KYUNG HEE UNIVERSITY

Department of Computer Science & Engineering,
KHU, South Korea

PhD Dissertation Presentation

Semantic Sequence Contraction and Expansion for Data Interoperability

Fahad Ahmed Satti
Fahad.satti@oslab.khu.ac.kr

Advisors: Prof. Sungyoung Lee, PhD
Prof. TaeChoong Chung, PhD

PRESENTATION AGENDA

INTRODUCTION

- Background
- Motivation
- Problem Statement
- Taxonomy
- Related Work

PROPOSED SOLUTION

- Solution 1: Sequence Classification
- Solution 2: Schema Alignment
- Solution 3: Semantic Reconciliation-on-Read

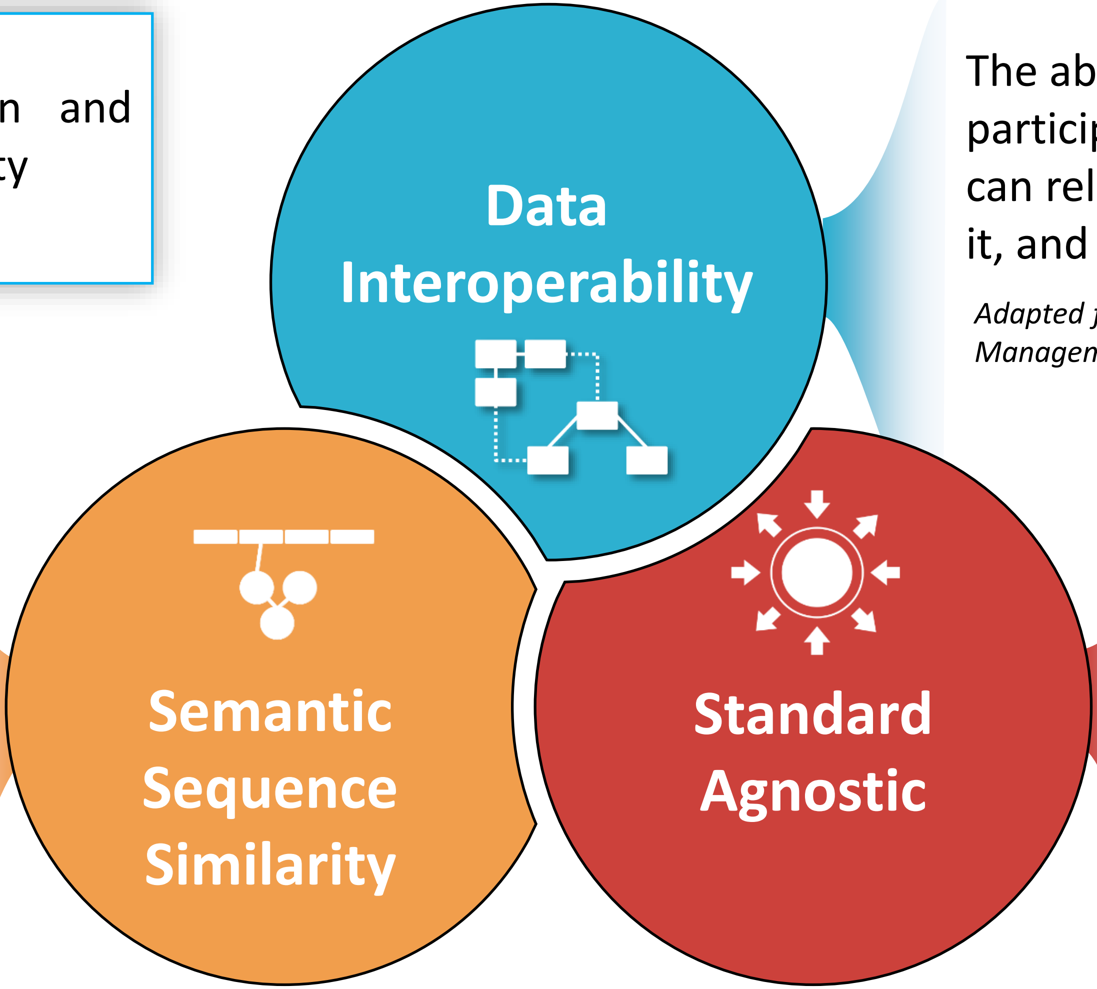
EXPERIMENTS & RESULTS

- Dataset
- Experimental Setup
- Results & Discussion

CONCLUSION

Background

Semantic Sequence Contraction and Expansion for Data Interoperability



The ability with which, two or more participating systems or components can reliably exchange data, interpret it, and use it.

Adapted from IEEE 610.12, HL7 and Healthcare Information Management Systems Society (HIMSS)

Determine if any two given entities are similar or dissimilar based on their respective, hidden meaning.

- Concept Dictionaries (Traditional, Expert Driven)
- Positional Context (Modern, Machine Driven)

A methodology, which works independently of any developed or under-development standards.

e.g. ICD-10, SNOMED-CT, LOINC, HL7 CDA, OpenEHR, HL7 FHIR

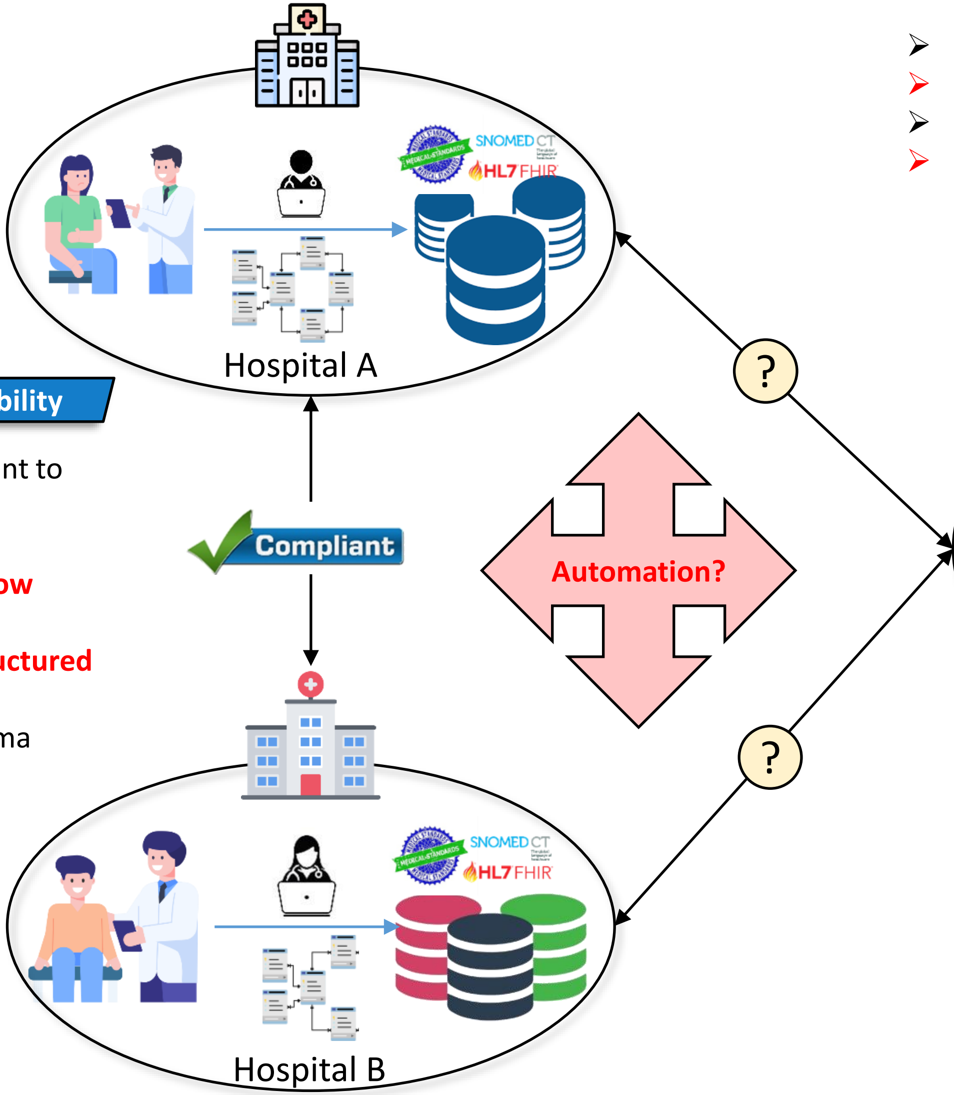
Motivation

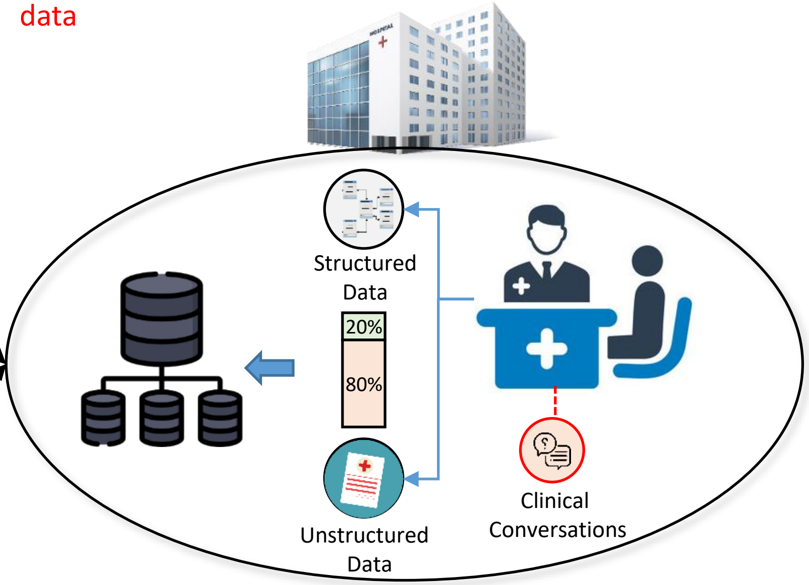
Conformance for Adhoc Schema ?

