrieval query. Changed versions are linked with their respective data elements for supporting traceability of medical records. The version elements contain the timestamp of change, author information, and the changed data, corresponding to the data element. In this way, the SRoR model is able to re-build a comprehensive medical profile of the patient. This theoretical representation provides the foundational elements of the SRoR engine. It provides the necessary infrastructure for providing data level interoperability, in particular and supporting healthcare interoperability, in general. In the next section we present the implementation details for building the SRoR engine.

6.1.2 Implementation

SRoR Storage

Implementation of the prototype SRoR storage form has been achieved by consolidating information from three medical systems(M), OpenEMR patient reports, 100,000 patient data set from EMRBOTS [152] and our custom implementation of expert driven medical diagnostic system(Krsiloemr). This platform is based on Hadoop, with HDFS acting as the main storage medium, while Apache Hive is used to temporarily create the SRoR schema(shown in Fig. 6.4) and fetch all records for the patient. The SRoR hadoop deployment is composed of 1 master and 2 slave nodes with 1.8TB HDFS size, 20 MB block size, Block Replication of 3. The master node has 64GB ram, while the slave nodes have 32GB ram. Each unit of this cluster has 4 core AMD Ryzen 3 2200G processor([153]), and has CentOS([154]) 7.5 as the operating system.

The SRoR storage form as shown in Fig. 6.3 (a), is stored in form of text files in HDFS, which

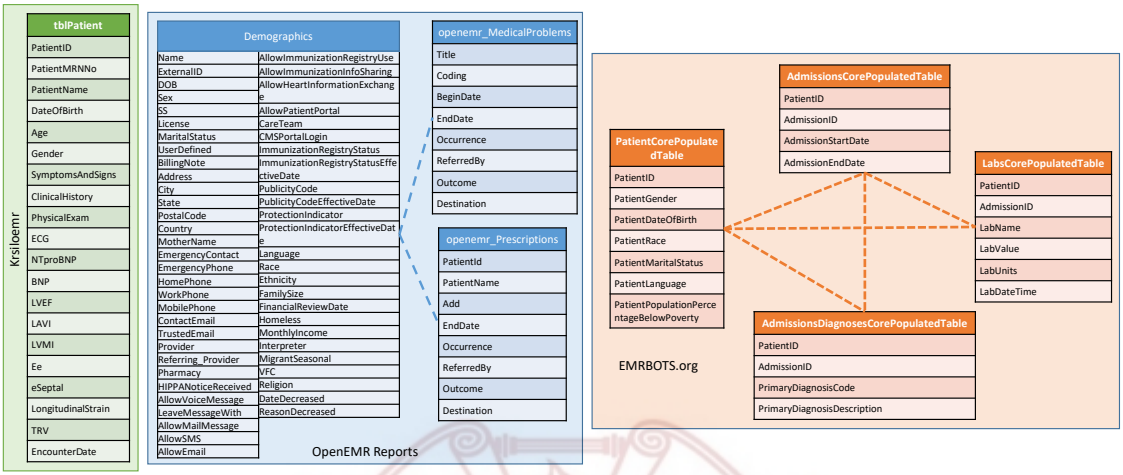

Figure 6.4: Schemas for Medical Fragments participating in SRoR

in turn, contain various medical fragments in semi-structured form. With Hive we temporarily create a schema, utilizing the semi-structured elements (the identifiers) and perform complex queries, which are then converted into MapReduce operations. Each patient is assigned a global identifier(i_{SRoR}) using a 128 bit UUID which maps each patient's firstname, lastname, and date of birth with a related medical fragment (f_m). Medical Data Archive, stores the medical fragment in block form, where many medical fragments are combined together into one file (identified by the global id). The medical fragments, in turn, contains, the unique identifier, as available in the L-Store (different versions of the same medical fragment, will have the same identifier). Additionally, it contains a type element, which is used to identify the schema of the medical fragment and will be used later on for using the correct, ontological map for transformation. Each fragment also contains a locally unique version identifier, which is used for managing instance evolution and verification purposes. Starting with 40 real patients in Krsiloemr and 12 patients for openEMR with various medical problems, we generated medical fragments for 80,000 patients. Each patient has 1 openemr-Demographic Report, and is randomly assigned another 29 medical fragments amongst

Krsiloemr, openemr-MedicalProblems, and openemr-Prescriptions. After 7 iteration and including the 100,000 patient dataset from EMRBOTS, the data store now contains 115,737,428(a little over 115 million) records, corresponding to 390,101 patients. The dataset from EMRBOTS was slighlty modified to include 'PatientName'(since this is required for our approach), before being serialized into a SRoR compliant format. The schema for our three participating medical systems is shown in Fig. 6.5.

Through our experiments, we were able to determine that the most feasible strategy to store

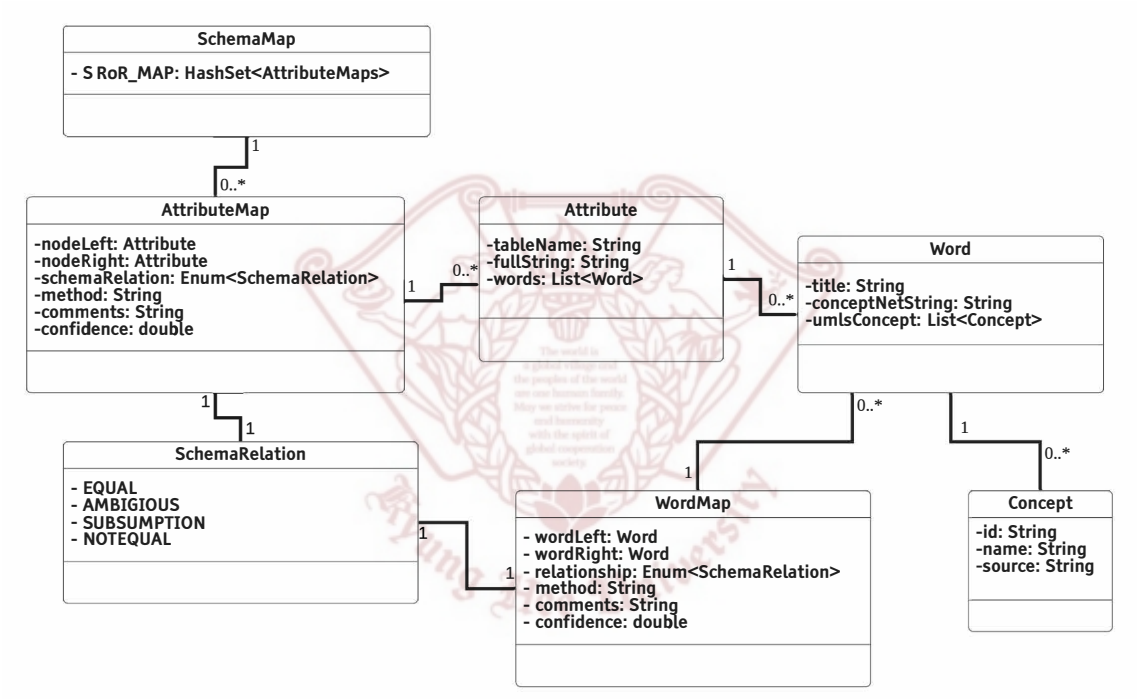

Figure 6.5: Modeling SRoR Maps

these fragments, along with L-Store metadata, in HDFS is by using a 1 file-per-transaction strategy [155]. In this strategy, we consolidate various medical fragments, from 1 transaction(similar to data buffering) into 1 metadata, 1 data, and 1 connector file. The metadata file, contains the meta information for the L-Store, the connector file contains index entries for mapping i_{SRoR} identifiers to i_f identifiers, and the data file contains the medical fragment, corresponding to each i_f identifier.In this way we can store a large amount of data in relatively smaller number of files. This strategy enables the most preferred way of data processing using MapReduce operations,

Table 6.2: Hive Queries

ine **Id**	**Query**	**Description**
ine Q1	select medicalfragmentidx.fragmentid, uhpridx.firstname, uhpridx.lastname, uhpridx.dob, uhpridx.gid from medicalfragmentidx,uhpridx where medicalfragmentidx.gid=uhpridx.gid AND uhpridx.firstname="Harry" AND uhpridx.lastname="Potter" AND uhpridx.dob="19880708";	Selects the fragment id, patient's first name, patient's last name, patient's date of birth, and global identifier, from the L-Store, for user named "Harry Potter" who was born on 19880708.
ine Q2	select * from uhpr where fragmentid in (select fragmentid from medicalfragmentidx where gid=(select gid from uhpridx where firstname="Harry" AND lastname="Potter" AND dob="19880708"));	Selects the medical fragments from UHPr storage form, by matching the global identifier for the patient named "Harry Potter", who was born on 19880708.
ine Q3	select fragmentid from medicalfragmentidx where gid=(select distinct(gid) from uhpridx where firstname="Harry" AND lastname="Potter" AND dob="19880708")	Select the fragment id from L-Store for the patient named "Harry Potter" who was born on 19880708, selecting only distinct global identifiers first.
ine Q4	select distinct(fragmentid) from medicalfragmentidx where gid=(select gid from uhpridx where firstname="Harry" AND lastname="Potter" AND dob="19880708")	Select only the unique fragment id from L-Store for the patient named "Harry Potter" who was born on 19880708.
ine Q5	select * from uhpr where fragmentid in (select fragmentid from medicalfragmentidx where gid=(select distinct(gid) from uhpridx where firstname="Harry" AND lastname="Potter" AND dob="19880708"));	Selects the medical fragments from UHPr storage form, by matching the distinct global identifier with the fragment id for the patient named "Harry Potter", who was born on 19880708.
ine Q6	select * from uhpr where fragmentid in (select distinct(fragmentid) from medicalfragmentidx where gid in (select gid from uhpridx where firstname="Harry" AND lastname="Potter" AND dob="19880708"));	Selects the medical fragments from UHPr storage form, by matching the distinct fragment identifier with the global id(s) for the patient named "Harry Potter", who was born on 19880708.
ine		

with small number of large sized files. [156, 157] As a result of this process, the SRoR is able to achieve transactional consistency. For data processing we then move the relevant records into memory by employing a temporary external table (schema-on-read), created using Hive. Using simple Hive Query Language (HiveQL) based queries (as shown in Table 6.2) we are able to retrieve the medical fragments belonging to a particular user.

SRoR Model

Structured output of the SRoR storage engine is presented in the form of SRoR model. This representation is retrieved after applying SRoR Maps, as semantic bridges, between the various attributes of the participating schema. In order to implement the SRoR Model, a java based application reads the medical fragment data from a csv file, generated from Hive. It also reads

the SRoR Maps JSON file and loads the Schema Maps in memory. Then based on the name of the schema for each medical fragment, it reads the appropriate schema map, and generates the graph form of the SRoR model. Based on user request, the SRoR Model generator can either add the all AttributeMaps belonging to an attribute in the output graph, or it can read the use the AttributeMaps linking the source and target attributes and apply transformation, if the confidence score of the mapping is above some user defined threshold. The class dia-

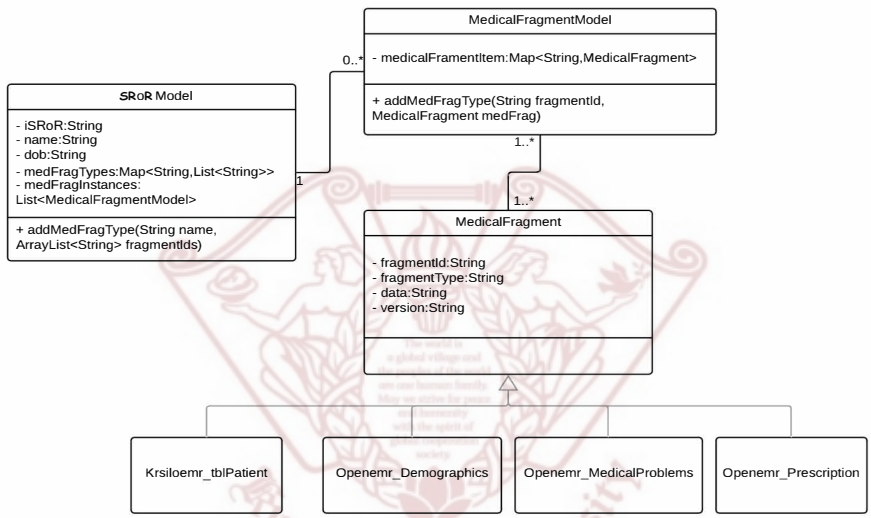

Figure 6.6: Class diagram, representing the SRoR model building application

gram for this model is shown in Fig. 6.6. It uses the UhprModel as the base class holding the root node, and MedicalFragmentModel, holding a map of fragmentId and the medical fragments. The MedicalFragment itself, is a parent class of our 8 specialized medical systems: ‘ Krsiloemr_tblPatient ’, ‘ Emrbots_PatientCorePopulatedTable ’, ‘ Openemr_Demographics ’, ‘ Openemr_MedicalProblems ’, ‘ Openemr_Prescription ’, ‘ Emrbots_LabsCorePopulatedTable ’, ‘ Emrbots_AdmissionsCorePopulatedTable ’, and ‘ Emrbots_AdmissionsDiagnosisCorePopulatedTable ’.

The same application then transforms the SRoR object form into JSON based graph form. Here, loading the SchemaMap into memory and its deserialization into object form took 988 millisec-

onds. For 1 patient with 30 records, the semantic linking process took under 3 seconds. While the semantic transformation process for the same user too 404 milliseconds. The SRoR model, in this graph form, is then transformed into a user friendly format, as shown in Fig. 6.7.

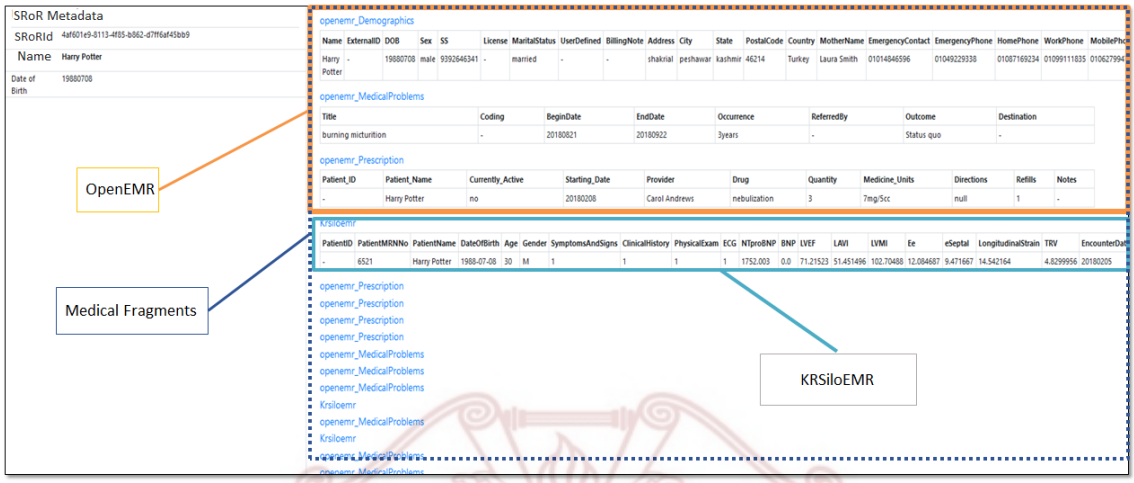

Figure 6.7: SRoR Results for selected user

Availability of Data and Software

All code(for creating, transforming, and view), some sample data(minus the EMRBots data set), and results related to this version of the SRoR are available in a public GitHub repository(https://github.com/desertzebra/UHPr).

Chapter 7

Results and Evaluation

7.1 Experimental Setup

7.1.1 Sequence Contraction

Data Acquisition

In order to collect and prepare the initial conversational dataset, we first obtained official consent of the two participating medical centers (DHQ-Kotli and Care+ MC-Islamabad) in Pakistan to collect data for this study. Two practicing physicians, then recorded their conversations with patients and guardians at these hospitals. Overall, 148 unique clinical interactions were collected from DHQ-Kotli and 19 from Care+ MC-Islamabad. Each participant signed a consent form before the start of the conversation and was explained the necessity of this research work, verbally as well. Since the conversations between the physicians and patients/guardians were conducted in a local language (Urdu), three human transcribers were hired to transcribe the contents of each conversation, and translate it into English. Two transcribers processed 74 audio files each, while one processed 19 conversations. All three transcribers were female with at least 14 years of education[1]. The English text of these conversations were then anonymized by switching patient names, from them. Next, we removed the introductory explanations of the study, from each text. This text is sent for pre-processing using methodology from Section 4.2.1 and produces results which are presented in Section 7.2.1.

After pre-processing the set S is split into three parts, with 508 sentences used for fine-tuning DistilBERT and creation of the enriched sequences (E) used in MASS, 464 sentences for threshold

[1] Annotator 1 has a Bachelors in Business Administration from Bahria University, Islamabad, Pakistan.
Annotator 2 has a Bachelors in Computer Software Engineering from Foundation University, Rawalpindi, Pakistan.
Annotator 3 has a Bachelor of Dental Surgery from Lahore Medical and Dental College, Lahore, Pakistan

selection, and 1281 sentences for concept extraction, schema mapping, and expert verification.

Model Training for Sequence Encoding

In order to convert the textual sequences obtained from conversations into embedding vectors, optimized for sentence similarity in the medical domain, we fine-tuned the base DistilBERT uncased model with a custom annotated dataset. To create this dataset, we first created a combination of the sequence set with itself ($S \times S$), to produce a set of unique sequence pairs. With 508 sequences in S, the combination set produced 129,272 pairs. For each pair, we then manually marked the two sequences as similar if they were intuitively equal and dissimilar otherwise. A pair of sequences, such as "how old is he? 5 years" and "whats her age? She's 15 years old" are semantically similar, however, the pair "what is child's name? h*****" and "the child has cough" are dissimilar. This produced a set of 6,464 similar sequences. We then randomly selected 6,464 dissimilar sequences from this set to produce a balanced dataset of 12,928 pairs. These pairs were further split into 70% training instances and 30% validation ones.

We tested various hyperparameters[2], to optimize the sentence similarity evaluation, eventually selecting the batch size of 32, the "Sparse Categorical Cross Entropy" loss function, "Sparse Categorical Accuracy" as the evaluation metric, and AdamW optimizer [158], with an initial learning rate of 1e-4, 10% warmup steps, and 12 epochs. As a result of this fine-tuning activity, our model shows an accuracy of 95% on the test instances.

Threshold Selection for Sequence Classification

In order to determine the optimal cosine similarity above which a test sequence can be classified as semantically similar to MASS, we evaluate 100 thresholds between 0.0 to 1.0 with a step size of 0.01. At each step, we calculate the area under ROC (AuROC), which provides a numeric measure to evaluate the performance of correct and incorrect classification for the original positive instance. Hence, for all evaluations, when a test instance has cosine similarity equal to or greater than α, with any instance from MASS, it is considered as medically aligned. After plotting all the values,as shown in Figure 7.1, the best AuROC is achieved at 0.87. This is the value of α, which

[2]The details of the fine-tuning process is left out to keep this manuscript concise

is used for classifying a test instance as similar to one of the instances in MASS.

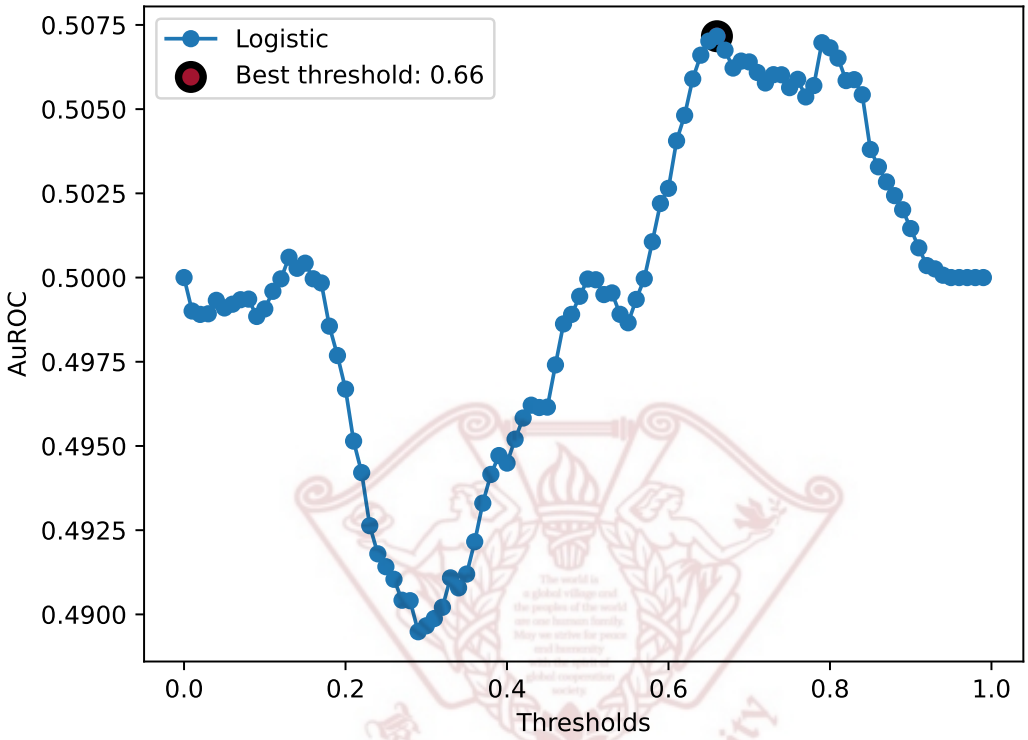

Figure 7.1: Plot between AuROC and threshold values between 0.0 and 1.0.

7.1.2 Sequence Expansion

In order to identify the set of attributes, we used six participating medical systems, including patient reports from OpenEMR (s_1), 100,000 patient records from EMRBOTS (s_2) [139], custom database design by Pan et. al(s_3) for supporting regional clinics and health care centers in China [140], clinical knowledge discovery tool MedTAKMI-CDI (s_4) [141], a custom implementation for cardiovascular disease management (s_5) [142], and the semi-structured schema obtained from the application of Sequence Contraction method (s_6). For each of these six systems we simulated medical data acquisition by generating over 115 million patient records, which are converted into a semi-structured form and stored in Hadoop Distributed File System (HDFS). Further details

of this generation process follow in Section 7.1.3. Medical fragments, thus produced, follow various schema design patterns supporting a variety of valid relational storage architectures. Such as, s_1, s_2 and s_4 are represented by creating a separate medical fragment for each participating table, s_3 utilizes its medical fragment to generate a linked record (from a linked object graph), where by the attributes can refer to other objects, mimicking the application of explicit foreign keys, and s_5 and s_6 represent a flat table structure. The code to generate this data set is available at "uhp_map_generation"[3]. Using the medical fragments file, we then generate the semantically enriched attribute [4], which contains the suffixes and their concepts corresponding to each EMR data attribute. The resulting set of enriched attributes are temporarily stored in a JSON file, which is then read by the same application to partially generate the schema maps. This process is used to create 48,826 distinct pairs of attributes across s. Each pair also contains the "relationshipList", which stores the results of fuzzy string matching [159] [5] between the attribute names. The JSON file thus produced is then used by a python script to generate the semantically enriched sequences and their embedded vectors using Word2Vec, and 7 transformer based NLI models [143]. The NLI models include the fine-tuned DistilBERT-base-uncased from 7.1.1, all-MiniLM-L12-v2, all-mpnet-base-v2, all-MiniLM-L6-v2, all-distilroberta-v1, multi-qa-distilbert-cos-v1, and multi-qa-MiniLM-L6-cos-v1.

The embedding vectors produced by encoding the enriched sequences are then compared using cosine similarity. The rationale behind switching the applications at various stages is to cache the results and create checkpoints for restarting any failed stages, easily. Additionally, since python provides better support for easy generation of embedding vectors, it was thus preferred over the Java based implementation, which is otherwise very beneficial for other tools. These applications were executed on a workstation running Ubuntu 20.04.2 on top of AMD Ryzen 3 2200G, and 32GB ram.

[3] https://github.com/desertzebra/UHP_v4/tree/main/uhpr_storage
[4] https://github.com/desertzebra/UHP_v4/tree/main/uhp_map_generation
[5] Java Library: https://github.com/xdrop/fuzzywuzzy

Table 7.1: Evaluation criteria for each iteration

Id	Description	Metric
C1	Time taken to insert SRoR medical fragment file into HDFS	Time
C2	Time taken to insert medical fragment bridging information, linking global id(g_id) with fragment id($f_id) into HDFS	Time
C3	Time taken to insert SRoR patient index part of L-Store into HDFS	Time
C4	Time taken to create SRoR table schema in Hive	Time
C5	Time taken to create medical fragment bridging table schema in Hive	Time
C6	Time taken to create SRoR patient index table schema in Hive.	Time
C7	Time taken to retrieve all fragment ids for 1 user	Time
C8	Time taken to retrieve all medical fragments for 1 user	Time

7.1.3 Semantic Reconciliation-on-Read (SRoR)

Starting with a set of 2.4 million synthesized medical fragments against 80,000 patients, we performed 7 iterations to increase the data and test the three metrics. Data for the first 6 iterations is based on 40 real patients in Krsiloemr and 12 patients for openEMR. In iteration 7, we used the EMRBOTS dataset of 100,000 patients. In each iteration, we evaluated 8 timeliness criteria to evaluate the performance of data insertion into HDFS (corresponding to SRoR storage form), creation of temporary schema in Hive, and timeliness of data retrieval (corresponding to the transformation process from SRoR storage form to model form). These are shown in Table 7.1. In order to test the actual transformation of medical fragments from SRoR storage form to the model form, we executed Q1 and Q2 in iteration 1-5, while Q3, Q4, Q5, and Q6 in iteration 6 and 7, to retrieve medical fragment ids and medical fragments, respectively. The queries and their description is shown in Table 6.2. These were repeated 10 times to strengthen the results. The evaluation results of these iterations and the relationship of the evaluated criteria across them is discussed as follows:

In the first iteration we started by generating medical fragments for 100 patients, with 20 medical fragments per patient. Total number of medical fragment instances for the user "Harry Potter", who was born on 19880708 were 30(same as iteration 0). The results for executing Q1 and Q2 10 times, for criteria C7 and C8 respectively, is shown in Fig. 7.2. The average time taken for C7 is 28.8528seconds and for C8 is 119.1014seconds. In the second iteration, the number of new patients was increased to 10,000, with each having 20 medical fragments. Executing Q1 and Q2,

Figure 7.2: Iteration 1-5 results for C7 and C8 after executing Q1 and Q2

10 times, for criteria C7 and C8, respectively, yielded the results shown by Fig. 7.2. The average time taken for C7 is 28.4869seconds, while for C8 is 121.4805 seconds.The total number of rows returned by these operations were 30(same as iteration 0). In the third iteration, 40,000 new patients with 20 medical fragments each was generated. The results for this iteration are shown in Fig. 7.2. The average time for C7 is 30.9533 seconds and for C8 is 128.011 seconds. In the fourth iteration, 80,000 new patient records were generated, with 30 fragments for each. As shown in Fig. 7.2, the average time for C7 is 33.0076 seconds and for C8 is 139.1931 seconds. Similar to the previous iteration, 80,000 new patients with 30 fragments each were added as a new SRoR storage file in the HDFS. As indicated by the results shown in Fig. 7.2 there is only a slight increase in the amount of time consumed by Hive. With an average time of 33.7804 seconds for C7 and 148.0349 seconds for C8, there is a slight increase of 0.7728 seconds for parsing the medical fragment identifiers and a relatively larger increase of 8.8418 seconds for retrieving the SRoR storage forms. Here, the former can be explained by the small size of each row, while the latter is the result of processing a large amount of text, especially in the raw data column.

The main aim behind iteration 6, was to evaluate the accuracy of the SRoR when new medical

fragments for a particular patient are added. In this iteration we generated an additional 40 medical fragments for our selected patient. It is also important to point out here that while the SRoR platform and the selected queries, allow for non-unique gid(i_{SRoR}), the theoretical model is based on these being unique for individual patients. As a result, the gid of the new fragments was also matched with the already existing one. The number of medical fragments for the selected patient were increased to 70(These exist in two distinct files for SRoR, and L-Store with the split 30-40). On executing Q1, the total number of rows returned were 140 in 32.918 seconds. The results indicated that each fragment id was repeated twice, which is the result of multiple "Map" operations, converging without consolidating their records. While this is not an erroneous execution, it is still undesirable for our use case. As a result, we switched the queries to Q3, Q4, Q5, and Q6. Executing in two sets of 10 repetitions each, we first calculated the results of Q3 and Q5, followed by 10 repetitions of Q4 and Q6.

In the first case, used the distinct function on the inner most query, which would produce a set of unique gid (which is 1 only), further used to retrieve the fragment ids and eventually the medical fragments. The results for this case are shown in Fig. 7.3 (a). On the other hand, Fig. 7.3 (b), shows the results of the second case, whereby the distinct function was applied on the outer query in Q4/middle query in Q6, to produce the unique set of fragment ids, eventually used for retrieving the medical fragments. The distinct operation is executed via the "Reduce" operation in Hive, which consolidates the results, leading to 70 correct records, every single time. In iteration 7 we introduced the large dataset from EMRBOTS into the platform after tweaking it to include randomly generated patient names(a requirement for our platform). The dataset contains 100,000 new patients, along with their corresponding record of 107,535,388 fragments(from Admission, Admission Diagnosis, and Labs table). Average time for C7 is 264.9 seconds and for C8 seconds. Here again, there was a substantial increase in the query execution time, as shown in Fig. 7.4. However the returned results were error free and conform with the platform scalability, discussed in the following section.

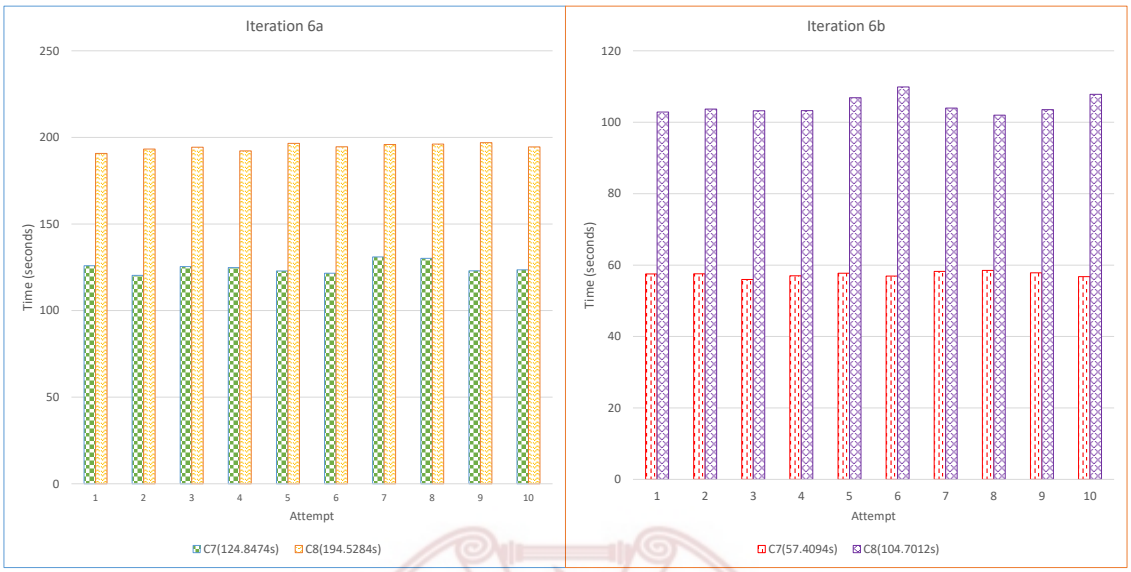

Figure 7.3: . Iteration 6 (a) results for C7 and C8 after executing Q3 and Q5 and (b) for C7 and C8 after executing Q4 and Q6

7.2 Results

7.2.1 Sequence Contraction

In order to evaluate the correctness of our methodology, and its conformance to the challenges stated in Section 4.1, we conducted several experiments. Some of the most important results are presented as follows.

Pre-Processing

The conversational instances, in text form, were pre-processed to convert them into the set of sequences S. This set is then divided into four parts, described in Table 7.2. Here, we have used the same data for fine-tuning our DistilBERT model (used for sequence encoding) and to create the MASS. For threshold selection and evaluation, unseen data was used.