- **Resource constraints** (especially in the developing world)
- Small-Mid scale Hospitals and clinics
- Medical data with non-standard (adhoc) schema
- Unstructured data represents **80%** of medical data
- Linking unstructured data with structured medical data

Standard-Driven Data Interoperability

- **Expert Driven** schema alignment to exchange data between heterogeneous sources.
- Getting to an agreement is a **slow** process
- Operates on (standardized) **structured data**
- No automatic support for schema **evolution**





Clinical Conversations

- Primary point of data collection and inference
- Without capturing this data source
  - Some data can be lost due to cognitive load
  - Redundant effort required to digitize EMR
  - Restricts effective utilization in the developing world

Motivation

Standard-Driven Data Interoperability

- Expert Driven
- Highly Accurate
- Useful in the long run
- Getting to an agreement is a slow process
- Operates on structured data
- No automatic support for schema evolution

Operational Gaps

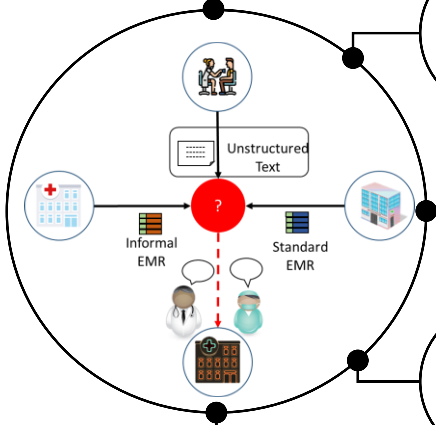
- Difficult for small health centers to support the upgradation of existing systems
- 80% of medical data is in unstructured form (Oliver 2016)

Active Communities

- CIMI(HL7 + OpenEHR)
- Yosemite Group
- SNOMED-CT + LOINC

Standard-Agnostic Data interoperability

A methodology, which works independently of any formally defined standard schema or otherwise.



Remove redundancies in collecting clinical history and conducting medical tests

Machine-driven, automatic solution, which supports **schema evolution**

Operates on **structured** and **unstructured** data.

Jumpstart **standard compliance** for small to mid scale hospitals and clinics

Reduce stress on hospitals and clinics in the developing world

Problem Statement

Problem statement

Existing Health data interoperability solutions are **expert-driven** and **standard dependent**, loosing a plethora of data residing in **informal schema** and **unstructured format**, and hindering the achievement of **Ubiquitous Healthcare**.

Goal

To identify and extract **clinical data** in a form consumable by various data engines for storage, usage, or exchange.

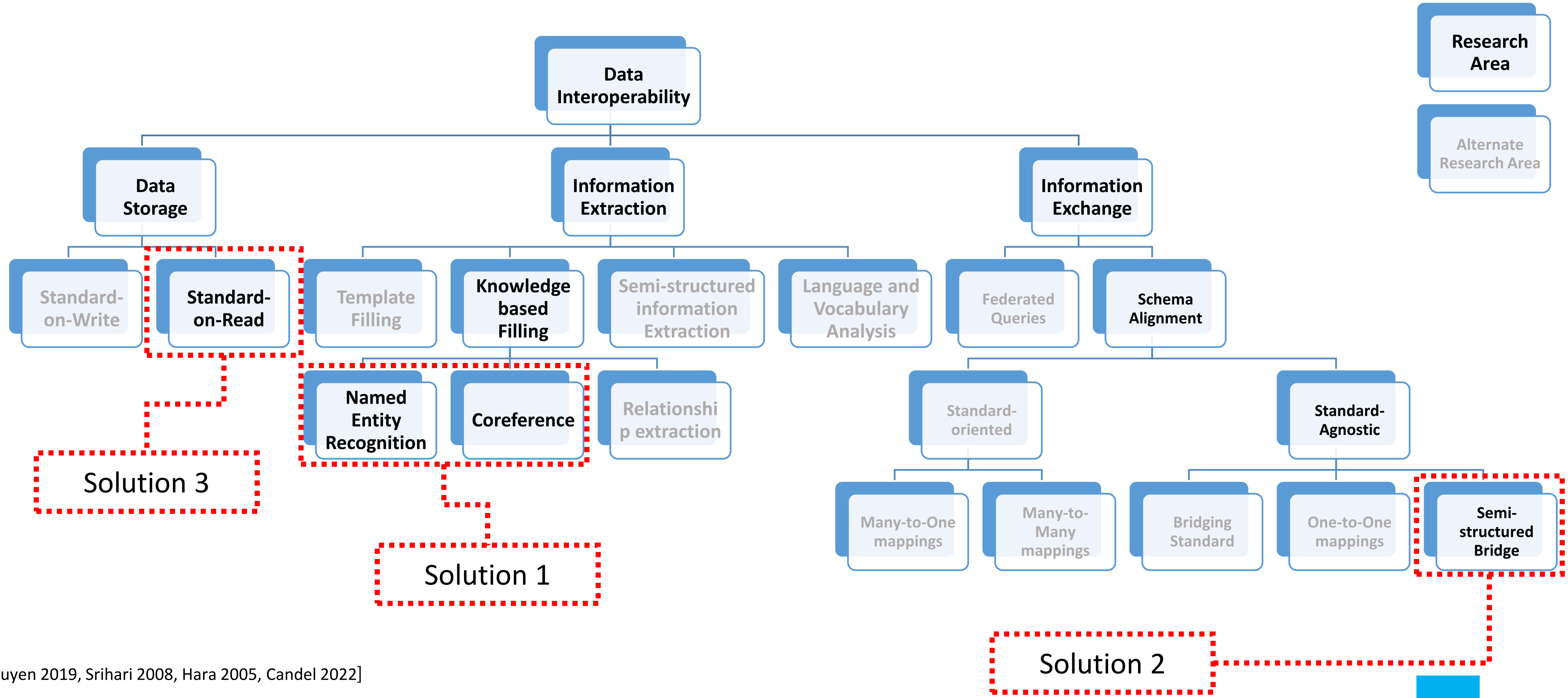
Standard-Agnostic

Data Interoperability

Challenges

- Challenge 1: **Identify** & **Extract** clinical **attributes** and their **values** from unstructured text
- Challenge 2: **Automatically align heterogeneous** structured and semi-structured **schema**
- Challenge 3: **Design** a **scalable** infrastructure, **automating** data interoperability.

Research Taxonomy



[Nguyen 2019, Srihari 2008, Hara 2005, Candel 2022]

Related Work

Literature Survey for Sequence Contraction

	Research	Method	Advantages	Limitations
Unstructured to Structured Data (Sequence Contraction)	Du (2019)	○ Uses Bi-LSTM with CRF to first identify the sequence of interest containing a symptom and identify the target symptom ○ Dual RNN based Seq2Seq model for identifying the similarity between utterances and existing attributes-values.		The utterances of symptoms must be **sequential** due to the seq2seq model which relies on the in-order occurrences of symptoms.
	Lin (2019)	○ Use Bi-LSTM with a global attention mechanism to get the contextual information from document level and corpus level. The hidden layers are then re-encoded and decoded by CRF to recognize the symptoms. ○ A symptom graph is used for symptom classification	Utilizes the semantics at document level and corpus level to identify the context of the data	○ Only works on limited **pre-defined items** (authors showed results only for “upper respiratory infection”, “functional dyspepsia”, “infantile diarrhea” and “bronchitis”). ○ The utterances of symptoms must be **sequential** due to the use of symptom graph
	Du (2020)	Proposed a deep learning-based approach to extract medically relevant attributes from EMR	Uses ALBERT model, which provides much better results than the traditional LSTM-CRF model.	Difficult to generalize the solution without **model retraining**
	Zhang (2020)	Utilizes Candidate Attribute-value pairs and their status, to calculate similarity between Bert based encoded vectors for utterances and the candidates	Takes into account both the statements and question/answer type of utterances.	Only works with **existing Candidates** and is **unable to extract unseen** medical artifacts

Challenge 1: Limitations of existing work

- Most have used **a small** set of **pre-defined attributes** which lack generalization and require intensive human efforts and time.