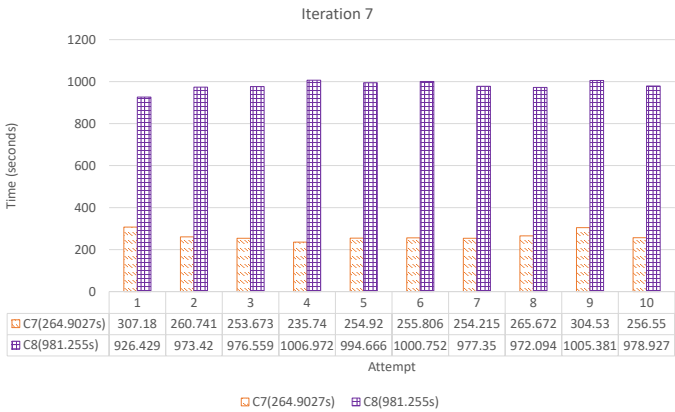

Figure 7.4: . Iteration 7 results for C7 and C8 after executing Q4 and Q6

Table 7.2: Dataset division in terms of its usage

Conversations	**Sentences**	**Activity**
30	508	DistilBERT fine-tuning & MASS Creation
30	464	Threshold Selection
88	1281	Testing labeled sequences
19	827	Testing unlabeled sequences

Model Development

In order to create MASS, we processed 508 sentences and created 190 instances for E. Each instance is a partial representation of some sequences from the training dataset, marked by identifiers, such as [CLS], [SEP], and [MASK]. Thus the sequence, "what is your issue? sir i am having severe flu and cough along with little fever", has the following corresponding enriched sequence, "[CLS] what is your issue? [SEP] sir i am having severe [MASK] and [MASK] along with little [MASK];;Sign or Symptom:cough,Disease or Syndrome:flu,Finding:fever;;umls". Here the three elements of e are separated by ";;", where the first is used for generating the embedding vector, the second contains the labels ("Sign or Symptom:....:fever"), and the third part contains the extraction function ("umls").

Each instance must have at least 1 label pertaining to some text within the original text. As illustrated in Figure 7.5 (a), MASS has 30 instances with multiple labels, and 160 with a single label. The count of unique labels in E is shown in Figure 7.5 (b), where 65 instances utilize regular expressions for 5 types of labels, while 7 labels from UMLS are used with the UMLS type extractor function.

The extractor function can be either a regular expression or UMLS based. In the case of the former, we add a regular expression that is closer to a pattern in the original text and is also useful to extract values from the target text as well.

$$(old|age)(.*)?\ \ ?(heis|hes|sheis|shes)?(?P < Age > .*)(years|month)?(.*)? \tag{7.1}$$

Consider the regular expression in Equation 7.1, which is the extractor function associated with the text, "how old are you? 18 years old". This can be used to extract "18 years old" not only from the source text but also from a target sentence such as, "and how old is he? 3 years" from the labeled test dataset. The extracted value, in this case, is "3 years".

Evaluation of the sequence similarity model

In order to evaluate the performance of our fine-tuned DistilBERT-base-uncased model, we have used the Semantic Textual Similarity benchmark (STSb) dataset [160]. The test dataset[6] contains, 1379 sentence pairs, which have been built from news items (500 instances), captions (625 instances), and forum (254 instances).

For each sentence pair instance, we first encoded the textual sentences to create embedding vector using[7] the pretrained all-mpnet-base-v2 model, the pretrained DistilBERT-base-uncased model, and the proposed fine-tuned DistilBERT-base-uncased model. We then calculated the cosine similarity between the embedding vectors. The similarity measure is then rescaled from 0-1 to 0-5, so as to identify the annotated labels for each sentence pair. Then we calculated the Pearson Correlation (r) between the computed similarity and the ground truth for the STSb dataset. The final results on the STSb test dataset for squared Pearson Correlation (r^2) are shown in Figure 7.6

[6] http://ixa2.si.ehu.eus/stswiki/index.php/STSbenchmark
[7] without downstream training

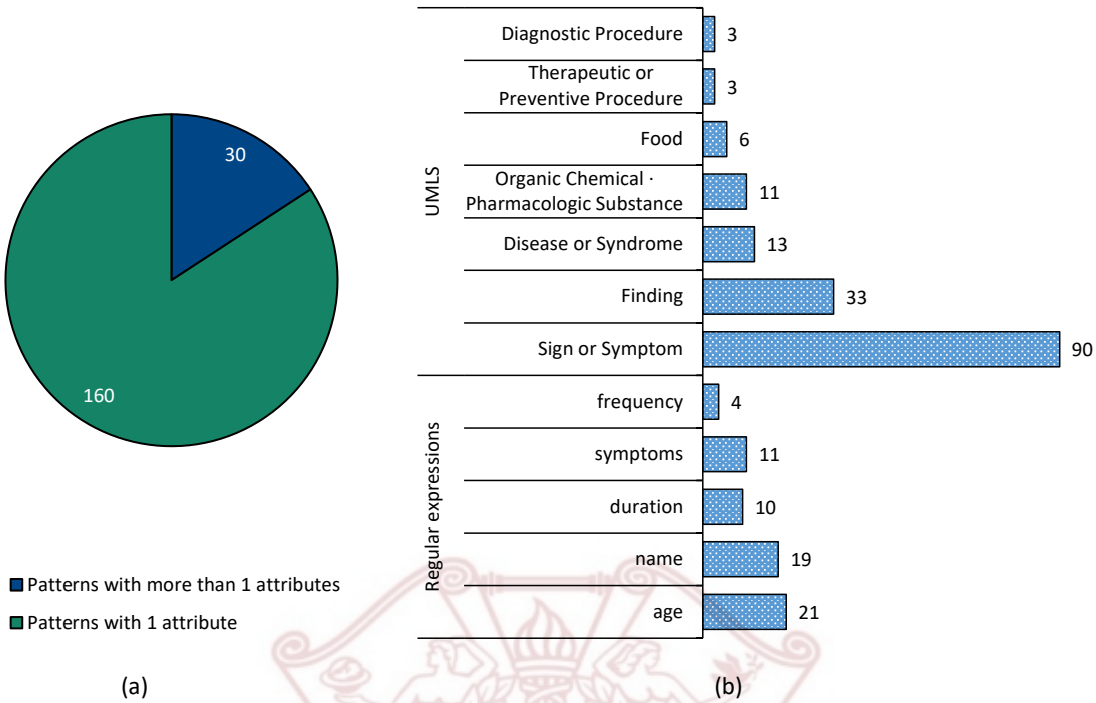

Figure 7.5: Statistical information for MASS instances (a) showing the ratio of e with single vs multiple attributes, and (b) showing the unique labels and their extraction methodology.

(a) - (c). On this cross-domain dataset, the performance of the pretrained all-mpnet-base-v2 model at r^2 of 0.70, far exceeds the pretrained DistilBERT-base-uncased model at r^2 of 0.31, which is itself higher than the fine-tuned DistilBERT-base-uncased model at r^2 of 0.22.

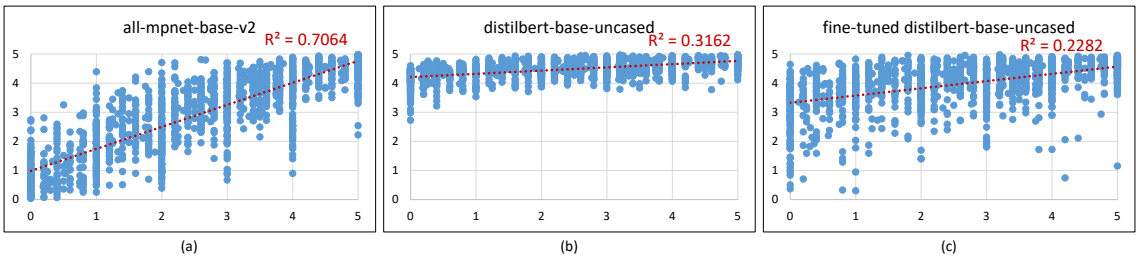

Figure 7.6: Squared Pearson Correlation (r^2) of semantic sequence similarity between the annotated similarity and similarity computed by (a) by pretrained all-mpnet-base-v2 model, (b) pretrained DistilBERT-base-uncased model, and (c) fine-tuned DistilBERT-base-uncased model.

Table 7.3: Performance evaluation of the proposed methodology on labeled test instances and its comparison with the baseline methodology

Methodology	Sequence ClassificationThreshold	Accuracy	Precision	Recall	F1 Score
Proposed Approach	0.87	52.96%	69.34%	29.44%	41.33%
Baseline Approach	0.49	44.47%	44.79%	30.33%	36.17%
SciBERT	0.38	48.81%	50.89%	37.54%	43.21%

Dataset Validation

In order to fullfill the "*Challenge 1*" stated in Section 4.1, whereby the attribute name's should be correctly identified, we have used a labeled Test dataset containing 1,281 sequences (1,201 unique), obtained from 88 conversations. Out of these, 546 sequences are in the "False" class and 655 sequences have an associated label and are used to form the "Truth" class. Using our custom DistilBERT model, we encoded each test sequence before matching it with MASS. Overall, this process performs 243,390 vector comparisons. Using the similarity threshold α of 0.87, we are able to filter the low-performing comparisons and end up with 68,414 instances. We then apply the value extraction function to these instances. By using simple token-based local caching, we store the semantic types obtained from UMLS against unigram and bigram tokens from the text. Since querying UMLS is a slow process and is prone to blocked traffic, caching is very important. The cache contains semantic types for 3,977 tokens, out of which 1,144 are unigram tokens and 2,833 are bigram tokens. With 10,270 words in the test dataset, only 11% unigram tokens are semantically equivalent to atleast one concept in UMLS. For 1201 test instances, with sequence similarity greater than α, with at least one instance from MASS, and some extracted value, the accuracy is 72.94%, and F1 score is 68.23% (as shown by "Proposed Approach" in Table 7.3).

To compare our results with a baseline model, we replaced our fine-tuned DistilBERT model with a pre-trained sentence similarity model, "all-mpnet-base-v2". This model is trained on over 1 billion tuples and provides the best results[8] for Sentence Embeddings (69.57% on 14 diverse datasets) and Semantic Search (57.02% on 6 diverse datasets) tasks. Using the same strategy of identifying the attribute name correctly and validating the existence of a corresponding value, we evaluated the results of the baseline, referred to as "Baseline Approach" in Table 7.3.

These results show that the baseline methodology shows better performance in resolving

[8]https://www.sbert.net/_static/html/models_en_sentence_embeddings.html.

"*Challenge 1*". It is better in terms of precision and accuracy than the baseline approach

7.2.2 Sequence Expansion

The validity of our proposed approach presented in Section 5 has been evaluated using several techniques including comparison of the proposed semantic matching process with fuzzy string matching, embedded vector generation and comparison using Word2Vec, and 10 BERT nli models.

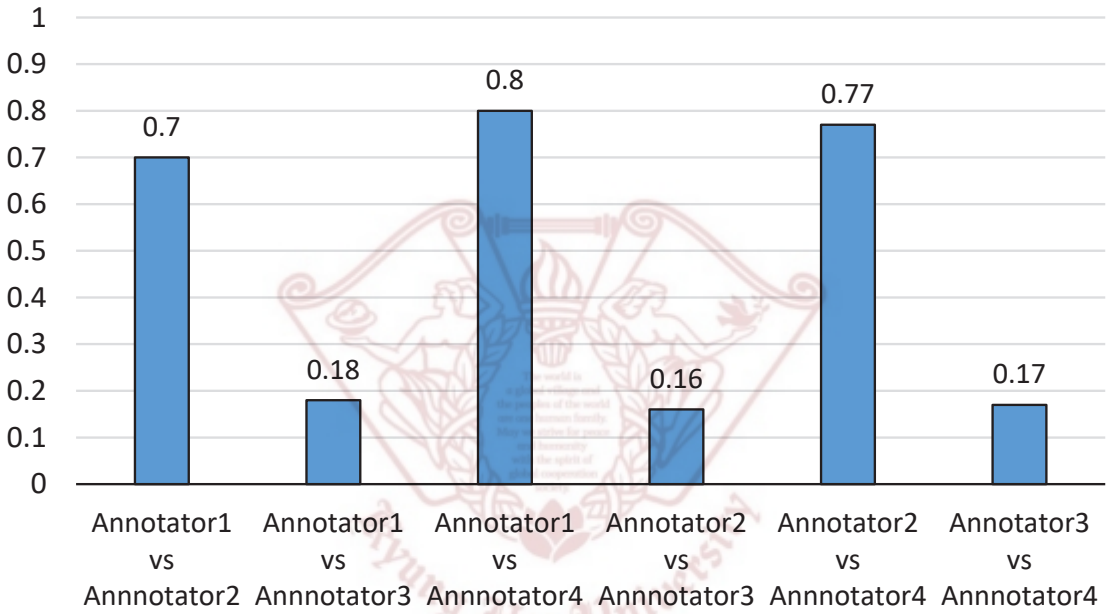

Figure 7.7: Cohen's Kappa (κ) score among the four annotators

Dataset Annotation

In order to compare our computed models with ground truth and to identify the best thresholds for classifying each instance as "equal", "related", or "unrelated" four human annotators were utilized to anonymously, score the similarity of each pair of attribute names. In order to support this process, we first repurposed one of our generated data matrix by marking all attribute pairs belonging to the same schema with the symbol "-". Following this, the annotators marked each cell corresponding to a pair of attributes (conversely, each attribute pair corresponds to two cells with the positioning of the pair-participants swapped; which is used for clarity and identify correct

Table 7.4: Annotations performed by the four annotators on five medical schema

Method	Total Matches	Marked as Equal	Marked as Related	Marked as Unrelated	Not Marked
Annotator1	40698	238	109	40351	0
Annotator2	40698	241	116	40341	0
Annotator3	40698	260	2103	38182	153
Annotator4	40698	225	62	40400	11

relationships between the attribute on left and attribute on right), by determining the similarity in terms of dissimilar as "0", exactly similar as "1", row attribute as child of column attribute as "$<$", row attribute as a parent of the column attribute as "$>$", and finally, unknown as " ". The data sheets generated after this extensive human effort have been made available for other researchers[9].

These sheets, additionally contain some missing values, which were left out by the annotators but in order to maintain their originality, these values were not filled; instead during our evaluation for these datasets, the missing values were considered as having the score "0". Using κ, we evaluated the inter-rater agreement of these annotations, which have been visualized in Figure 7.7. It can be seen in this plot, that "Annotator3" has very small correlation with the other 3 annotators. This difference can be traced back to the number and type of annotations performed by each annotator, which is shown in Table 7.4. The "Annotator3" has marked 2103 cells as related (one of $>$, $<$, or) and left 153 as empty. Even in the presence of these differences, it is pertinent to include the data for all annotators in order to avoid any bias.