Related Work

Literature Survey for Sequence Expansion

	Research	Method	Advantages	Limitations
Schema Alignment (Sequence Expansion)	Bulygin (2018)	Devised an ontology and schema matching based approach by combining lexical and conceptual semantic similarity with various ML algorithms.	The authors have testing various ML algorithms, including Naïve Bayes, Logistic Regression, and Gradient Boosted Tree.	○ Only operates on entities of **pre-defined ontologies**. ○ All entities are matched using **naïve** comparison.
	Nozaki (2019)	Utilized instance-based matching and Word2Vec to create embedding vectors and calculate similarity of attributes across heterogeneous databases.	Operates on heterogeneous databases	○ Word2Vec suffers from **Out of Vocabulary** problem. ○ Only limited experiments, which do not take into account the concepts behind the values
	Yousfi (2020)	○ Proposed an XML schema matcher, which uses conceptual semantic techniques, to transform schemas into set of words, measures each words context. ○ Similar words are identified based on relatedness score using WordNet.	Operates on heterogeneous xml documents	○ A **well-defined XSD** is necessary ○ Only works on well formed **markup** languages ○ **Relatedness** score of WordNet is an **old technique**, which has been replaced by the seq2seq based semantic similarity
	Kersloot (2020)	Reviewed several NLP algorithms for clinical text mappings onto ontological concepts.	The authors revealed that over one fourth of the NLP algorithms used were not evaluated and have no validation.	Systematic Review only

Challenge 2: Limitations of existing work

- Most solutions require a well-defined schema, which correctly and completely identifies each entity
- Out of vocabulary problem can greatly limit the performance of the whole technique
- Model trained on a specific dataset are unable to generalize

Related Work

Literature Survey for Semantic Reconciliation-on-Read

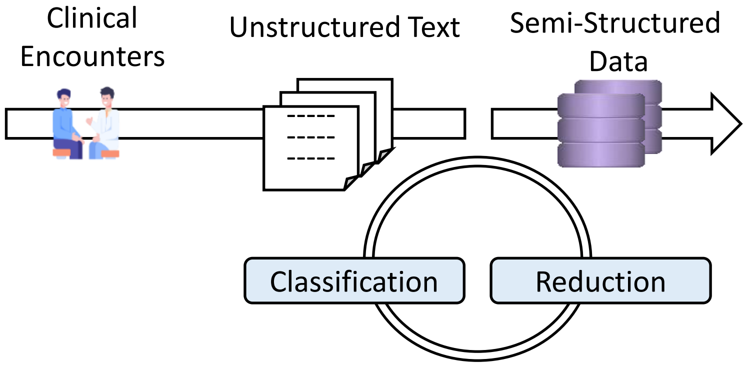
	Research	Method	Advantages	Limitations
Semantic Reconciliation-on-Read	LinkEHR (2019)	○ Uses well defined archetypes to provide a semantic and syntactic transformation engine with large input from knowledge engineer. ○ Depends on HL7 CDA and OpenEHR ○ Federated query model which is based on one-to-one mapping	Provides good alignment between HL7 CDA and openEHR.	○ **Standard dependent** ○ Data retrieval dependent on how well the **transformation** definitions are. ○ No **traceability** of healthcare records. ○ Schema **evolution** necessitates expert input
	OBDA (2018)	○ OBDA, utilizes a well formed ontology to which all participating system must conform to. ○ Federated query model, which does not store any data	○ Does not store data, so the source data always reflects the most recent updates ○ One to one mapping, allows any consumer or producer to provide a conformance map only once	○ All systems must comply with **their standard** ○ Data retrieval dependent on how well the **transformation** definitions are. ○ No **traceability** of healthcare records. ○ Schema **evolution** necessitates expert input
	HSB (2015)	○ Similar to OBDA, however the producers and consumers are loosely coupled with each other ○ Transformation services from well-defined standard form to an internal format is required for exchanging data.	Service Bus architecture hides the details of the participating system from others	○ Participating systems can comply to any system, however they should be able to **transform** the data at their ends.

Challenge 3: Limitations of existing work

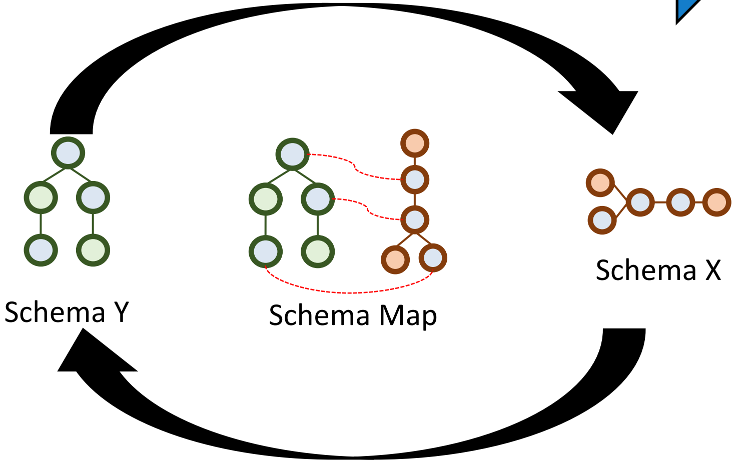
- ➢ Most solutions require a well-defined schema, which correctly and completely identifies each entity
- ➢ No traceability of health records
- ➢ Schema evolution necessitates expert input to resolve any new interoperability problems

Research Map

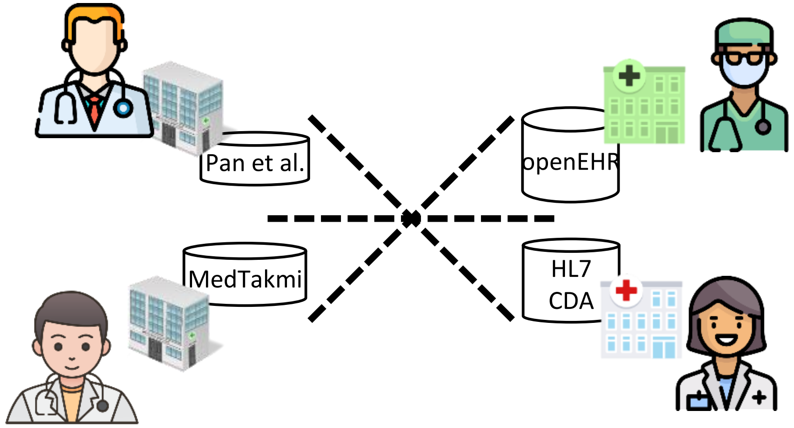
Challenge 1
Unstructured text to semi-Structured data



Challenge 2
Alignment between schema



Challenge 3
Semantic reconciliation-on-Read



Existing Solutions

1. Bi-LSTM with CRF instance of interest and Dual RNN based Seq2Seq model for value identification. Du (2019), Lin (2019)
2. BERT based approaches. Zhang (2020), Du (2020)

Limitations

- Mainly focus on a small set of attributes
- Lacks generalization.
- Require local ontologies

Existing Solutions

1. Ontology based approach. (Bulygin 2018)
2. Instance-based matching and Word2Vec. (Nozaki 2019)
3. conceptual semantic technique working on XSD. (Yousfi 2020)

Limitations

- Require a well-defined schema
- Out of vocabulary problem
- Lacks generalization

Existing Solutions

1. Federated query model HL7CDA and OpenEHR. (LinkEHR 2019)
2. Federated query model, with 1-1 mapping. (OBDA 2018)
3. Health Service Bus with loose 1-1 mapping. (HSB)

Limitations

- Require a well-defined schema
- No traceability of health records
- No support for schema evolution

Challenges and Proposed Solutions

Goal

To identify and extract **clinical data** in a form consumable by various data engines for storage, usage, or exchange.

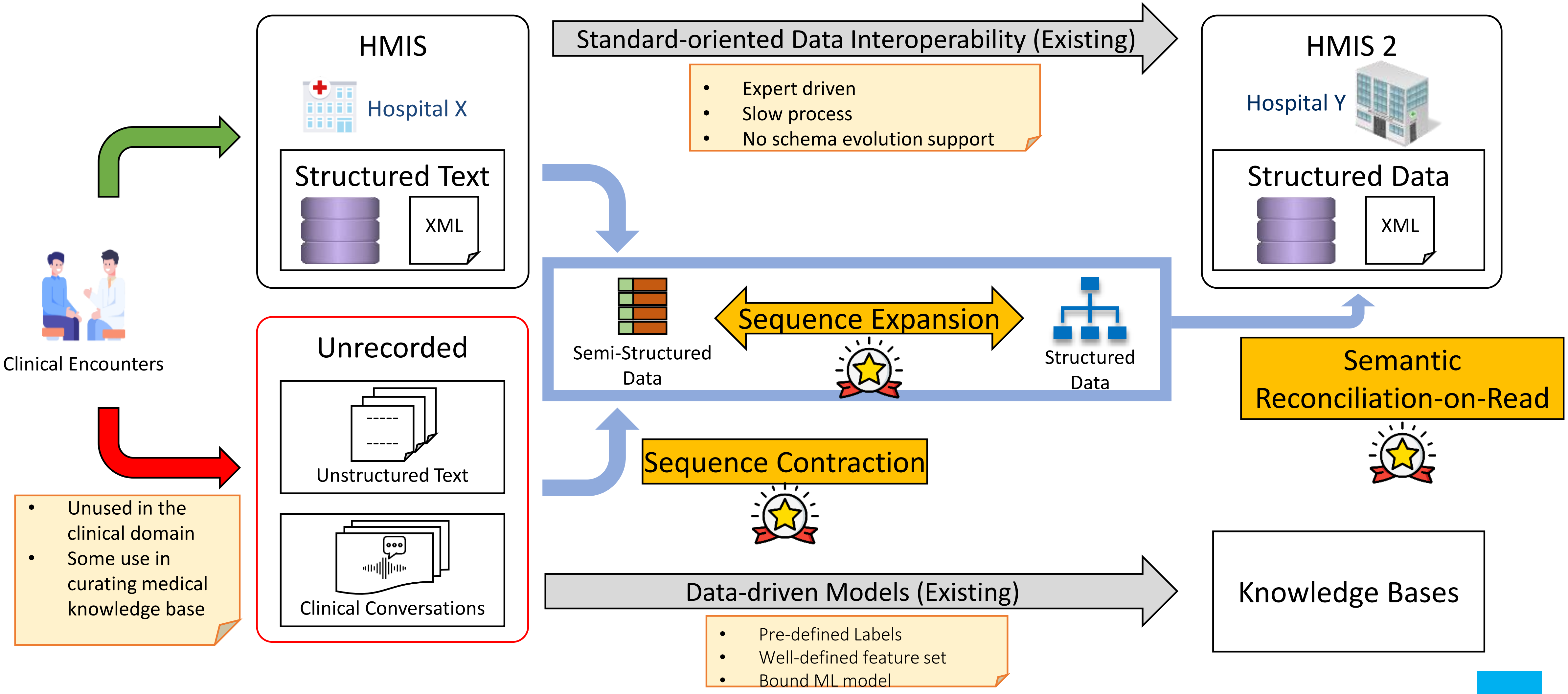
Challenges	**Proposed Solutions**	**Objectives**
C1 **Identify** & **Extract** clinical **attributes** and their **values** from unstructured text.	**S1: Sequence Contraction** Transfer Learning to classify sequences and application of syntactic and semantic extractors for creating attribute-value pairs.	Find attributes and values from unstructured data.
C2 **Automatically align heterogeneous** structured and semi-structured **schema**	**S2: Sequence Expansion** Semantic similarity of sequences, built from attribute names, using phrasal n-grams and concept enrichment.	Align Attributes with heterogeneous schema for data format transformation
C3 **Design** a **scalable** infrastructure, **automating** data interoperability.	**S3: Semantic Reconciliation-on-Read** A semi-structured data archiving and processing framework.	Design a practical platform which supports mapping evolution and low resource usage.