This annotated data was then processed to replace all related entries with "0.5" ($class_{0.5}$) and all "-" with "0" ($class_0$), while the values for similar at "1" ($class_1$) and "0" ($class_0$) for dissimilar were kept the same. This conversion was then used to produce a consolidated dataset of 40,698 attribute pairs using mode scores of all annotators for each cell. We also tested average scores between the annotators, but that would produce scores between "0", "0.5", and "1", greatly increasing the number of classes for classification. Hence the maximum agreement between the annotators maintains the final label values within these three classes, which become easier to evaluate. Additionally, the original dataset and its mode consolidated form is biased in favour of class "0", since most attribute pairs are not related to each other. This dataset is then split into development and testing partitions with a ratio of 70:30. The development partition is used for

[9] https://github.com/desertzebra/EMR-Interoperability/tree/master/Implemenation/Data/Annotated

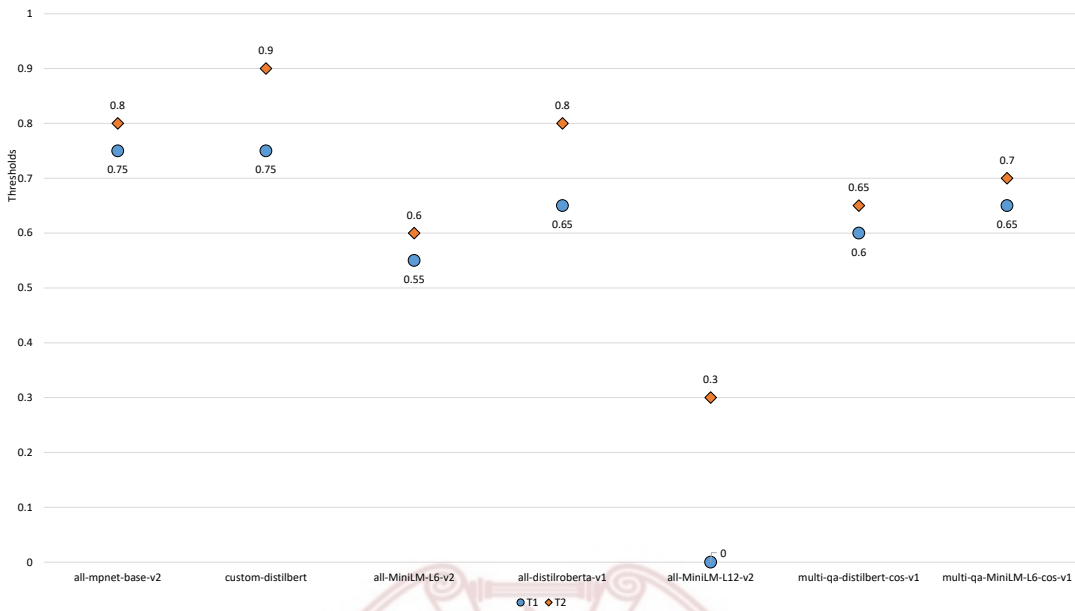

Figure 7.8: Thresholds selection using MCC scores where t1 indicates the similarity threshold between $class_0$ and $class_{0.5}$, and t2 indicate the similarity threshold between $class_{0.5}$ and $class_1$.

threshold selection based on the best MCC score for identifying class "equal", followed by best scores for class "related" and finally best of class "unrelated". The optimal threshold thus achieved is used to classify the instances of the test dataset, which is finally evaluated on its MCC and F_1 measure.

Figure 7.9 shows a heatmap of the semantic similarity between the attributes of the six participating medical schema, as marked by the four annotators. The grey color indicates that no similarity was calculated because the corresponding cell pertains to the same attribute in row and column. Blue color indicates, the two attributes are not similar, yellow indicates some semantic similarity and red indicates the two attributes are equal.

Threshold Selection

A good text classification methodology is dependent on the correct choice of a threshold, which can maximize the target class participation. In case of independent labels, area under the precision recall curve can provide this optimal measure, however as in our case, for dependent classes on a biased dataset the MCC, is better [144]. Since our aim is to apply an optimal text similarity

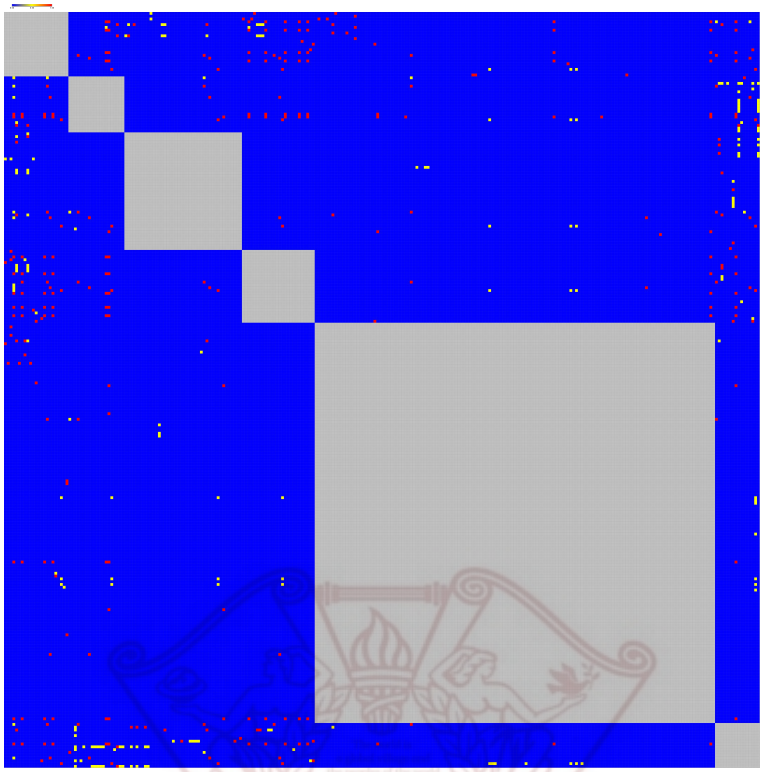

Figure 7.9: A Heat Map showing the semantic similarity between the attributes, as indicated by the (mode of) annotated values.

classifier to resolve this multi-class problem ($class_0$, $class_{0.5}$, and $class_1$), we have to test various threshold scores for separating the instances between $class_0$ and $class_{0.5}$ (t_1), and then $class_{0.5}$ and $class_1$ (t_2). Additionally, since our aim is to correctly identify the similar attribute instances, it is pertinent to maximize the classification performance of $class_1$ (similar), followed by $class_{0.5}$ (related), and finally $class_0$ (unrelated). With a step size of *0.05* ($step$), and starting from t_1 as 0.0 and t_2 as t_1 + $step$, we move the thresholds until t_2 reaches *1.0*, followed by increase in t_1 by step size. Eventually, t_1, reaches *0.95* and t_2 reaches *1.0*, at which point, the process stops. This is to ensure that t_1 remains behind t_2, for all iterations, measuring MCC score, for the 9 models. These models include, "Fuzzy_Wuzzy", "Word2Vec", and 7 transformer based models, specialized for the text semantic similarity measurement task. The optimal thresholds achieved by each of these models is shown in Figure 7.8.

Threshold values for Word2Vec are placed at the lower end of the spectrum indicating a very

Table 7.5: Performance matrix for individual classes using one vs all binarization technique

Model	Class Positive	Class Negative	Accuracy	Precision	Recall	F-1	MCC
FUZZY_MATCH	$class_0$	$class_{0.5}$,$class_1$	0.98	1.00	0.99	0.99	0.37
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.98	0.20	0.59	0.29	0.33
Word2Vec	$class_0$	$class_{0.5}$,$class_1$	0.24	0.99	0.24	0.38	0.01
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.15	0.10	0.12	0.12
	$class_1$	$class_0$,$class_{0.5}$	0.35	0.01	**0.84**	0.01	0.03
bert-base-nli-stsb-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	0.99	0.99	0.37
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.22	0.50	0.30	0.32
bert-large-nli-stsb-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	1.00	1.00	0.51
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.41	0.50	0.45	0.45
roberta-base-nli-stsb-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	0.99	1.00	0.40
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.26	0.50	0.34	0.36
roberta-large-nli-stsb-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	1.00	1.00	1.00	1.00	0.61
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	**1.00**	**0.59**	0.50	**0.54**	**0.54**
distilbert-base-nli-stsb-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	0.99	0.99	0.38
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.23	0.50	0.32	0.33
bert-base-nli-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	0.99	0.99	0.35
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.20	0.50	0.29	0.31
bert-large-nli-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.28	1.00	0.27	0.43	0.05
	$class_{0.5}$	$class_0$,$class_1$	0.99	0.34	0.90	0.49	0.55
	$class_1$	$class_0$,$class_{0.5}$	0.57	0.00	0.50	0.01	0.01
roberta-base-nli-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.96	1.00	0.97	0.98	0.23
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.96	0.08	0.50	0.13	0.19
roberta-large-nli-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	0.99	0.99	0.39
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.20	0.50	0.28	0.31
distilbert-base-nli-mean-tokens	$class_0$	$class_{0.5}$,$class_1$	0.99	1.00	0.99	0.99	0.36
	$class_{0.5}$	$class_0$,$class_1$	1.00	0.00	0.00	0.00	0.00
	$class_1$	$class_0$,$class_{0.5}$	0.99	0.21	0.50	0.30	0.32

large number of instances are classified as similar (above similarity score of *0.1*), while a small number of instances (with similarity score *0.05*) are classified as dissimilar. Similarly, the $class_{0.5}$ lies within the similarity threshold of *0.05* similarity points.

It can be observed that the threshold for selecting the related class is above *0.5* points for most semantic textual similarity models. In case of all-MiniLM-L12-v2, the lower threshold is however at *0.0* while the higher threshold is at *0.3*. This indicates, that this model is not able to identify the similar instances, however for the equal instances and dissimilar instances, that model would produce good results. The custom-DistilBERT-base-uncased model, which has been fine-tuned

on the clinical conversations achieves a lower threshold at *0.75* and higher at *0.9*. Similarly, the all-mpnet-base-v2 pretrained model achieves a lower threshold at *0.75* and higher at *0.8*. Since these thresholds have been computed to maximize the MCC score for $class_1$, followed by a maximization of MCC for $class_{0.5}$, thus a the difference between t_1 and t_2 is relatively small, capturing the lower number of equal and similar instances, as annotated by the human experts. Similarly, in order to capture a large number of dissimilar instances, represented by $class_0$, the lower threshold t_1 is at a high similarity point.

These results show a general trend of how the cosine similarity varies/maintains itself, against embedded vectors generated from various pre-trained models. In absolute terms, however these threshold values provide the mechanism for classifying the test dataset, which is evaluated for performance in the next subsection.

Model Evaluation

On unseen test dataset with thresholds selected in the previous step and the 9 models, we measured the performance score using one vs all binarization of the multi-classes. As evident in Table 7.5, very high values of accuracy are visible across all models with all three positive classes. In all, except the case of Word2Vec, precision and recall also show values close to *1.0*. However, these measures are very misleading, since the dataset is greatly biased in favour of $class_0$.

In terms of F_1 measure $class_{0.5}$ shows the worst possible results, independently, with all except Word2Vec and bert-large-nli-mean-tokens having a score of *0.0*. All-mpnet-base-v2 provides the best F_1 measure at *0.49*. These metrics are thus not useful to gauge the performance of the evaluated models.

Instead, focusing on the MCC score, provides a good picture of the model performance for individual classes when all other instances are negative.

Finally, we evaluated the overall κ coefficient and MCC score to evaluate the performance of each model on the test dataset. These scores range between [-1,1], providing a measure quantifying the accuracy of the classifier to correctly predict correct and incorrect instances.

As shown in Figure 7.10, the models Word2Vec (at 0.02), all-MiniLM-L12-v2 (at 0.06), and the fine-tuned DistilBERT-base-uncased (at 0.09), with κ score between [0,0.20] indicate random

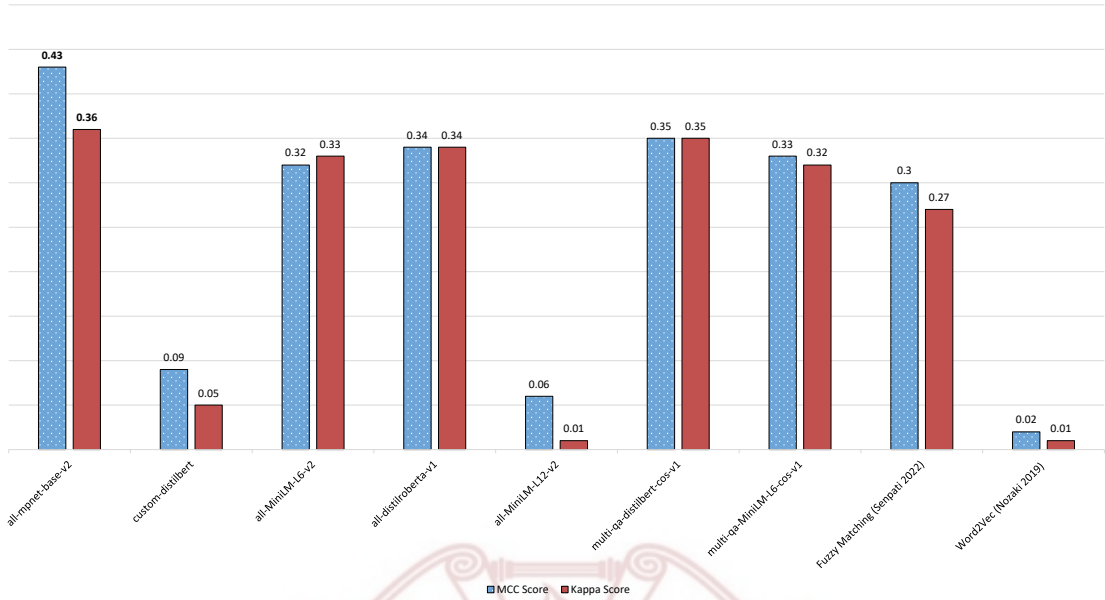

Figure 7.10: Performance evaluation of various models using MCC and kappa (κ) scores.

classification. while κ score between [0.21,0.39], achieved by all-mpnet-base-v2, all-MiniLM-L6-v2, all-distilroberta-v1, multi-qa-distilbert-cos-v1, multi-qa-MiniLM-L6-cos-v1, and Fuzzy Matching show only minimal agreement with the annotated data [161]. This effect is due to the imbalance nature of the dataset. Relatively, the best results, when evaluated in terms of MCC are achieved by all-mpnet-base-v2 pretrained semantic textual similarity model, which indicate a good, balanced agreement between the annotated dataset and the computed one.

Figure 7.11 shows a heatmap of the semantic similarity between the attributes of the six participating medical schema. The grey color indicates that no similarity was calculated because the corresponding cell pertains to the same attribute in row and column. Blue color indicates, the two attributes are not similar, yellow indicates some semantic similarity and red indicates the two attributes are equal.

7.2.3 Semantic Reconciliation-on-Read

The evaluation results are presented in the following sections.

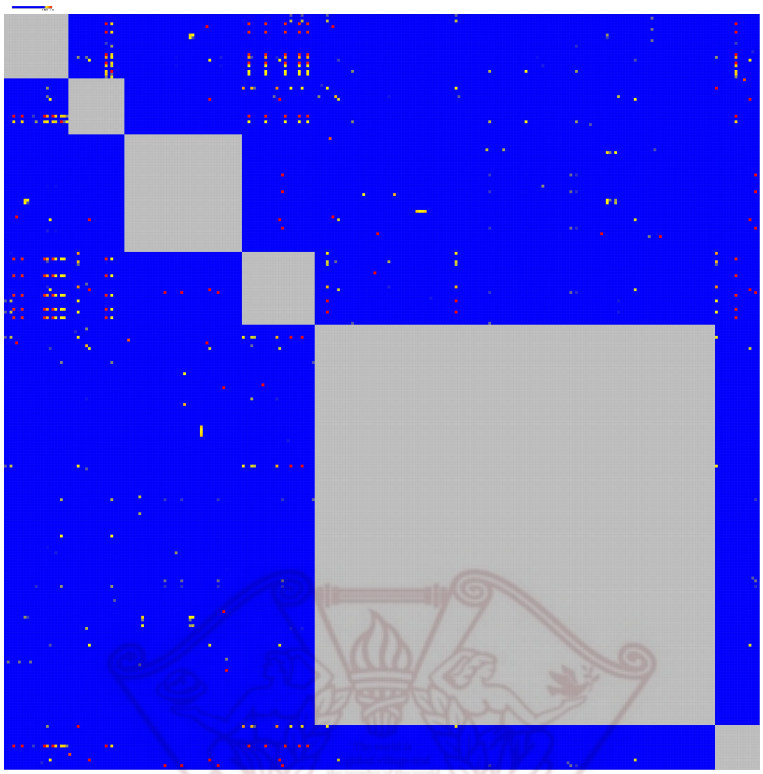

Figure 7.11: A Heat Map showing the semantic similarity between the attributes, computed using all-mpnet-base-v2 pretrained model.

Timeliness

In order to evaluate the timeliness aspect of the SRoR, we analyzed the time taken in each iteration to store the medical fragments and their associated metadata into HDFS. As shown in Fig. 7.12 (a), there is a general increasing trend in the amount of time consumed, in relation with an increase in the amount of records. In iteration 1 and iteration 6, the time consumed by C1 and C3 is almost the same. For iteration 2 there is approximately 200% increase, while in iteration 3, 4, and 5 there is a 300% increase. For C2, in all iterations the difference remains within 0.402 seconds. This variation is explained by the increasing file size involved in each iteration, as shown in Table 7.6. For criteria C4, C5, C6, all six iterations showed similar execution time. This is due to the fact that in creating a table, Hive only performs basic indexing operations, thereby creating a logical schema, which is unaffected by the amount of actual data or files in the system.