Proposed Methodology

Idea Diagram

Key Idea

- Create Sequences from Unstructured text and attribute names
- Define a set of true sequences, enriched with semantic concepts
- Apply semantic similarity to classify unseen data
- Transform the classified instances into required results



- Expert driven
- Slow process
- No schema evolution support

- Unused in the clinical domain
- Some use in curating medical knowledge base

- Pre-defined Labels
- Well-defined feature set
- Bound ML model

Proposed Methodology: Novelty

Semi-Structured Data

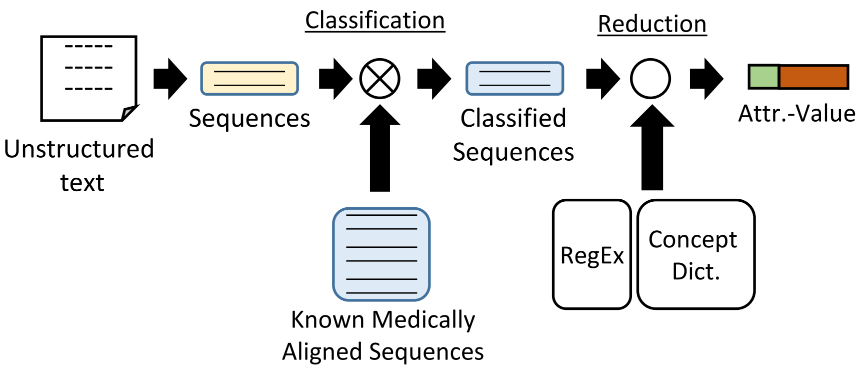
Structured Data

Semi-Structured Schema

Schema Map

Solution 1: Sequence Contraction

Identify relevant medical data from Clinical Conversations



Novelty

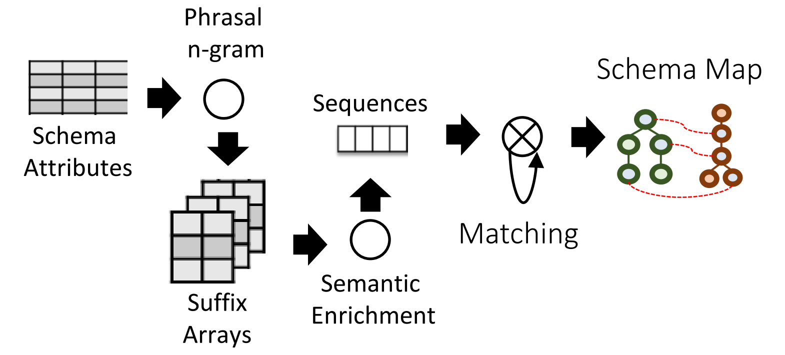
- Semantic Similarity based Classification
  - Transfer Learning to classify sequences
  - Easily extendable
  - Less training requirements
- Reduction
  - syntactic and semantic extractors for creating attribute-value pairs.
  - Captures syntactic artifacts like name, age, etc.
  - Utilizes conceptual semantics to enrich reduction

Limitation

- Requires expert intervention to build the set of known medically aligned sequences and Regular expressions

Solution 2: Sequence Expansion

Convert attribute to sequences with suffixes and concepts



Novelty

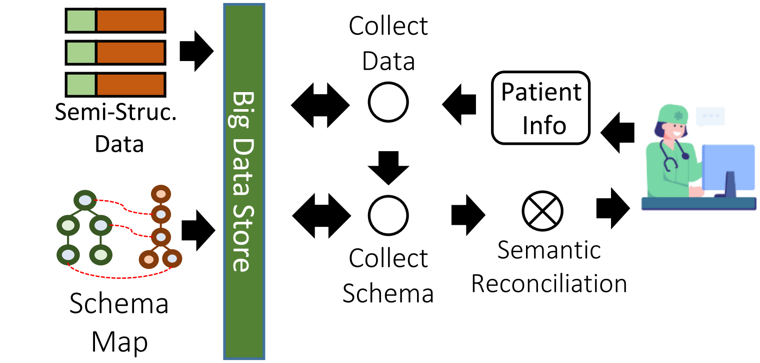
- Phrasal n-gram
  - Identify hidden words within attribute names
  - Handles adhoc naming conventions
- Semantic Enrichment
  - Utilizes semantic concepts for enriched matching
- Matching
  - Unsupervised
  - m-m matching between enriched sequences

Limitation

- No simple pathway for attributes with atomic names

Solution 3: Semantic Reconciliation-on-Read

Semi-structured data archiving and processing framework



Novelty

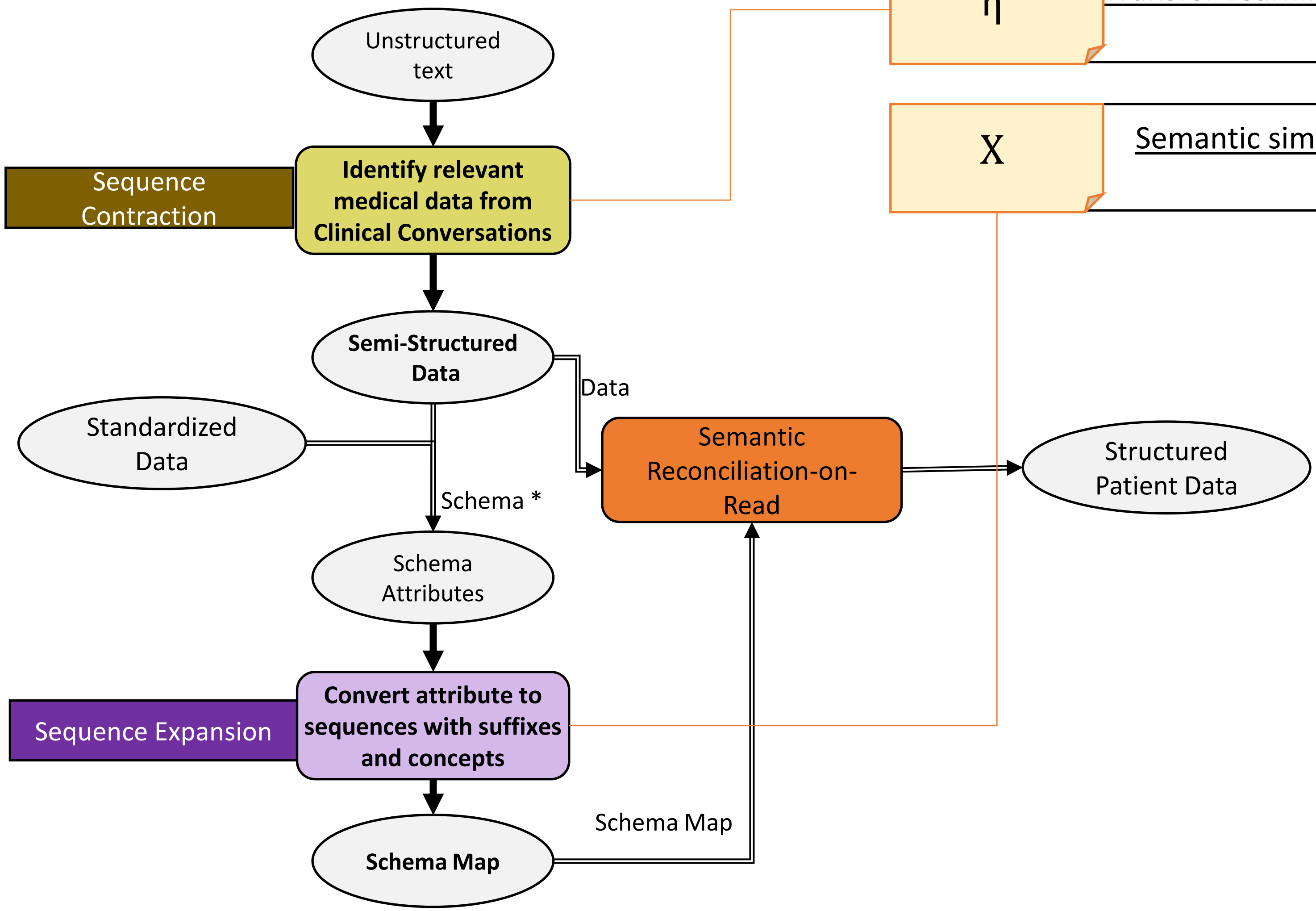
- Big Data Store
  - Low resource requirements at the end-nodes (hospitals/clinics)
  - Data archiving to prevent data loss
  - Evolvable schema-maps
- On demand transformation
  - Conversion to any standard
  - When required, utilize latest schema-map
  - Supports 1-1 and 1-m mappings

Limitation

- Resource requirements for the Big Data Store will be high

Proposed Methodology

Abstract Workflow



η — Transfer Learning to classify sequences and application of syntactic and semantic extractors for creating attribute-value pairs.