Fig. 7.12 (b). Shows this trend, with only one corner case in iteration 1, which is most likely, an

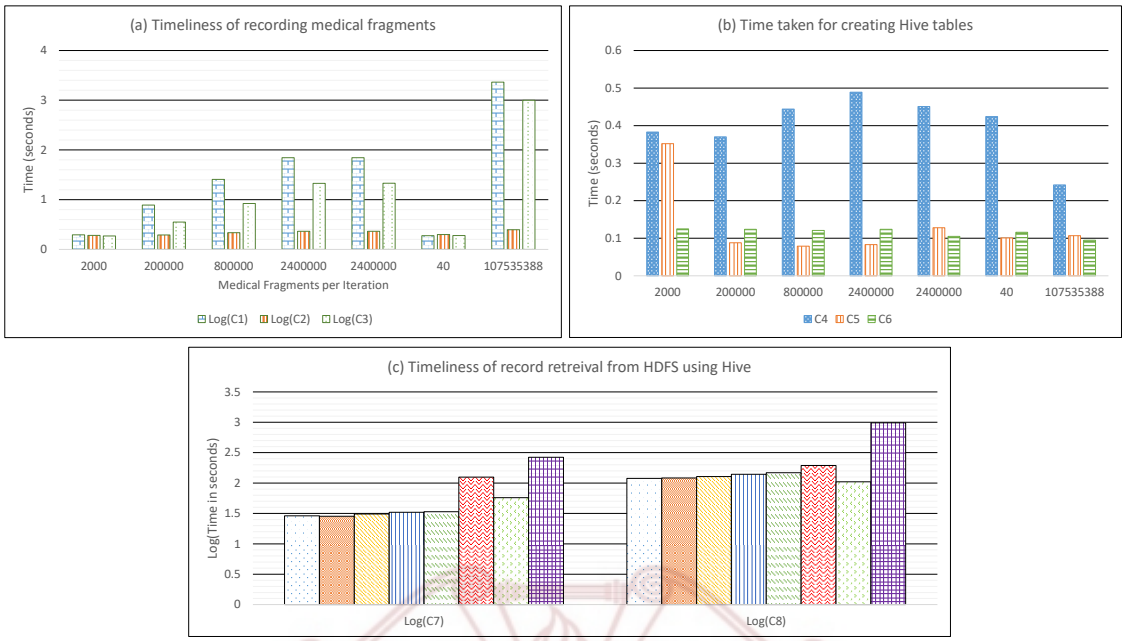

Figure 7.12: (a)Timeliness of recording medical fragments and their metadata in HDFS, (b)Time taken by Hive to create tables [C4, C5, C6], (c) Timeliness of retrieving medical fragments

outlier. Finally, for C7 and C8, we took the average time of 10 queries, as discussed earlier and analyzed the results, which also showed a general increasing trend, till Q4 and Q6, dramatically changed the results. This trend can be explained as a by-product of an unintended optimization. The result of this analysis is shown in Fig. 7.12 (c). Summarizing these results, it is evident that the rate of increase in file size and medical records has a very small impact on the rate variations of C1, C2, C3, C7, and C8. While there is no impact on C4, C5, and C6 criteria. This indicates that the SRoR platform is able maintain timeliness of data storage and retrieval, while also supporting high scalability.

Scalability

As discussed earlier, from our 7 iterations, we have been able to stress test the storage platform, eventually recording over 116 million medical fragments for slightly over 390,000 patients. The storage strategy here, is very important as Hadoop and by extension Hive are really good at processing a small number of large sized files. As shown in Fig. 7.13, the platform is not only able to scale up when adding new patients and their associated medical fragments but has also proved

Table 7.6: HDFS file size comparison for the medical fragments produced in Iteration 1-7

Iteration	Total Medical Fragments	File size for C1 (Kb)	File size for C2 (Kb)	File size for C3 (Kb)
1	2,000	659	6	181
2	200,000	66,260	580	18,059
3	800,000	264,923	2,320	72,242
4	2,400,000	755,295	4,639	216,617
5	2,400,000	755,417	4,639	216,608
6	40	13	1	4
7	116560948	25752400	7263	11118380

successful in scaling the medical fragments of an already existing patient. In particular between iteration 6 and 7, when there was a 14-fold increase in data, only 9-fold increase in querying time was observed.

Accuracy

For our test case of retrieving records of the user “Harry Potter” born on “19880708”, the SRoR has shown 100% accuracy in all 7 iterations, albeit with some adjustment in the 6th iteration. However, even in the case where our particular query returned more results than expected, it did only double up every correct value. This has been explained earlier as a lack of consolidation for the results, which once applied, returned the correct results. Another associated caveat here is the somewhat tightly controlled nature of the sampled data. Even though the data was synthesized (partially based on 52 real patient data), producing over 116 million records, no patient with the same name and data of birth was repeated. However, in real world that may not be the case leading to the challenge of sparse data, which we will discussed in the Section8.6.

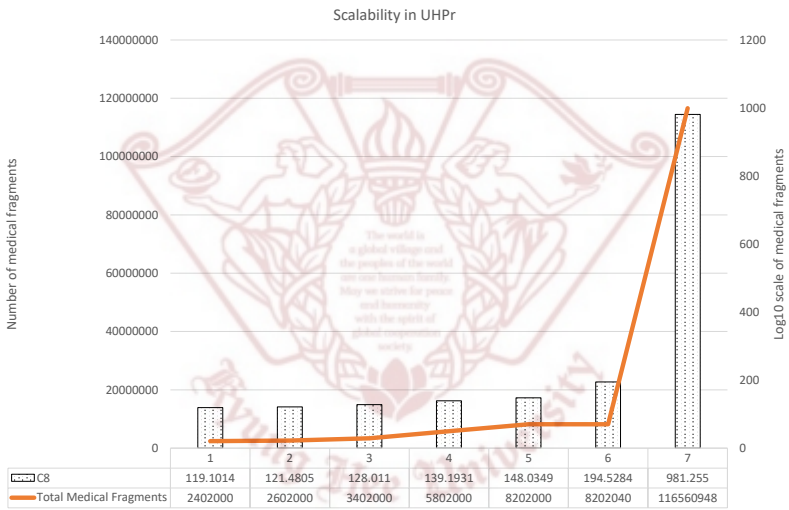

Figure 7.13: Scalability in SRoR

Chapter 8

Discussion

8.1 Sequence Classification its Implications

Identifying structured data, with attribute names and values from the domain of healthcare is a challenging task due to the difficulty in obtaining the source data and its sensitivity. Once the data has been obtained, several operational challenges in processing the text and extracting a structured representation from it, in a timely manner is a nontrivial task.

The results presented in this manuscript indicate two important implications. Firstly, while the pre-trained textual similarity models are able to identify similarity between various, cross-domain texts, it is pertinent to apply domain adaptation before utilizing these models as a part of a domain-specific, solution. In particular, the evaluation of sentence similarity task on the complete STSb dataset, the pre-trained all-mpnet-base-v2 model and the pre-trained DistilBERT-base-uncased model show excellent agreement, in terms of the achieved Pearson Correlation. On the other hand, for the same dataset, and using the fine-tuned DistilBERT-base-uncased model, which was subsequently used in this study, indicates mediocre results. It is also important to note here that the STSb dataset has been collected from news items, image captions, and forum discussions, which produces text from various different domains. The performance of a domain-specific model is bound to be reduced on this dataset, especially when compared to the models trained on a large

Secondly, the classification threshold and semantic similarity score between the instances of MASS and the test sequences is of great importance. The threshold used by the proposed method is at “0.87”, while the one used by the baseline method is “0.49”. These values have been calculated using a dedicated portion of the dataset, with AuROC determining the optimal threshold value for semantic similarity. This increase in performance is due to the fact that with a lower threshold, a larger number of test sequences will be classified, and cause an increase in the num-

ber of actually true instances being found. It will also cause an increase in the number of incorrect results being found, thereby causing an imbalance between the precision and recall evaluations. In the real world, the test sequences are un-labeled and have to be manually verified by a clinician, before they can because a permanent part of a patient's medical record. Thus a balance has to be established between the accuracy and the number of classified instances. Essentially, this is what our proposed methodology achieves over the baseline methodology. With just 837 instances our proposed method achieves an accuracy of 52.96%. However, when the threshold is reduced from "0.87" to "0.6" the accuracy increases by less than 2% and an additional 622 instances are classified, which would nearly double the verification time for the physician. In scenarios, where it is important to capture a wide variety of correct instances and the focus is on improving the performance in terms of recall, a multi-modal approach, such as the one presented in [162] can be used.

These results can be further improved by increasing the volume of MASS and adding more medically aligned sequences and attribute labels into it, adding more and better defined extraction functions, and improving the text segmentation.

8.2 Value Extraction via Patterns vs UMLS

As an example, consider the sequence, "The patient has acute fever". Here, when we split this sequence into unigrams and use each non-stop word token to check the UMLS browser, we find that "Patient" has a semantic type of "Patient or Disabled Group", "acute" has "Temporal Concept ", "fever" has "Sign or Symptom" and "Finding", as determined by UMLS. With a bigram lookup, "acute fever" has a semantic type of "Disease or Syndrome" and "Finding", and "patient acute" has an approximate match with "Finding". These are only some of the concept types associated with the unigram and bigram tokens, provided by UMLS. However, "Finding" is the semantic concept in the label of the trained sequence, matching with this test sequence, we can easily make remote calls to the UMLS API and identify the best matching values, for "Finding" in it.

On the other hand, the sentence "I am 8 years old", is better identified through the use of regular expressions. This is because the semantic concepts associated with the unigrams "8", "years", "old", and bigrams, "8 years", "years old", are not able to identify the value for the

attribute “age”.

8.3 Validation on labeled data vs Verification on unlabeled data

As evident in Section 7.2.1, we obtain very different results on the labeled data (72.94% accuracy) compared with the results on unlabeled data (65.38% correct value identification). However, these results are not comparable due to two reasons. Firstly, in the case of labeled data, our evaluation focused on identifying the performance of sequence classification and key identification only, while in the case of unlabeled data, our aim was to identify the correctness of the value in a timely manner. Hence, even if the value identified in the former case was incorrect, our evaluation identified it as correct. To determine a value as correct, a human expert has to verify it, which for the labeled data part was too tedious.

Secondly, the data used to develop MASS and the labeled sequence set was obtained from DHQ-Kotli, with the physician dealing with pediatric patients in an outpatient department. As a result, all encounters in those conversations, are short and deal with relatively minor problems. On the other hand, the physician at Care+ MC-Islamabad was handling patient guardians in the neo-natal department. As a result, the encounters at Care+ MC-Islamabad are longer and contain many administrative and operational instructions, which are not useful to obtain an appropriate key-value pair for this study.

8.4 Clinical perspective on formalizing the encounters

The clinical adage that about two-thirds of diagnoses can be made on the basis of history alone has retained its validity despite the technological advances of the modern hospital. Once a rapport is built between a physician and a patient it helps boost the self-esteem of ill patients who are already struggling with their illnesses. The correct guidance by the physician is always relieving for the patients but for that, the art of interviewing a patient should be mastered [163]. Objective questioning is a helpful tool in guiding patients and reaching the right diagnosis. It is also very helpful to cut-down unnecessary investigations which are a waste of time and money for the patients. While open-ended questions give the patients and their attendants the opportunity to

explain the symptoms in detail, they often lead to cognitive overload and necessitate continuous note-taking and recording so as not to miss any important detail. Instead, for the physician, it is better to utilize short, targeted questions, and for the patient to provide detailed answers, so that contextual information can be collected [164]. Additionally, by recording these conversations, and extracting a correct summary from them, a lot of time and money can be saved for the medical center, physician, and patient.

8.5 Schema Alignment

In text classification, production and use of a well annotated corpus for supervised and semi-supervised learning is of utmost importance. The same is also useful for evaluating the performance of unsupervised learning techniques. In the real world, the production and maintenance of these corpora is an expensive task, often requiring extensive human effort and conformance to ethical principles, which can restrict access to critical data for the researchers. While there are many factors, influencing this reality, one of the most critical is the perception and cognitive ability of an expert user to subjectively assign a label to an instance [165, 166]. Data validity is especially important in the domain of healthcare, where the acquisition, curation, and sharing processes are all encapsulated by the need to ensure correctness as well as privacy and security of the user. Consequently, the availability of healthcare data, its accuracy, and transparency are major concerns for most researchers associated with this domain [167]. It is not only important to access the data but also to understand how it was produced, the caveats associated with it, and any assumptions made during or after its acquisition. In the case of our annotated dataset, the instances have been labeled by four human experts (two medical practitioners and two computer science graduates), using their subjective knowledge. One example of this subjective classification, is evident in the raw form of the data instances labeled by “Annotator3”. According to “Annotator3” the relationship between the term “AdmissionId” and “ClinicalHistory” is parent-child. While the selection of this label can be debated from a subjective view point, changing or removing it or any other label, from the (objective) view point of the computing methodology would be incorrect [168]. As a result, the annotations were kept anonymous so as not to induce any bias. Thus, the complete annotated data in its original form became the basis of our threshold selection and model evaluation methodology.

A mode based voting mechanism was then used to resolve the differences between the annotators. The consolidated true dataset was then formed based on the agreed upon label by atleast 3 annotators. As pointed out in [165], the net effect of such a voting mechanism is an increase in the precision of the machine learning classifiers in lieu of, their accuracy. As shown in the results and discussed further on, due to the bias nature of our dataset, accuracy measure is replaced by MCC. The lower scores of agreement between the (annotated data) true labels and predicted labels, have thus been evaluated in a contextual and relative manner.

Throughout this research work, the choice of performance metrics used for threshold selection and model evaluations are also driven by the dataset's nature. Even before annotation, the dataset is bias in favour of unrelated attributes. As established by the human experts and the machine learning model, for 254 attributes involved in 40698 possible pairs only a little over 300 similar instances are found. In these circumstances performance metrics, such as accuracy, precision, and recall are meaningless. These metrics are unable to account for the imbalanced datasets and provide an incorrect view of the classifier's accuracy. Instead metrics such as F_1, MCC, and κ can provide a true picture of the classifier's accuracy. These metrics are also well suited for evaluating multi-class classifiers, using one vs all or one vs one binarizations of the dataset, as well as consolidating the results into a single measure. In our experimentation, we also evaluated the Area Under the Receiver Operating Characteristic (AUROC) curve and Area Under the Precision Recall (AUPR) curve, as a threshold selection metric. These graphs are well suited for independent classes as shown by [169]. Additionally, in our case it is important to maximize the identification of similar attribute pairs ($class_1$), followed by related ones ($class_{0.5}$) and finally the unrelated ones($class_0$). AUROC and AUPR were thus replaced with our current approach for threshold selection. The benefit of using this kind of dependent classification is its usefulness in practice to identify a small set of similar attribute pairs, which can be used to establish positive results.

8.6 Patient Identification

Patient identification number is considered one bottleneck for cross sharing of the patient information among different medical organizations. The proposal of a single identifier across the country to identify patients in every medical organization would be one restricted solution. But the im-

plementation of this strategy worldwide, still needs to be seen. Many covid 19 fatalities that were having underlying medical conditions could have been saved, if patient identification was performed properly across the countries. The problem of unique indexing can be explained by a simple question, raised by one of the reviewers of our work, "What happens when there are two individuals named Harry Potter and born on 19880708?". This is one of the key research in the field of information systems. Also known as the entity resolution problem, in a Big Data environment, this problem is especially important, given the schema less storage and the large volume of items, qualifying as an entity [170].

There are two perspectives of this particular challenge. Firstly, the problem of disambiguation, whereby two different individuals from the real world, must remain so in the digital world as well. Secondly, due to sparse data, we may not always have the complete picture leading to one real world user, having multiple digital profiles. The problem might look trivial with an obvious solution to incorporate some more unique attributes like patient's address, or a hash of the patient's demographics, or an email or a phone. However, for one thing this would lead to a cyclic argument, whereby no amount of extra attributes would be enough for a universal solution. Pattern recognition technique such as the one presented in [171], which performs a similarity analysis, while keeping the computation with-in database can prove to be useful in our setting as well.

8.7 Data Verification

Another challenge towards achieving complete data interoperability is the lack of a comprehensive and easy to use data verification platform. This is partially due to the veracity of medical data. As discussed in the motivation section, it is not possible to expect the over-worked medical experts to provide complete information. Instead a system of incentive based verification along with distributed voting, crowd sourcing or Blockchain could prove to be successful here. Another related aspect of this problem, is verification of semantic matching and semantic integration, in order to provide semantic reconciliation at data, process, and knowledge levels.

Data verification is made more complicated because of the occurrence of duplicate records in the patient index. Similarly, duplicates can occur when multiple records of the same patient are created in a medical system. This will not provide full medical history to the medical staff, restricting the

quality of care. Confusions can also occur when same ID is provided to multiple patients. This can be very risky as the history of one of the patients should be the combination of the two patients with similar IDs. In addition, inaccuracy of data can be another challenge for data verification. For example, inaccurate data collection at the current or previous registration process at same or different medical organizations.

8.8 Security and Privacy

Due to the very sensitive nature of the healthcare domain, data privacy is a major challenge, which requires implementation of very precise and comprehensive methodologies and policies for preventing any unauthorized access [172]. This includes providing an authentication and authorization procedure, maintaining integrity of the data, keeping the patient records confidential, maintaining availability, and disallowing non-repudiation [173]. Security and privacy is one of the most critical factors for any information system in general and an interoperable one in particular. This involves the questions such as whom to share, how to share, why to share, and how much to share? This also is related with another debate about who is real owner of the data (patient, one of the participating medical organizations, or all of the medical organization).

While ample solutions do exist which can help resolve this problem, identifying and using the one with least impact on the timeliness and scalability is the main concern, here. Additionally, depending upon the abstraction level at which the platform, like SRoR, is deployed, it may be necessary to take into account multiple legislation and organizational policies. e.g. compliance with Health Insurance Portability and Accountability Act of 1996 is required in the US, while in the EU medical record management systems must comply with General Data Protection Regulation 2016/679.

Chapter 9

Conclusion and Future Direction

9.1 Conclusion

Integrated healthcare systems can drastically increase the quantity and quality of healthcare services, available to the general public. Through a combination of wellness and medical interventions, geared towards each patient's unique medical history and by incorporating the best treatment plans from similar interventions in the past, patient-oriented healthcare has now become a reality. Advances in ICT in general, and AI, and Big Data, in particular have led to the creation of several foundational platforms, supporting the Ubiquitous Healthcare.

In reality, however, many technical and operational gaps exist between the tools and technologies, available to the healthcare providers in the developed world, versus those in the developing world. In technical terms, state-of-the-art design principals underlying the creation of standardized HMIS, such as HL7 or OpenEHR based communication standards, SNOMED-CT or LOINC based terminological standards, and many others, have not gained universal acceptance and are severely lacking in many systems, utilized in the developing world. A prevalence of quick and trivial solutions, have led to the development of customized solutions, utilizing adhoc data schema, in some of the large hospitals in the developing world. Operationally, in the developing world, a very large portion of the healthcare services in general and public healthcare in particular, suffers from large patient loads and a lack of resources available to the physicians and other healthcare providers to effectively utilize digital platforms for recording clinical interactions.

The semantic textual similarity task, typically, utilizes the positional semantics of words in a text sequence to determine its context. Using this contextual information, two distinct text sequences can then be encoded into embedding vectors and their similarity can be determined using some distance or similarity metric, such as cosine similarity, manhattan distance or others.

Leveraging this theoretical background and based on the very good empirical results achieved by transformer based models to solve the semantic textual similarity task in the general domain, this dissertation, provides a solution to a real-world problem, in the specific (healthcare) domain. Theoretically, the first two solutions proposed in this dissertation, the Sequence Contraction approach and the Sequence Expansion approach, provide a mechanism to build appropriate text sequences, which can be used for the semantic textual similarity task in the clinical, NLP domain.

Practically, the automated solutions proposed in this dissertation, aim to create an interoperable healthcare environment, which can remove redundancies in the data acquisition process, leading to the provision of rich clinical histories by integrating multi-modal patient data. This approach, resolves the subjective nature of the clinical history enabled by the cognitive limitations of the patients and jump starts standard compliance for the small and mid scale hospitals. In particular, this dissertation looks at three problems, related to these technical and operational gaps, the resolution to which, can provide several benefits to the healthcare community at large, and greatly reduce the stress on healthcare resources. Firstly, we propose a semantic similarity based approach to automate the process of collecting relevant medical artifacts from unstructured clinical conversations. This approach, utilizes a classification and reduction based approach, to first reduce the sample space from a plethora of unstructured text to medically aligned sequences and then, further reduces each medically aligned sequence to a medically aligned attribute value pair through the Sequence Contraction process. Secondly, a schema alignment approach, based on unsupervised matching of sequences, obtained from attribute names is presented. Here, the main contribution presented in this dissertation is the transformation of an attribute name to a machine understandable sequence, built from a contextually enriched Suffix Array. Thirdly, the integration of these two methodologies into an end-to-end framework is proposed, which supports Semantic Reconciliation in real-time, while also providing data and schema evolution management.

The empirical results presented in this dissertation provide the foundational backing to an automated standard-agnostic mechanism for achieving Data Interoperability. The Sequence Contraction results were obtained on real-world clinical conversation collected from two hospitals in Pakistan. As such, the performance achieved by the proposed fine-tuned DistilBERT-base-uncased model, in terms of its accuracy at 52.96% and precision at 69.34% is higher than other pre-trained

sequence similarity models. The key finding here, is the correctness of the proposed approach, which classifies unseen text sequences, to identify the probable medical sequences, which can then be reduced into attribute-value pairs, using syntactic and contextual semantic approaches. The underlying Machine Learning models can be replaced with other counterparts, which can provide solutions in the general domain or are better adapted, however, the classification and reduction based workflow presented in this dissertation is the minimum viable approach necessary to automate the process of converting unstructured clinical text into (semi) structured medical data.

Similarly, as a result of the Sequence Expansion approach, the conversion of attribute names based on Suffix Arrays, and inclusion of contextual semantics, produces text sequences appropriately configured for unsupervised matching task. Here, the pre-trained allmpnet-base-v2 model achieves an MCC score of 0.43 and a Kappa score of 0.36. Significantly, these results indicate that for a set of 270 attribute names, obtained from 6 medical schema, when the comparison between disjoint attributes is made, the results obtained by the proposed computed method is identical to those produced by expert labeling.

Finally, these two solutions, are integrated into an end-to-end framework, which collects medical data from various structured and unstructured sources (Sequence Contraction operates on the later), and manages the creation, storage, and evolution of Schema Maps (using the Sequence Expansion approach). Based on the curation design of Big Data, the proposed approach from solution 3, provides Semantic Reconciliation of source medical data, using the latest available Schema Map between the source and target schema, to build consumable medical data, conforming to a target schema, temporarily. Such a temporary conversion, ensures that the original medical data remains near to its original form, and as the schema design and/or the Schema Maps change, there is no negative impact on the medical data. Scalability analysis of the proposed framework, implemented using Hadoop engine and Hive based data retrieval layer, indicates that relevant patient medical data can be retrieved in an error free and timely manner, even with millions of medical records. In particular, 116 million medical records are processed in 981 seconds (16 minutes), which is much better than the traditional relational data engine based approach. Additionally, the NoSQL based data engine proposed here, is well suited for providing a single storage point for a variety of data schema, which can be processed with MapReduce or Graph processing to support

standard-agnostic Data Interoperability.

9.2 Future Direction

The solutions proposed in this dissertation provide theoretical foundations, backed by empirical results, to design an end-to-end framework supporting medical Data Interoperability and by extension, Ubiquitous Healthcare. These solutions provide a proof-of-concept for using transfer learning and re-purposing generic models prepared for textual semantic similarity task, for classification of medically aligned sequences. As discussed earlier, the ML models, used and proposed in this dissertation, act as a placeholder for the semantic similarity models, and as such, these can be replaced with similar transformer based models, such as RoBERTa or GPT 2/3.

Additionally, an increase in the amount of real-world data would be beneficial to enrich MASS, further fine-tune the ML model, and identify a variety of new attributes and their extraction methodology. Currently, we have used basic regular expressions to extract syntactic text artifacts corresponding to the value of identified attributes, and UMLS based conceptual semantics to identify text artifacts similar to the identified attributes. This mechanism can be updated to automatically create regular expressions using state-of-the-art ML models, such as the multilingual BLOOM. Similarly, the conceptual elements can be extended to capture a larger variety of semantic concepts from UMLS.

For Schema Alignment, an increase in number of participating medical systems, can further enhance the interoperability of the existing medical fragments, due to an increase in the number of many-to-many alignments.

Finally, many of the steps, utilized in the proposed solutions can be automated to reduce the number of manual steps, required to prepare the data, and ensure the produced information's validation.

Bibliography

[1] M. L. Kiah, A. Haiqi, B. B. Zaidan, and A. A. Zaidan, "Open source emr software: Profiling, insights and hands-on analysis," *Computer Methods and Programs in Biomedicine*, vol. 117, no. 2, pp. 360–382, 2014.

[2] K. Schwarze, J. Buchanan, J. C. Taylor, and S. Wordsworth, "Are whole-exome and whole-genome sequencing approaches cost-effective? a systematic review of the literature," *GENETICS in MEDICINE*, vol. 00, no. August 2017, 2018.

[3] B. Mesko, "The role of artificial intelligence in precision medicine," *Expert Review of Precision Medicine and Drug Development*, vol. 2, no. 5, pp. 239–241, 2017.

[4] G. Cappon, G. Acciaroli, M. Vettoretti, A. Facchinetti, and G. Sparacino, "Wearable continuous glucose monitoring sensors: a revolution in diabetes treatment," *Electronics*, vol. 6, no. 3, p. 65, 2017.

[5] T. N. Gia, M. Ali, I. B. Dhaou, A. M. Rahmani, T. Westerlund, P. Liljeberg, and H. Tenhunen, "Iot-based continuous glucose monitoring system: A feasibility study," *Procedia Computer Science*, vol. 109, pp. 327–334, 2017.

[6] J. Lahtiranta, "Mediator – enabler for successful digital health care," *Finnish Journal of eHealth and eWelfare*, vol. 9, no. 4 SE - Scientific articles, 11 2017.

[7] T. Shaw, M. Hines, and C. Kielly-Carroll, "Impact of digital health on the safety and quality of health care," 2018.

[8] W. A. Khan, M. Hussain, K. Latif, M. Afzal, F. Ahmad, and S. Lee, "Process interoperability in healthcare systems with dynamic semantic web services," *Computing*, vol. 95, no. 9, pp. 837–862, 2013.

[9] M. G. Kahn, J. S. Brown, A. T. Chun, B. N. Davidson, D. Meeker, P. B. Ryan, L. M. Schilling, N. G. Weiskopf, A. E. Williams, and M. N. Zozus, "Transparent reporting of data quality in distributed data networks," *Egems*, vol. 3, no. 1, 2015.

[10] L. Samal, P. C. Dykes, J. O. Greenberg, O. Hasan, A. K. Venkatesh, L. A. Volk, and D. W. Bates, "Care coordination gaps due to lack of interoperability in the united states : a qualitative study and literature review," *BMC Health Services Research*, pp. 1–8, 2016.

[11] S. young Jung, K. Lee, and H. Hwang, "Recent trends of healthcare information and communication technologies in pediatrics: a systematic review," *Clinical and Experimental Pediatrics*, vol. 65, no. 6, p. 291, 2022.

[12] I. T. Union, *Global Connectivity Report 2022*. International Telecommunication Union, 2022.

[13] HL7, "Health Level 7 Version 3 (HL7v3) Product Suite," 2017. [Online]. Available: https://www.hl7.org/implement/standards/product{_}brief.cfm?product{_}id=186

[14] SNOMED, "SNOMED Clinical Terminologies," 01 2020. [Online]. Available: http://www.snomed.org/snomed-ct/five-step-briefing

[15] LOINC, "Learn LOINC," 2018. [Online]. Available: https://loinc.org/learn/

[16] CIMI, 2015. [Online]. Available: http://www.opencimi.org/

[17] S. Kempe and D. Booth, "SmartData Webinar: Yosemite Project for Healthcare Information Interoperability," 2015. [Online]. Available: https://www.dataversity.net/smartdata-webinar-the-yosemite-project-for-healthcare-information-interoperability/

[18] A. Geraci, F. Katki, L. McMonegal, B. Meyer, J. Lane, P. Wilson, J. Radatz, M. Yee, H. Porteous, and F. Springsteel, *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*. IEEE Press, 1991.

[19] R. Berryman, N. Yost, N. Dunn, and C. Edwards, "Data interoperability and information security in healthcare," *Transactions of the International Conference on Health Information Technology Advancement*, vol. 26, 2013.

[20] World Health organization, "Global strategy on digital health 2020-2025," 2021. [Online]. Available: https://apps.who.int/iris/bitstream/handle/10665/344249/9789240020924-eng.pdf

[21] H. Beks, O. King, R. Clapham, L. Alston, K. Glenister, C. McKinstry, C. Quilliam, I. Wellwood, C. Williams, A. W. Shee *et al.*, "Community health programs delivered through information and communications technology in high-income countries: Scoping review," *Journal of Medical Internet Research*, vol. 24, no. 3, p. e26515, 2022.

[22] G. Coppersmith, "Digital life data in the clinical whitespace," *Current Directions in Psychological Science*, vol. 31, no. 1, pp. 34–40, 2022.

[23] B. Kaplan, "Revisiting health information technology ethical, legal, and social issues and evaluation: telehealth/telemedicine and covid-19," *International journal of medical informatics*, vol. 143, p. 104239, 2020.

[24] D. Furtado, A. F. Gygax, C. A. Chan, and A. I. Bush, "Time to forge ahead: The internet of things for healthcare," *Digital Communications and Networks*, 2022.

[25] M. Chandra, K. Kumar, P. Thakur, S. Chattopadhyaya, F. Alam, and S. Kumar, "Digital technologies, healthcare and covid-19: insights from developing and emerging nations," *Health and Technology*, pp. 1–22, 2022.

[26] R. Rahman and C. K. Reddy, "Electronic health records: A survey." *Healthcare Data Analytics*, vol. 36, p. 21, 2015.

[27] O. Müller, I. Junglas, S. Debortoli, and J. vom Brocke, "Using text analytics to derive customer service management benefits from unstructured data," *MIS Quarterly Executive*, vol. 15, no. 4, pp. 243–258, 2016.

[28] B. Percha, "Modern clinical text mining: A guide and review," *Annual Review of Biomedical Data Science*, vol. 4, pp. 165–187, 2021.

[29] K. S. Jones, "Natural language processing: a historical review," *Current issues in computational linguistics: in honour of Don Walker*, pp. 3–16, 1994.

[30] D. Khurana, A. Koli, K. Khatter, and S. Singh, "Natural language processing: State of the art, current trends and challenges," *arXiv preprint arXiv:1708.05148*, 2017.

[31] H. Dalianis, *Clinical text mining: Secondary use of electronic patient records*. Springer Nature, 2018.

[32] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual keyword extraction and summarization: State-of-the-art," *Information Processing & Management*, vol. 56, no. 6, p. 102088, 2019.

[33] D. Putthividhya and J. Hu, "Bootstrapped named entity recognition for product attribute extraction," in *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 2011, pp. 1557–1567.

[34] H. Yan, T. Gui, J. Dai, Q. Guo, Z. Zhang, and X. Qiu, "A unified generative framework for various ner subtasks," 2021, pp. 5808–5822. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.451

[35] Q. Wang, L. Yang, B. Kanagal, S. Sanghai, D. Sivakumar, B. Shu, Z. Yu, and J. Elsas, "Learning to extract attribute value from product via question answering: A multi-task approach," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 47–55.

[36] H. Xu, W. Wang, X. Mao, X. Jiang, and M. Lan, "Scaling up open tagging from tens to thousands: Comprehension empowered attribute value extraction from product title," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 5214–5223.

[37] G. Zheng, S. Mukherjee, X. L. Dong, and F. Li, "Opentag: Open attribute value extraction from product profiles," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1049–1058.

[38] R. Ghani, K. Probst, Y. Liu, M. Krema, and A. Fano, "Text mining for product attribute extraction," *ACM SIGKDD Explorations Newsletter*, vol. 8, no. 1, pp. 41–48, 2006.

[39] K. Roy, P. Goyal, and M. Pandey, "Attribute value generation from product title using language models," in *Proceedings of The 4th Workshop on e-Commerce and NLP*, 2021, pp. 13–17.