χ — Semantic similarity of sequences, built from attribute names, using phrasal n-grams and concept enrichment.

Unstructured corpus C

$$\exists C \wedge \exists \eta | \forall c \in C.\, \eta(c) \rightarrow p \vee \phi | p \in C$$

$$p =< p_a, p_v > | p_a \models p_v$$

Structured corpus C

$$\exists S \wedge \exists \zeta | \zeta(S) \rightarrow Q | Q = \{q\}$$

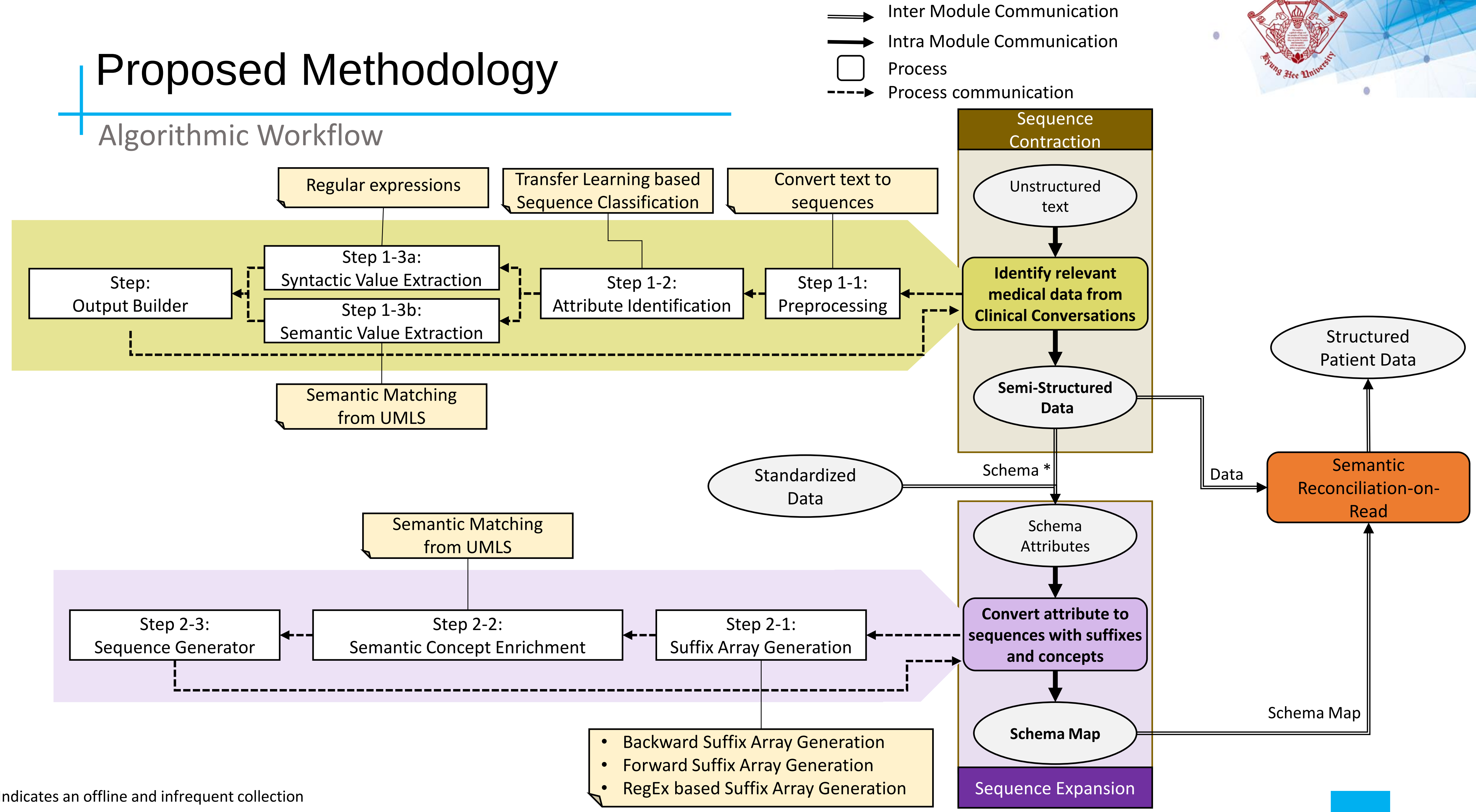
$$q =< q_a, q_v > | q_a \models q_v$$

$$\chi(p,q) = \begin{cases} 1 \ \ if \ \ (p_a = q_a) \\ \sim \ \ if \ \ (p_a \cong q_a) \\ 0 \ \ otherwise \end{cases} \quad \}$$

* Indicates an offline and infrequent collection

Proposed Methodology

Algorithmic Workflow

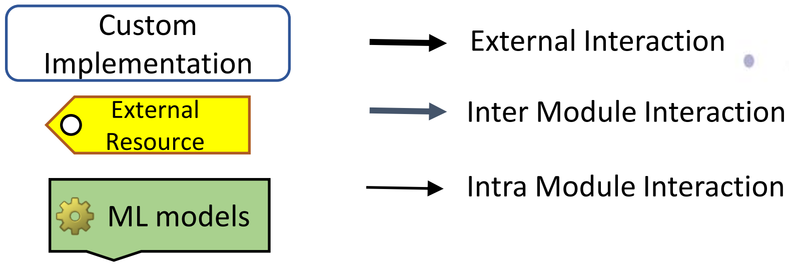


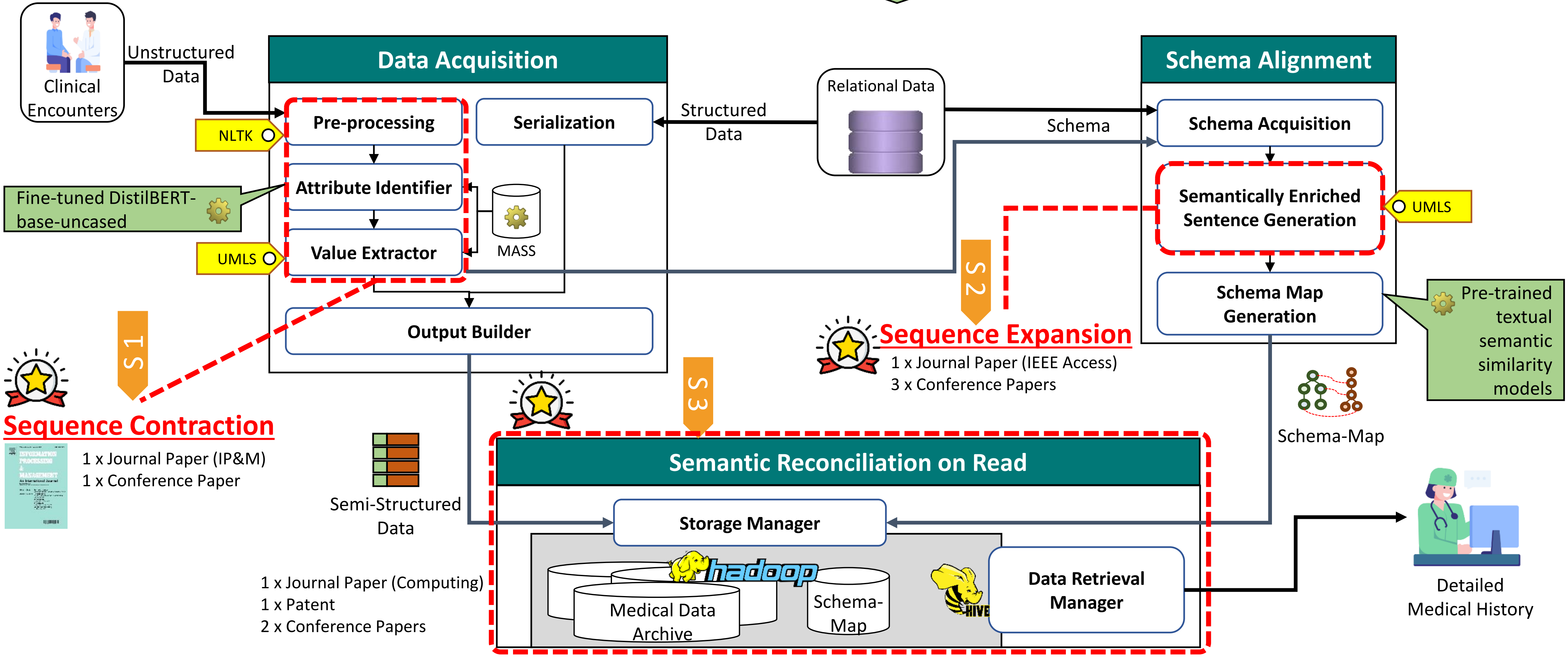
Inter Module Communication
Intra Module Communication
Process
Process communication
Sequence Contraction
Unstructured text
Identify relevant medical data from Clinical Conversations
Semi-Structured Data
Regular expressions
Transfer Learning based Sequence Classification
Convert text to sequences
Step 1-3a: Syntactic Value Extraction
Step: Output Builder
Step 1-2: Attribute Identification
Step 1-1: Preprocessing
Step 1-3b: Semantic Value Extraction
Semantic Matching from UMLS
Structured Patient Data
Standardized Data
Schema *
Data
Semantic Reconciliation-on-Read
Schema Attributes
Semantic Matching from UMLS
Step 2-3: Sequence Generator
Step 2-2: Semantic Concept Enrichment
Step 2-1: Suffix Array Generation
Convert attribute to sequences with suffixes and concepts
Schema Map
Schema Map
Backward Suffix Array Generation
Forward Suffix Array Generation
RegEx based Suffix Array Generation
Sequence Expansion