[40] L. Yang, Q. Wang, Z. Yu, A. Kulkarni, S. Sanghai, B. Shu, J. Elsas, and B. Kanagal, "Mave: A product dataset for multi-source attribute value extraction," in *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, 2022, pp. 1256–1265.

[41] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan, "Domain adaptation of rule-based annotators for named-entity recognition tasks," in *Proceedings of the 2010 conference on empirical methods in natural language processing*, 2010, pp. 1002–1012.

[42] J. P. Chiu and E. Nichols, "Named entity recognition with bidirectional lstm-cnns," *Transactions of the association for computational linguistics*, vol. 4, pp. 357–370, 2016.

[43] D. Vandic, J.-W. Van Dam, and F. Frasincar, "Faceted product search powered by the semantic web," *Decision Support Systems*, vol. 53, no. 3, pp. 425–437, 2012.

[44] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification," *Lingvisticae Investigationes*, vol. 30, no. 1, pp. 3–26, 2007.

[45] A. Mikheev, M. Moens, and C. Grover, "Named entity recognition without gazetteers," in *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, 1999, pp. 1–8.

[46] E. Aramaki, Y. Miura, M. Tonoike, T. Ohkuma, H. Masuichi, and K. Ohe, "Text2table: Medical text summarization system based on named entity recognition and modality identification," in *Proceedings of the BioNLP 2009 Workshop*, 2009, pp. 185–192.

[47] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of machine learning research*, vol. 12, no. ARTICLE, pp. 2493–2537, 2011.

[48] X. Ling and D. S. Weld, "Fine-grained entity recognition," in *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

[49] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, and H. Liu, "A comparison of word embeddings for the biomedical natural language processing," *Journal of biomedical informatics*, vol. 87, pp. 12–20, 2018.

[50] A. Narayanan, A. Rao, A. Prasad, and B. Das, "Character level neural architectures for boosting named entity recognition in code mixed tweets," in *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*. IEEE, 2020, pp. 1–6.

[51] D. Zeng, C. Sun, L. Lin, and B. Liu, "Lstm-crf for drug-named entity recognition," *Entropy*, vol. 19, no. 6, p. 283, 2017.

[52] J. Yan, N. Zalmout, Y. Liang, C. Grant, X. Ren, and X. L. Dong, "Adatag: Multi-attribute value extraction from product profiles with adaptive decoding," in *ACL/IJCNLP (1)*, 2021.

[53] K. Mehta, I. Oprea, and N. Rasiwasia, "Latex-numeric: Language agnostic text attribute extraction for numeric attributes," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, 2021, pp. 272–279.

[54] Y. Si, J. Wang, H. Xu, and K. Roberts, "Enhancing clinical concept extraction with contextual embeddings," *Journal of the American Medical Informatics Association*, vol. 26, no. 11, pp. 1297–1304, 2019.

[55] R. Chew, M. Wenger, J. Guillory, J. Nonnemaker, A. Kim *et al.*, "Identifying electronic nicotine delivery system brands and flavors on instagram: Natural language processing analysis," *Journal of medical Internet research*, vol. 24, no. 1, p. e30257, 2022.

[56] M. Du, W. Wang, S. Wang, and B. Xu, "A unified framework for attribute extraction in electronic medical records," in *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*, 2020, pp. 1–7.

[57] C. Sun, Z. Yang, L. Wang, Y. Zhang, H. Lin, and J. Wang, "Biomedical named entity recognition using bert in the machine reading comprehension framework," *Journal of Biomedical Informatics*, vol. 118, p. 103799, 2021.

[58] S. Althubaiti, Ş. Kafkas, M. Abdelhakim, and R. Hoehndorf, "Combining lexical and context features for automatic ontology extension," *Journal of biomedical semantics*, vol. 11, no. 1, pp. 1–13, 2020.

[59] K. Nozaki, T. Hochin, and H. Nomiya, "Semantic schema matching for string attribute with word vectors," in *2019 6th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*. IEEE, 2019, pp. 25–30.

[60] A. Yousfi, M. H. El Yazidi, and A. Zellou, "xmatcher: Matching extensible markup language schemas using semantic-based techniques," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 8, pp. 655–665, 2020.

[61] L. Bulygin, "Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem," in *Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains"(DAMDID/RCDL'2018)*, 2018, pp. 245–249.

[62] G. H. Martono and S. Azhari, "Review implementation of linguistic approach in schema matching," *International Journal of Advances in Intelligent Informatics*, vol. 3, no. 1, pp. 1–9, 2017.

[63] A. A. Alwan, A. Nordin, M. Alzeber, and A. Z. Abualkishik, "A survey of schema matching research using database schemas and instances," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, 2017.

[64] M. G. Kersloot, F. J. van Putten, A. Abu-Hanna, R. Cornet, and D. L. Arts, "Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies," *Journal of biomedical semantics*, vol. 11, no. 1, pp. 1–21, 2020.

[65] L. Xu and D. W. Embley, "Discovering direct and indirect matches for schema elements," in *Eighth International Conference on Database Systems for Advanced Applications, 2003.(DASFAA 2003). Proceedings.* IEEE, 2003, pp. 39–46.

[66] D. Zikos and N. Delellis, "CDSS-RM: A clinical decision support system reference model," *BMC Medical Research Methodology*, vol. 18, no. 1, pp. 1–14, 2018.

[67] M. Z. Ercan and M. Lane, "Evaluation of nosql databases for ehr systems," *25th Australasian Conference on Information Systems*, p. 10, 2014.

[68] W. L. Schulz, B. G. Nelson, D. K. Felker, T. J. Durant, and R. Torres, "Evaluation of relational and nosql database architectures to manage genomic annotations," *Journal of Biomedical Informatics*, vol. 64, pp. 288–295, 2016.

[69] R. Sánchez-De-Madariaga, A. Muñoz, R. Lozano-Rubí, P. Serrano-Balazote, A. L. Castro, O. Moreno, and M. Pascual, "Examining database persistence of iso/en 13606 standardized electronic health record extracts: Relational vs. nosql approaches," *BMC Medical Informatics and Decision Making*, vol. 17, no. 1, pp. 1–14, 2017.

[70] A. Celesti, M. Fazio, A. Romano, A. Bramanti, P. Bramanti, and M. Villari, "An oais-based hospital information system on the cloud: Analysis of a nosql column-oriented approach," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, no. 3, pp. 1–7, 2018.

[71] I. Balaur, M. Saqi, A. Barat, A. Lysenko, A. Mazein, C. J. Rawlings, H. J. Ruskin, and C. Auffray, "Epigenet: A graph database of interdependencies between genetic and epige-

netic events in colorectal cancer," *Journal of Computational Biology*, vol. 24, no. 10, pp. 969–980, 2017.

[72] N. Henke, J. Bughin, M. Chui, J. Manyika, T. Saleh, B. Wiseman, and G. Sethupathy, "The age of analytics: Competing in a data-driven world," *McKinsey Global Institute*, vol. 4, 2016.

[73] Ishwarappa and J. Anuradha, "A brief introduction on big data 5vs characteristics and hadoop technology," *Procedia Computer Science*, vol. 48, no. C, pp. 319–324, 2015.

[74] R. E. Gliklich, N. A. Dreyer, M. B. Leavy *et al.*, *Registries for evaluating patient outcomes: a user's guide*. Government Printing Office, 2014, no. 13.

[75] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, "Uk biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age," *PLoS Medicine*, vol. 12, no. 3, pp. 1–10, 2015.

[76] D. R. Blair, C. S. Lyttle, J. M. Mortensen, C. F. Bearden, A. B. Jensen, H. Khiabanian, R. Melamed, R. Rabadan, V. E. Bernstam, S. Brunak, L. J. Jensen, D. Nicolae, N. H. Shah, R. L. Grossman, N. J. Cox, K. P. White, and A. Rzhetsky, "A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk," *Cell*, vol. 155, no. 1, pp. 70–80, 2013.

[77] G. De Moor, M. Sundgren, D. Kalra, A. Schmidt, M. Dugas, B. Claerhout, T. Karakoyun, C. Ohmann, P. Y. Lastic, N. Ammour, R. Kush, D. Dupont, M. Cuggia, C. Daniel, G. Thienpont, and P. Coorevits, "Using electronic health records for clinical research: The case of the ehr4cr project," *Journal of Biomedical Informatics*, vol. 53, pp. 162–173, 2015.

[78] H. H. Nguyen, F. Mirza, M. A. Naeem, and M. Nguyen, "A review on iot healthcare monitoring applications and a vision for transforming sensor data into real-time clinical feedback," *Proceedings of the 2017 IEEE 21st International Conference on Computer Supported Cooperative Work in Design, CSCWD 2017*, pp. 257–262, 2017.

[79] N. B. Krishnan, S. S. S. Sai, and S. B. Mohanthy, "Real time internet application with distributed flow environment for medical iot," *Proceedings of the 2015 International Conference on Green Computing and Internet of Things, ICGCIoT 2015*, pp. 832–837, 2016.

[80] A. Geissbuhler, M. Kimura, C. A. Kulikowski, P. J. Murray, L. Ohno-Machado, H. A. Park, and R. Haux, "Confluence of disciplines in health informatics: An international perspective," *Methods of Information in Medicine*, vol. 50, no. 6, pp. 545–555, 2011.

[81] D. Khan, "Efficient semantic reconciliation for data interoperability among heterogeneous healthcare systems," Ph.D. dissertation, Department of Computer Engineering, Kyung Hee University, South Korea, 2015.

[82] G. Fanjiang, J. H. Grossman, W. D. Compton, P. P. Reid *et al.*, *Building a Better Delivery System: A New Engineering/Health Care Partnership*. National Academies Press, 2005.

[83] S. C. Denaxas, J. George, E. Herrett, A. D. Shah, D. Kalra, A. D. Hingorani, M. Kivimaki, A. D. Timmis, L. Smeeth, and H. Hemingway, "Data resource profile: Cardiovascular disease research using linked bespoke studies and electronic health records (caliber)," *International Journal of Epidemiology*, vol. 41, no. 6, pp. 1625–1638, 2012.

[84] H. Hemingway, G. S. Feder, N. K. Fitzpatrick, S. Denaxas, A. D. Shah, and A. D. Timmis, "Conclusions and implications for clinical practice and further research," in *Using nationwide 'big data'from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the ClinicAl disease research using LInked Bespoke studies and Electronic health Records (CALIBER) programme*. NIHR Journals Library, 2017.

[85] d. boyd and K. Crawford, "Six provocations for big data," *SSRN Electronic Journal*, vol. 123, 09 2011.

[86] N. Askham, D. Cook, M. Doyle, H. Fereday, M. Gibson, U. Landbeck, R. Lee, C. Maynard, G. Palmer, and J. Schwarzenbach, "The six primary dimensions for data quality assessment defining data quality dimensions," *DAMA UK Working Group*, 2013.

[87] M. C. Sanchez-Gomez, K. Dundon, and X. Deng, "Evaluating Data Quality of Newborn Hearing Screening." *Journal of early hearing detection and intervention*, vol. 4, no. 3, pp. 26–32, 2019. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31911952https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6945814/

[88] H. Hemingway, G. S. Feder, N. K. Fitzpatrick, S. Denaxas, A. D. Shah, and A. D. Timmis, "Using nationwide 'big data' from linked electronic health records to help improve outcomes in cardiovascular diseases: 33 studies using methods from epidemiology, informatics, economics and social science in the clinical disease research using linked besp," *Programme Grants for Applied Research*, vol. 5, no. 4, pp. 1–330, 2017.

[89] A. Geraci, F. Katki, L. McMonegal, B. Meyer, J. Lane, P. Wilson, J. Radatz, M. Yee, H. Porteous, and F. Springsteel, *IEEE standard computer dictionary: Compilation of IEEE standard computer glossaries*. IEEE Press, 1991.

[90] M. Clarke, J. De Folter, V. Verma, and H. Gokalp, "Interoperable end-to-end remote patient monitoring platform based on ieee 11073 phd and zigbee health care profile," *IEEE Transactions on Biomedical Engineering*, vol. 65, no. 5, pp. 1014–1025, 2018.

[91] H. Information and M. S. Society, "Definition of interoperability," p. 2013, 2013. [Online]. Available: http://www.himss.org/library/interoperability-standards/what-is

[92] P. Coorevits, M. Sundgren, G. O. Klein, A. Bahr, B. Claerhout, C. Daniel, M. Dugas, D. Dupont, A. Schmidt, P. Singleton, G. De Moor, and D. Kalra, "Electronic health records: New opportunities for clinical research," *Journal of Internal Medicine*, vol. 274, no. 6, pp. 547–560, 2013.

[93] J. Li, "A service-oriented approach to interoperable and secure personal health record systems," *Proceedings - 11th IEEE International Symposium on Service-Oriented System Engineering, SOSE 2017*, pp. 38–46, 2017.

[94] P. Pagano, L. Candela, and D. Castelli, "Data interoperability," *Data Science Journal*, vol. 12, no. 0, pp. GRDI19–GRDI25, 2013.

[95] S. A. Renner, J. G. Scarano, and A. S. Rosenthal, "Data interoperability: Standardization or mediation," *1st IEEE metadata conference*, pp. 1–8, 1996.

[96] C. Bizer and T. Heath, "Linked data - the story so far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009.

[97] F. Song, G. Zacharewicz, and D. Chen, "An ontology-driven framework towards building enterprise semantic information layer," *Advanced Engineering Informatics*, vol. 27, no. 1, pp. 38–50, 2013.

[98] P. R. da Silva and Ferreira, "Enabling agents to retrieve openehr-based health data through implementing hl7 communication with departmental information systems," 2012.

[99] OAEI, "Ontology Alignment Evaluation Initiative (OAEI)," 2020. [Online]. Available: http://oaei.ontologymatching.org/

[100] F. Pentaris, Y. Ioannidis, and I. Manifold, "Interoperability via mapping objects," *Proceedings of the Third DELOS Network of Excellence Workshop on Interoperability and Mediation in Heterogeneous Digital Libraries*, pp. 1–5, 2001.

[101] M. Vujasinovic, N. Ivezic, B. Kulvatunyou, E. Barkmeyer, M. Missikoff, F. Taglino, Z. Marjanovic, and I. Miletic, "Semantic mediation for standard-based b2b interoperability," *IEEE Internet Computing*, vol. 14, no. 1, pp. 52–63, 2010.

[102] F. Benaben, W. Mu, N. Boissel-Dallier, A. M. Barthe-Delanoe, S. Zribi, and H. Pingaud, "Supporting interoperability of collaborative networks through engineering of a service-based mediation information system (mise 2.0)," *Enterprise Information Systems*, vol. 9, pp. 556–582, 2015.

[103] (VeraTech for Health), "LinkEHR," 2019. [Online]. Available: https://linkehr.veratech.es/research.html

[104] HL7, "Health Level 7 Clinical Document Architecture (HL7 CDA)," 2010. [Online]. Available: https://www.hl7.org/implement/standards/product{_}brief.cfm?product{_}id=7

[105] J. A. Maldonado, D. Moner, D. Boscá, J. T. Fernández-Breis, C. Angulo, and M. Robles, "Linkehr-ed: A multi-reference model archetype editor based on formal semantics," *International Journal of Medical Informatics*, vol. 78, no. 8, pp. 559–570, 2009.

[106] C. Martínez Costa, M. Menárguez-Tortosa, and J. T. Fernández-Breis, "Clinical data interoperability based on archetype transformation," *Journal of Biomedical Informatics*, vol. 44, no. 5, pp. 869–880, 2011.

[107] J. A. Maldonado, C. M. Costa, D. Moner, M. Menárguez-Tortosa, D. Boscá, J. A. Miñarro Giménez, J. T. Fernández-Breis, and M. Robles, "Using the researchehr platform to facilitate the practical application of the ehr standards," *Journal of Biomedical Informatics*, vol. 45, no. 4, pp. 746–762, 2012.

[108] M. Marcos, J. A. Maldonado, B. Martínez-Salvador, D. Boscá, and M. Robles, "Interoperability of clinical decision-support systems and electronic health records using archetypes: A case study in clinical trial eligibility," *Journal of Biomedical Informatics*, vol. 46, no. 4, pp. 676–689, 2013.

[109] S. P. Gardner, "Ontologies and semantic data integration," *Drug Discovery Today*, vol. 10, no. 14, pp. 1001–1007, 2005.