* Indicates an offline and infrequent collection

Proposed Methodology

System Perspective



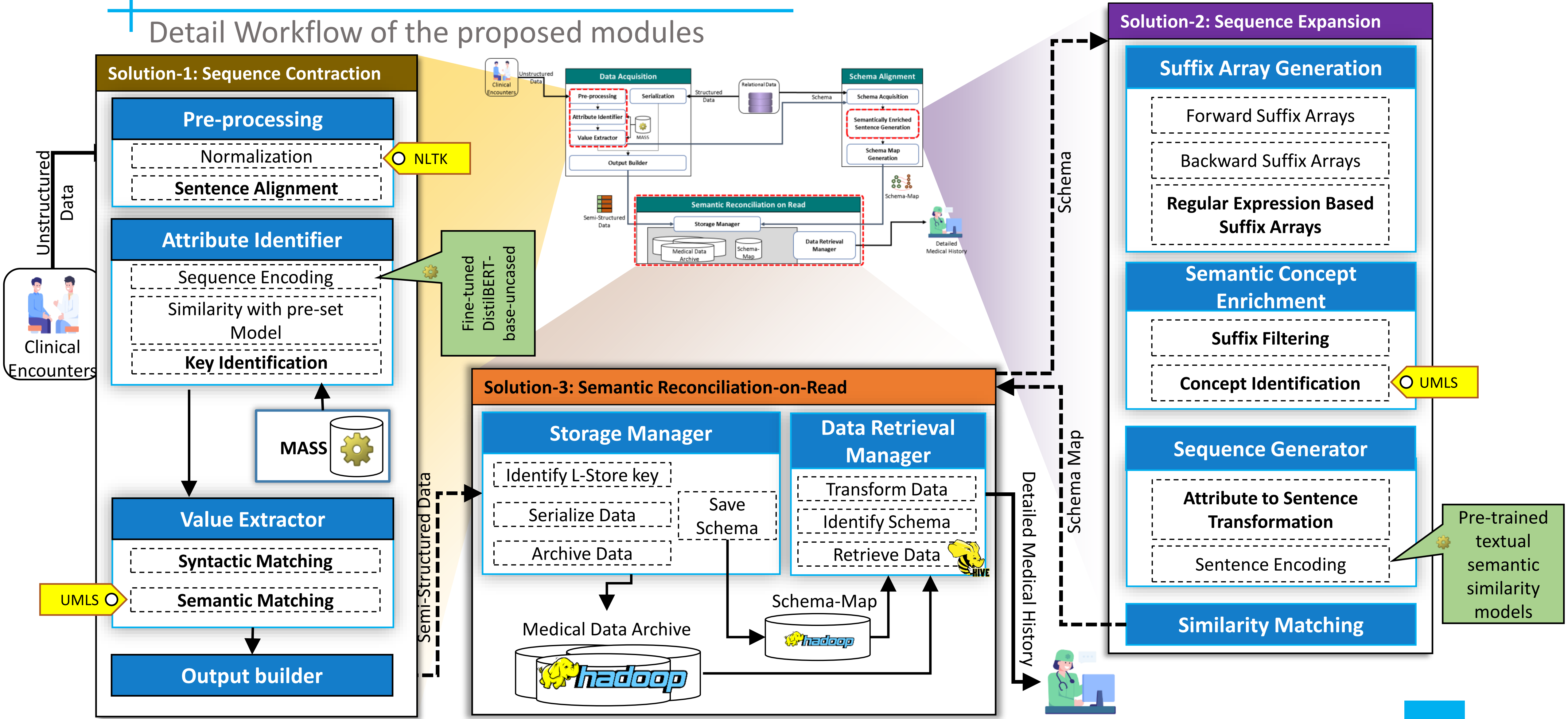


Proposed Methodology

Detail Workflow of the proposed modules

Inter Module Communication

External Communication



Solution-1: Sequence Contraction

- **Pre-processing**
  - Normalization
  - **Sentence Alignment**
  - NLTK
- **Attribute Identifier**
  - Sequence Encoding
  - Similarity with pre-set Model
  - **Key Identification**
  - Fine-tuned DistilBERT-base-uncased
- MASS
- **Value Extractor**
  - **Syntactic Matching**
  - **Semantic Matching**
  - UMLS
- **Output builder**

Unstructured Data

Clinical Encounters

Semi-Structured Data

Solution-2: Sequence Expansion

- **Suffix Array Generation**
  - Forward Suffix Arrays
  - Backward Suffix Arrays
  - **Regular Expression Based Suffix Arrays**
- **Semantic Concept Enrichment**
  - **Suffix Filtering**
  - **Concept Identification**
  - UMLS
- **Sequence Generator**
  - **Attribute to Sentence Transformation**
  - Sentence Encoding
  - Pre-trained textual semantic similarity models
- **Similarity Matching**

Schema

Schema Map

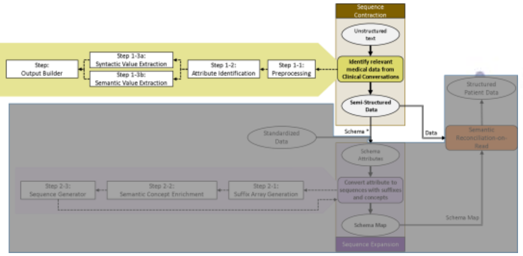
Solution-3: Semantic Reconciliation-on-Read

- **Storage Manager**
  - Identify L-Store key
  - Serialize Data
  - Archive Data
  - Save Schema
- **Data Retrieval Manager**
  - Transform Data
  - Identify Schema
  - Retrieve Data
  - HIVE
- Medical Data Archive
- hadoop
- Schema-Map
- hadoop

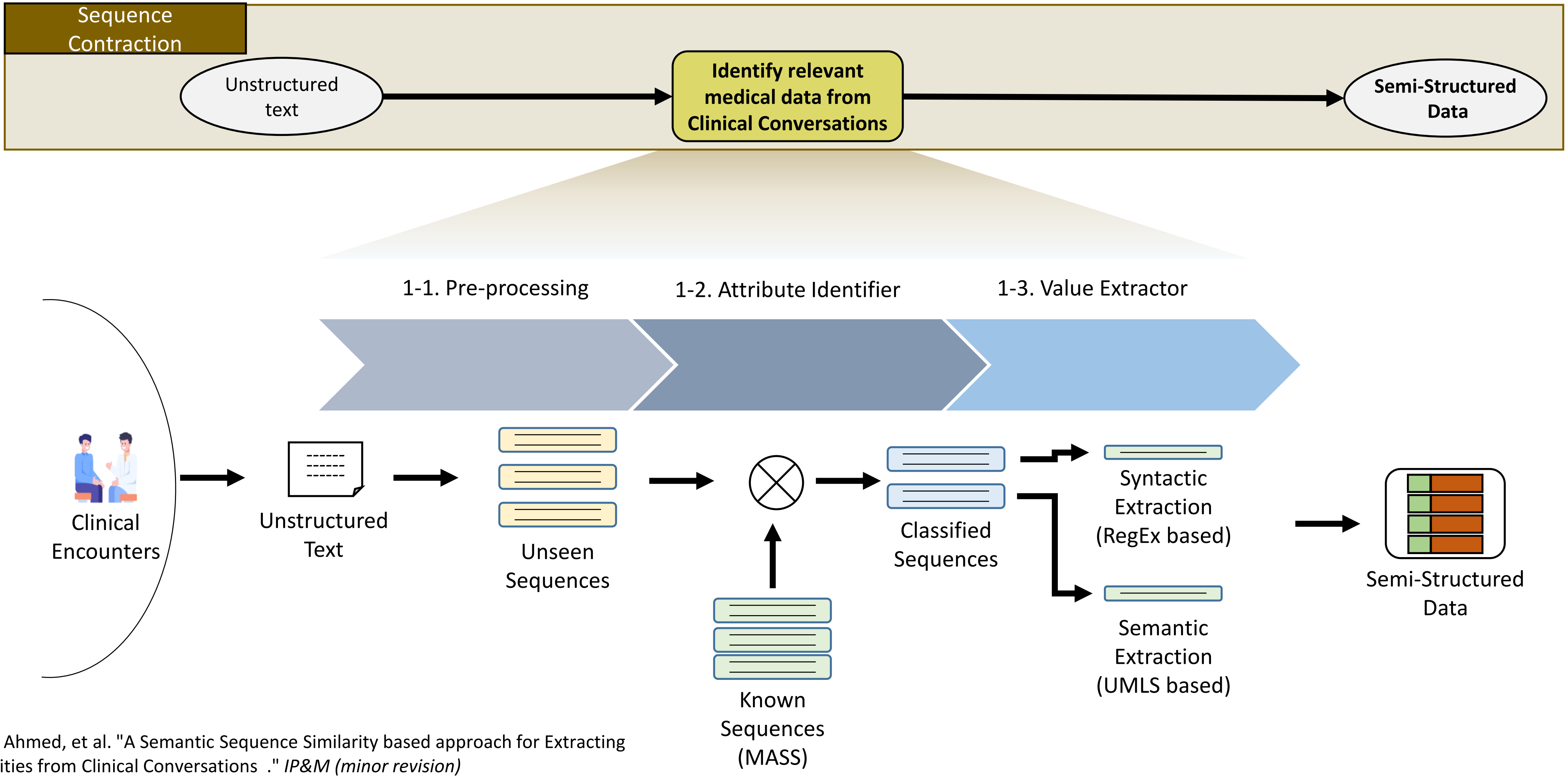
Detailed Medical History

Solution 1: Sequence Contraction

Abstract View



Sequence Contraction
Unstructured text
Identify relevant medical data from Clinical Conversations
Semi-Structured Data
Step: Output Builder
Step 1-3a: Syntactic Value Extraction
Step 1-3b: Semantic Value Extraction
Step 1-2: Attribute Identification
Step 1-1: Preprocessing
Structured Patient Data
Standardized Data
Schema *
Data
Semantic Reconciliation-on-Read
Schema Attributes
Convert attribute to sequences with suffixes and concepts
Step 2-3: Sequence Generator
Step 2-2: Semantic Concept Enrichment
Step 2-1: Suffix Array Generation
Schema Map
Schema Map
Sequence Expansion