[110] W3C - RDFCore Working Group, "RDF," 2014. [Online]. Available: https://www.w3.org/RDF/

[111] W3C - OWL Working Group, "OWL," 2012. [Online]. Available: https://www.w3.org/2001/sw/wiki/OWL

[112] W3C, "SPARQL," 2013. [Online]. Available: https://www.w3.org/TR/sparql11-overview/

[113] H. Zhang, Y. Guo, Q. Li, T. J. George, E. Shenkman, F. Modave, and J. Bian, "An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival," *BMC Medical Informatics and Decision Making*, vol. 18, no. Suppl 2, 2018.

[114] N. I. of Health(NIH), "National Cancer Institute(NCI) Thesaurus," 2020. [Online]. Available: https://ncithesaurus.nci.nih.gov/ncitbrowser/

[115] D. Calvanese, B. Cogrel, S. Komla-Ebri, R. Kontchakov, D. Lanti, M. Rezk, M. Rodriguez-Muro, G. Xiao, and Ó. Corcho, "Ontop: Answering SPARQL Queries over Relational Databases," *Semantic Web*, vol. 8, no. 3, pp. 471–487, 2016.

[116] D. Meridou, C. Patrikakis, A. Kapsalis, I. Venieris, P. Kasnesis, and D.-T. Kaklamani, "An event-driven health service bus," *MOBIHEALTH 2015 - 5th EAI International Conference on Wireless Mobile Communication and Healthcare - Transforming Healthcare through Innovations in Mobile and Wireless Technologies*, 2015.

[117] W3C, "Web Services Description Language (WSDL)," 2001. [Online]. Available: https://www.w3.org/TR/wsdl.html

[118] ——, "Simple Object Access Protocol (SOAP)," 2007. [Online]. Available: https://www.w3.org/TR/soap12/

[119] IBM, "IBM Integration Bus Healthcare Pack," 2015. [Online]. Available: https://www-01.ibm.com/common/ssi/cgi-bin/ssialias?infotype=an{&}subtype=ca{&}appname=gpateam{&}supplier=897{&}letternum=ENUS215-165

[120] HL7, "Fast Healthcare Interoperability Resources (FHIR)," 2019. [Online]. Available: https://www.hl7.org/fhir/overview.html

[121] A. Ryan and P. Eklund, "The health service bus: An architecture and case study in achieving interoperability in healthcare," *Studies in Health Technology and Informatics*, vol. 160, no. PART 1, pp. 922–926, 2010.

[122] MuleSoft, "Mule ESB," 2020. [Online]. Available: https://www.mulesoft.com/platform/soa/mule-esb-open-source-esb

[123] HL7, "Health Level 7 Version 2 (HL7v2) Product Suite," 2011. [Online]. Available: https://www.hl7.org/implement/standards/product{_}brief.cfm?product{_}id=185

[124] National Institute of Health, "Unified Modeling Language System (UMLS)," 2020. [Online]. Available: https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html{#}

[125] WHO, "ICD-11," 2019. [Online]. Available: https://icd.who.int/icd11refguide/en/index.html

[126] B. Dutta and H.-G. Hwang, "The adoption of electronic medical record by physicians: A prisma-compliant systematic review," *Medicine*, vol. 99, no. 8, 2020.

[127] C. J. F. Candel, D. S. Ruiz, and J. J. García-Molina, "A unified metamodel for nosql and relational databases," *Information Systems*, vol. 104, p. 101898, 2022.

[128] A. Gasparetto, M. Marcuzzo, A. Zangari, and A. Albarelli, "A survey on text classification algorithms: From text to predictions," *Information*, vol. 13, no. 2, 2022. [Online]. Available: https://www.mdpi.com/2078-2489/13/2/83

[129] L. Ismail, H. Materwala, A. P. Karduck, and A. Adem, "Requirements of health data management systems for biomedical care and research: Scoping review," *J Med Internet Res*, vol. 22, no. 7, p. e17508, Jul 2020.

[130] M. F. Abdullah and K. Ahmad, "The mapping process of unstructured data to structured data," in *2013 International Conference on Research and Innovation in Information Systems (ICRIIS)*. IEEE, 2013, pp. 151–155.

[131] F. A. Satti, M. Hussain, J. Hussain, S. I. Ali, T. Ali, H. S. M. Bilal, T. Chung, and S. Lee, "Unsupervised semantic mapping for healthcare data storage schema," *IEEE Access*, vol. 9, pp. 107 267–107 278, 2021.

[132] F. A. Satti, T. Ali, J. Hussain, W. A. Khan, A. M. Khattak, and S. Lee, "Ubiquitous health profile (uhpr): a big data curation platform for supporting health data interoperability," *Computing*, vol. 102, no. 11, pp. 2409–2444, 2020.

[133] Healthcare Information and Management Systems Society, "Definition of Interoperability," 2013. [Online]. Available: https://www.himss.org/sites/hde/files/d7/FileDownloads/HIMSS%20Interoperability%20Definition%20FINAL.pdf

[134] T. Ali, M. Hussain, W. A. Khan, M. Afzal, J. Hussain, R. Ali, W. Hassan, A. Jamshed, B. H. Kang, and S. Lee, "Multi-model-based interactive authoring environment for creating

shareable medical knowledge," *Computer Methods and Programs in Biomedicine*, vol. 150, pp. 41–72, 2017.

[135] "SNOMED Clinical Terminologies." [Online]. Available: http://www.snomed.org/snomed-ct/five-step-briefing

[136] S. Mohammed and J. Fiaidhi, *Ubiquitous Health and Medical Informatics: The Ubiquity 2.0 Trend and Beyond: The Ubiquity 2.0 Trend and Beyond*. IGI Global, 2010.

[137] F. A. Satti, W. Ali Khan, T. Ali, J. Hussain, H. W. Yu, S. Kim, and S. Lee, "Semantic bridge for resolving healthcare data interoperability," in *2020 International Conference on Information Networking (ICOIN)*, Jan 2020, pp. 86–91.

[138] X. Zhu, T. Li, and G. De Melo, "Exploring semantic properties of sentence embeddings," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 632–637.

[139] U. Kartoun, "A methodology to generate virtual patient repositories," *arXiv preprint arXiv:1608.00570*, 2016.

[140] L. Pan, X. Fu, F. Cai, Y. Meng, and C. Zhang, "Design a novel electronic medical record system for regional clinics and health centers in china," in *2016 2nd IEEE International Conference on Computer and Communications (ICCC)*. IEEE, 2016, pp. 38–41.

[141] A. Inokuchi, K. Takeda, N. Inaoka, and F. Wakao, "Medtakmi-cdi: interactive knowledge discovery for clinical decision intelligence," *IBM Systems Journal*, vol. 46, no. 1, pp. 115–133, 2007.

[142] T. Ali and S. Lee, "Reconciliation of snomed ct and domain clinical model for interoperable medical knowledge creation," in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2017, pp. 2654–2657.

[143] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural*

Language Processing. Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[144] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation," *BMC genomics*, vol. 21, no. 1, pp. 1–13, 2020.

[145] K. L. Gwet, *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC, 2014.

[146] M. Hussain, J. Hussain, M. Sadiq, A. U. Hassan, and S. Lee, "Recommendation statements identification in clinical practice guidelines using heuristic patterns," *2018 19th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, vol. 95, no. 10, pp. 152–156, 2018.

[147] H. Hemingway, F. W. Asselbergs, J. Danesh, R. Dobson, N. Maniadakis, A. Maggioni, G. J. Van Thiel, M. Cronin, G. Brobert, P. Vardas, S. D. Anker, D. E. Grobbee, and S. Denaxas, "Big data from electronic health records for early and late translational cardiovascular research: Challenges and potential," *European Heart Journal*, vol. 39, no. 16, pp. 1481–1495, 2018.

[148] T. Ali and S. Lee, "Reconciliation of SNOMED CT and domain clinical model for interoperable medical knowledge creation," *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 2654–2657, 2017.

[149] H. Liu and P. Singh, "Conceptnet—a practical commonsense reasoning tool-kit," *BT technology journal*, vol. 22, 06 2004.

[150] P. Jesus, C. Baquero, and P. Almeida, "ID Generation in Mobile Environments," pp. 1–4, 2006. [Online]. Available: http://hdl.handle.net/1822/36065

[151] P. Leach, Microsoft, M. Mealling, Refactored Networks, R. Salz, and I. DataPower Technology, "Rfc 4122," 2005. [Online]. Available: https://www.ietf.org/rfc/rfc4122.txt

[152] U. Kartoun, M. General, and H. Harvard, "A Methodology to Generate Virtual Patient Repositories," *CoRR*, vol. abs/1608.00570, 2016. [Online]. Available: http://arxiv.org/abs/1608.00570

[153] AMD, "AMD Ryzen 3," 2018. [Online]. Available: https://www.amd.com/en/products/apu/amd-ryzen-3-2200g

[154] The CentOS Project, "CentOS," 2020. [Online]. Available: https://wiki.centos.org/

[155] F. A. Satti, W. A. Khan, G. Lee, A. M. Khattak, and S. Lee, "Resolving data interoperability in ubiquitous health profile using semi-structured storage and processing," in *In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing (SAC'19)*. ACM, 2019, pp. 762–770.

[156] P. Gohil and B. Panchal, "Efficient ways to improve the performance of hdfs for small files," *Computer Engineering and Intelligent Systems*, vol. 5, no. 1, pp. 45–49, 2014.

[157] B. Gupta, R. Nath, G. Gopal, and K. K, "An efficient approach for storing and accessing small files with big data technology," *International Journal of Computer Applications*, vol. 146, no. 1, pp. 36–39, 2016.

[158] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[159] Adam Cohen, "Fuzzywuzzy: Fuzzy string matching in python," Jul 8th, 2011. [Online]. Available: https://chairnerd.seatgeek.com/fuzzywuzzy-fuzzy-string-matching-in-python/

[160] D. Cer, M. Diab, E. Agirre, I. Lopez-Gazpio, and L. Specia, "Semantic textual similarity-multilingual and cross-lingual focused evaluation," in *Proceedings of the 2017 SEMVAL International Workshop on Semantic Evaluation (2017). https://doi. org/10.18653/v1/s17-2001*, 2017.

[161] M. L. McHugh, "Interrater reliability: the kappa statistic," *Biochemia medica*, vol. 22, no. 3, pp. 276–282, 2012.

[162] M. Hussain, F. A. Satti, J. Hussain, T. Ali, S. I. Ali, H. S. M. Bilal, G. H. Park, S. Lee, and T. Chung, "A practical approach towards causality mining in clinical text using active transfer learning," *Journal of Biomedical Informatics*, vol. 123, p. 103932, 2021.

[163] H. K. Walker, W. D. Hall, and J. W. Hurst, "Clinical methods: the history, physical, and laboratory examinations," 1990.

[164] S. Chen, X. Guo, T. Wu, and X. Ju, "Exploring the online doctor-patient interaction on patient satisfaction based on text mining and empirical analysis," *Information Processing & Management*, vol. 57, no. 5, p. 102253, 2020.

[165] D. Reidsma and R. op den Akker, "Exploiting 'subjective'annotations," in *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, 2008, pp. 8–16.

[166] S. Velupillai, "Towards a better understanding of uncertainties and speculations in swedish clinical text–analysis of an initial annotation trial," in *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, 2010, pp. 14–22.

[167] M. Joshi and P. Bhardwaj, "Impact of data transparency: Scientific publications," *Perspectives in Clinical Research*, vol. 9, no. 1, pp. 31–36, 2018. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/29430415https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5799949/

[168] E. Pavlick and T. Kwiatkowski, "Inherent disagreements in human textual inferences," *Transactions of the Association for Computational Linguistics*, vol. 7, pp. 677–694, 2019.

[169] R. P. Espíndola and N. F. Ebecken, "On extending f-measure and g-mean metrics to multi-class problems," *WIT Transactions on Information and Communication Technologies*, vol. 35, 2005.

[170] L.-E. Axelsson, "Identify user profiles in information systems with unknown users - a database modelling approach," *International Journal of Public Information Systems*, vol. 2006, no. 2, pp. 19–32, 2006.

[171] A. Tashkandi, I. Wiese, and L. Wiese, "Efficient in-database patient similarity analysis for personalized medical decision support systems," *Big Data Research*, vol. 13, pp. 52–64, 2018.

[172] L. Coventry and D. Branley, "Cybersecurity in healthcare: A narrative review of trends, threats and ways forward," *Maturitas*, vol. 113, no. March, pp. 48–52, 2018.

[173] R. Priya, S. Sivasankaran, P. Ravisasthiri, and S. Sivachandiran, "A survey on security attacks in electronic healthcare systems," *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, vol. 2018-Janua, pp. 691–694, 2018.

Appendix A

List of Acronyms

Acronyms

In alphabetical order:

AI Artificial Intelligence

BERT Bidirectional Encoder Representations from Transformers

CA Condition Action

CC Condition Consequances

CDSS Clinical Decision Support System

CPG Clinical Practice Guideline

CTTM Casual Triple Trained Model

EHR Electronic Health Record

HIMSS Healthcare Information and Management Systems Society

HMIS Healthcare Management Information Systems

HL7 Health Level Seven International

ICT Information and Communication Technologies

KE Knowledge Engineer

ML Machine Learning

NBC Naive based Classifier

NER Named Entity Recognition

NGT Nominal Group Technique

NLP Natural Lanauge Processing

POS Part of Speech

RegEx Regular Expressions

SVM Support Vector Machine

UMLS Unified Medical Language System

Appendix B

List of Publications

B.1 International Journal Papers [6]

1 **Satti, Fahad Ahmed**, Musarrat Hussain, Jamil Hussain, Syed Imran Ali, Taqdir Ali, Hafiz Syed Muhammad Bilal, Taechoong Chung, and Sungyoung Lee. "Unsupervised Semantic Mapping for Healthcare Data Storage Schema." IEEE Access 9 (2021): 107267-107278.

2 **Satti, Fahad Ahmed**, Taqdir Ali, Jamil Hussain, Wajahat Ali Khan, Asad Masood Khattak, and Sungyoung Lee. "Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability." Computing 102, no. 11 (2020): 2409-2444.

3 Hussain, Musarrat, **Fahad Ahmed Satti**, Jamil Hussain, Taqdir Ali, Syed Imran Ali, Hafiz Syed Muhammad Bilal, Gwang Hoon Park, Sungyoung Lee, and TaeChoong Chung. "A practical approach towards causality mining in clinical text using active transfer learning." Journal of Biomedical Informatics 123 (2021): 103932.

4 Hussain, Musarrat, **Fahad Ahmed Satti**, Syed Imran Ali, Jamil Hussain, Taqdir Ali, Hun-Sung Kim, Kun-Ho Yoon, TaeChoong Chung, and Sungyoung Lee. "Intelligent knowledge consolidation: From data to wisdom." Knowledge-Based Systems 234 (2021): 107578.

5 Imran Ali, Syed, Bilal Ali, Jamil Hussain, Musarrat Hussain, **Fahad Ahmed Satti**, Gwang Hoon Park, and Sungyoung Lee. "Cost-sensitive ensemble feature ranking and automatic threshold selection for chronic kidney disease diagnosis." Applied Sciences 10, no. 16 (2020): 5663.

6 Ali, Syed Imran, Hafiz Syed Muhammad Bilal, Musarrat Hussain, Jamil Hussain, **Fahad Ahmed Satti**, Maqbool Hussain, Gwang Hoon Park, Taechoong Chung, and Sungyoung

Lee. ”Ensemble feature ranking for cost-based non-overlapping groups: A case study of chronic kidney disease diagnosis in developing countries.” IEEE Access 8 (2020): 215623-215648.

B.2 Domestic Journal Papers [1]

1 Musarrat Hussain, Taqdir Ali, Jamil Hussain, **Fahad Ahmed Satti**, Usman Akhtar, Jaehun Bang, Taeho Hur, Sun Moo Kang, Byeong Ho Kang, and Sungyoung Lee. ”Intelligent Medical Platform: IMP”, The Journal of The Korean Institute of Communication Science, 37, no. 9(2020): 3-17.

B.3 Patents [1]

4 Sungyoung Lee and **Fahad Ahmed Satti** ”APPARATUS FOR JUST-IN-TIME SEMANTIC RECONCILIATION OF MEDICAL DATA AND METHOD THEREOF, AND METHOD FOR GENERATING SCHEME MAP ON THE APPARATUS” *Korean Intellectual Property Office*, Application No. 1020200156787, Applied on: 2020.11.20.