Sequence Contraction

Unstructured text

Identify relevant medical data from Clinical Conversations

Semi-Structured Data

1-1. Pre-processing

1-2. Attribute Identifier

1-3. Value Extractor

Clinical Encounters

Unstructured Text

Unseen Sequences

Known Sequences (MASS)

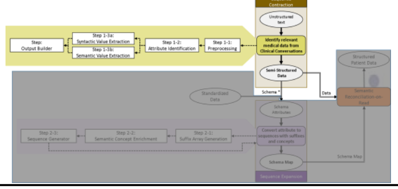
Classified Sequences

Syntactic Extraction (RegEx based)

Semantic Extraction (UMLS based)

Semi-Structured Data

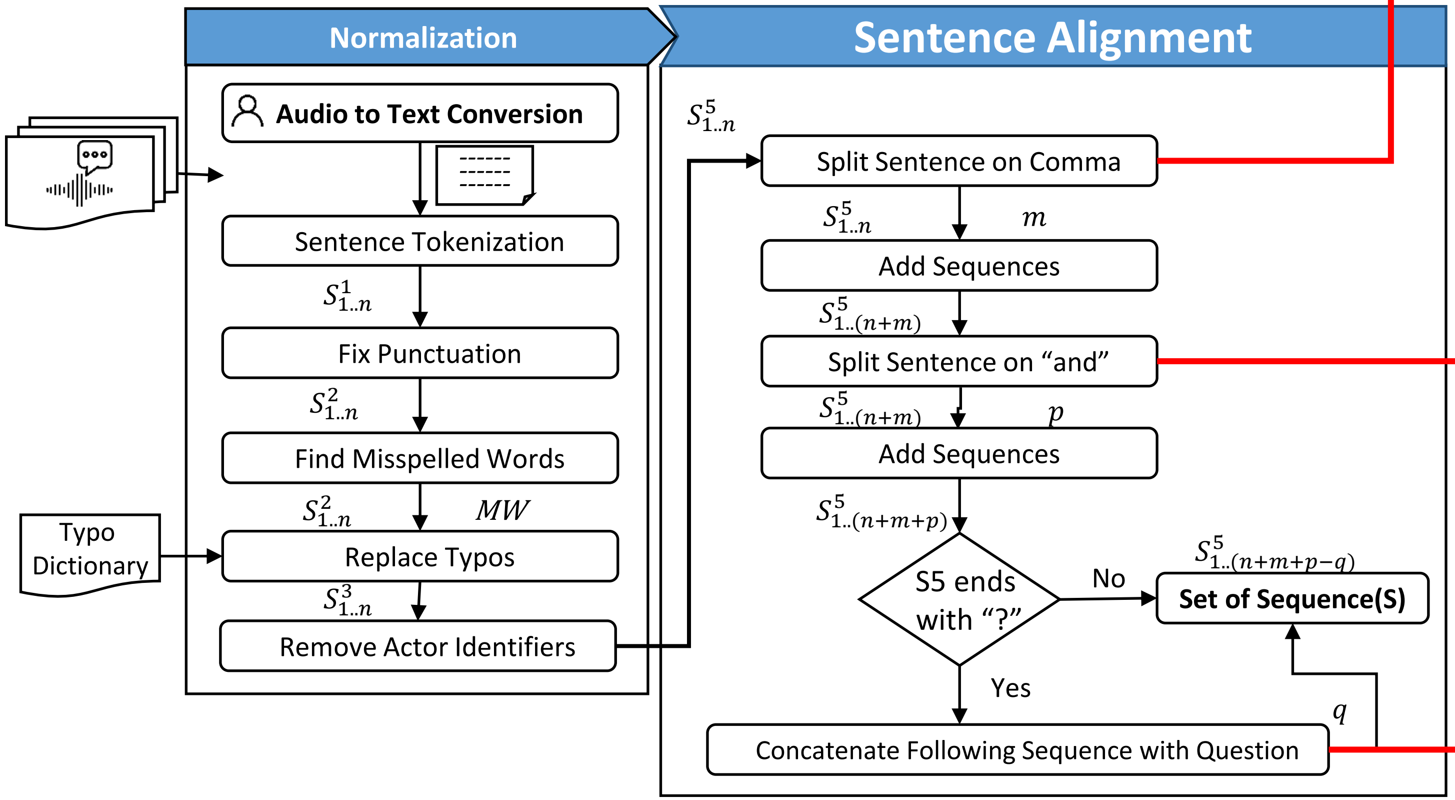
Satti, Fahad Ahmed, et al. "A Semantic Sequence Similarity based approach for Extracting Medical Entities from Clinical Conversations ." *IP&M (minor revision)*

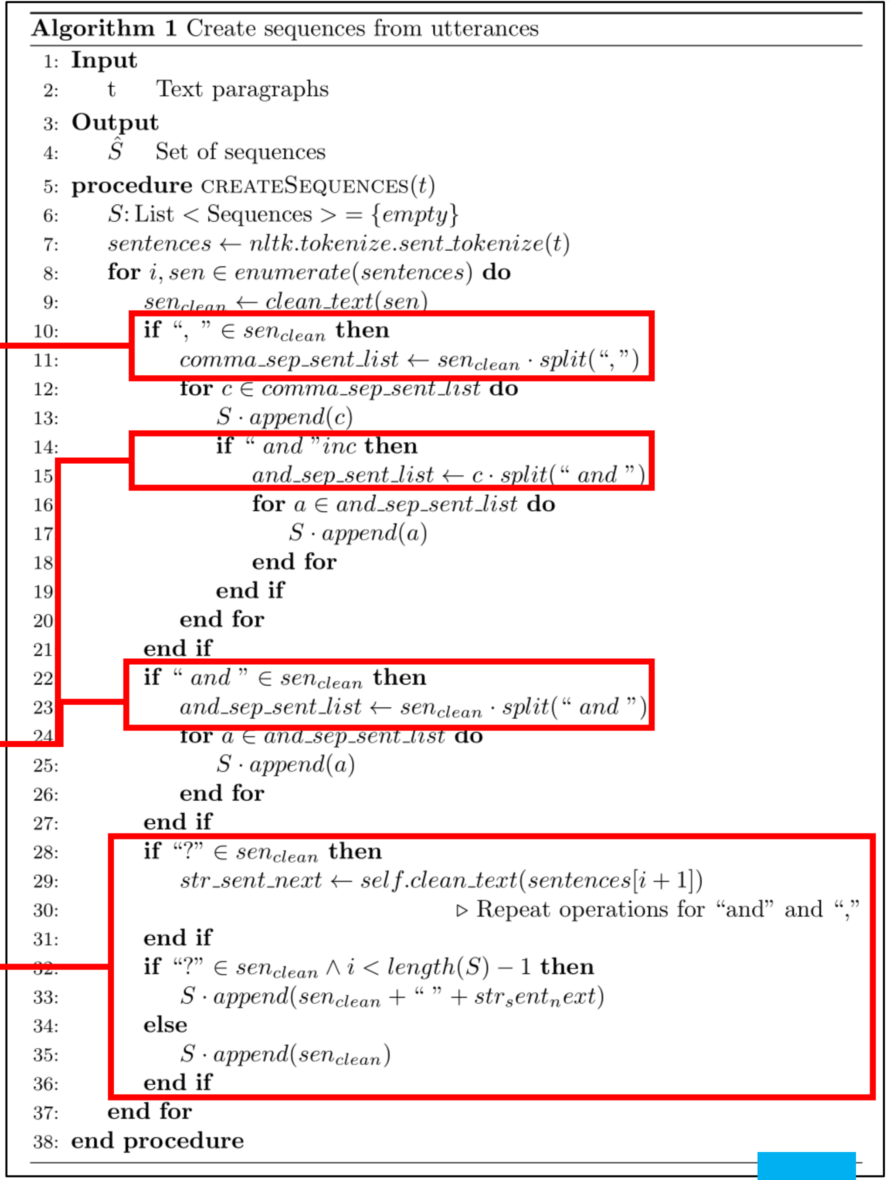


Solution 1-1: Pre-processing

Sequence Contraction

AIM	Benefits
Create sequences from text which contain both attributes and its values (Statements and Q/A)	• Shorter self contained sequences • Prioritizes shorter context over longer one • Faster processing



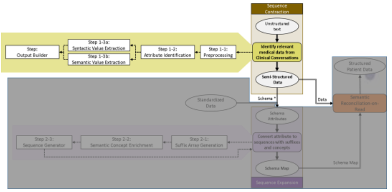


Algorithm 1 Create sequences from utterances

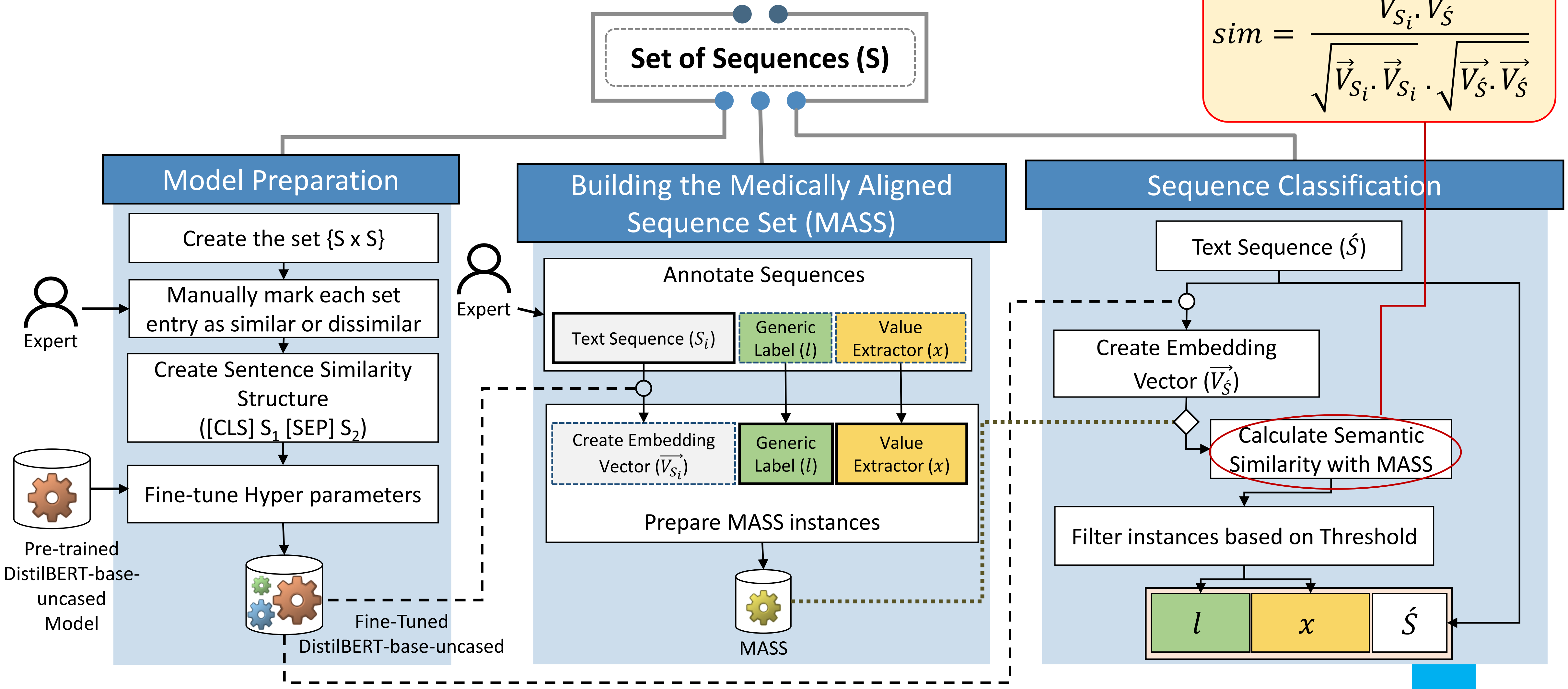
```
Input
    t    Text paragraphs
Output
    $\hat{S}$    Set of sequences
procedure CREATESEQUENCES($t$)
    $S$: List < Sequences > = $\{empty\}$
    $sentences \leftarrow nltk.tokenize.sent\_tokenize(t)$
    for $i, sen \in enumerate(sentences)$ do
        $sen_{clean} \leftarrow clean\_text(sen)$
        if “, ” $\in sen_{clean}$ then
            $comma\_sep\_sent\_list \leftarrow sen_{clean} \cdot split(“,”)$
            for $c \in comma\_sep\_sent\_list$ do
                $S \cdot append(c)$
                if “ and ”inc then
                    $and\_sep\_sent\_list \leftarrow c \cdot split(“ and ”)$
                    for $a \in and\_sep\_sent\_list$ do
                        $S \cdot append(a)$
                    end for
                end if
            end for
        end if
        if “ and ” $\in sen_{clean}$ then
            $and\_sep\_sent\_list \leftarrow sen_{clean} \cdot split(“ and ”)$
            for $a \in and\_sep\_sent\_list$ do
                $S \cdot append(a)$
            end for
        end if
        if “?” $\in sen_{clean}$ then
            $str\_sent\_next \leftarrow self.clean\_text(sentences[i+1])$
                                        ▷ Repeat operations for “and” and “,”
        end if
        if “?” $\in sen_{clean} \wedge i < length(S) - 1$ then
            $S \cdot append(sen_{clean} + “ ” + str_sent_next)$
        else
            $S \cdot append(sen_{clean})$
        end if
    end for
end procedure
```

Solution 1-2: Attribute Identifier

Sequence Contraction



$$sim = \frac{\vec{V}_{S_i} . \vec{V}_{\acute{S}}}{\sqrt{\vec{V}_{S_i} . \vec{V}_{S_i}} . \sqrt{\vec{V}_{\acute{S}} . \vec{V}_{\acute{S}}}}$$



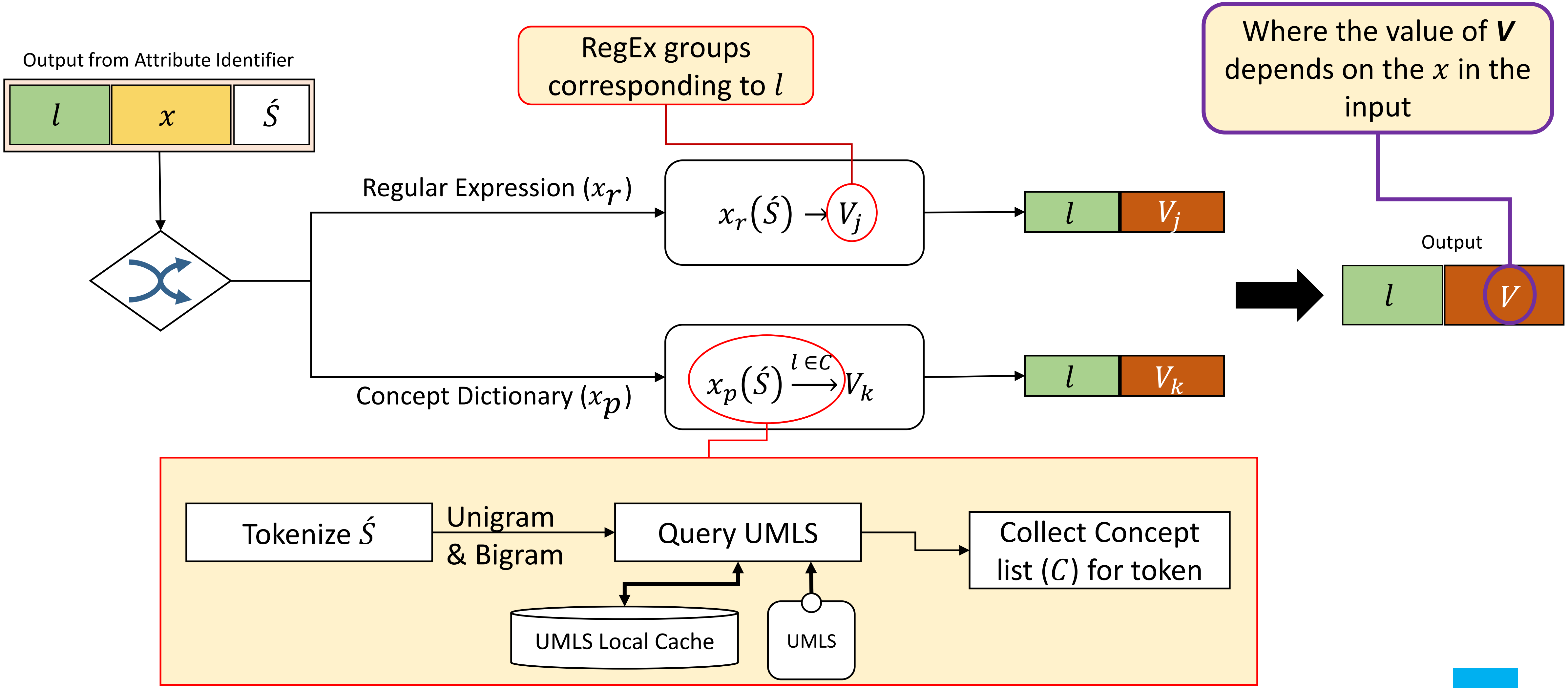
Set of Sequences (S)
Model Preparation
Create the set {S x S}
Manually mark each set entry as similar or dissimilar
Expert
Create Sentence Similarity Structure ([CLS] S_1 [SEP] S_2)
Fine-tune Hyper parameters
Pre-trained DistilBERT-base-uncased Model
Fine-Tuned DistilBERT-base-uncased
Building the Medically Aligned Sequence Set (MASS)
Expert
Annotate Sequences
Text Sequence (S_i)
Generic Label (l)
Value Extractor (x)
Create Embedding Vector ($\vec{V}_{S_i}$)
Generic Label (l)
Value Extractor (x)
Prepare MASS instances
MASS
Sequence Classification
Text Sequence ($\acute{S}$)
Create Embedding Vector ($\vec{V}_{\acute{S}}$)
Calculate Semantic Similarity with MASS
Filter instances based on Threshold
l
x
$\acute{S}$

https://huggingface.co/distilbert-base-uncased

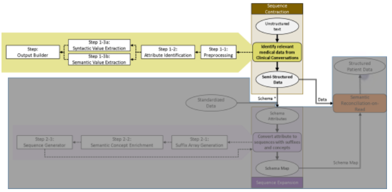
Solution 1-3: Value Extractor

Sequence Contraction

Concept Dictionary (x_p)



Output from Attribute Identifier
l
x
$\acute{S}$
RegEx groups corresponding to l
Regular Expression (x_r)
$x_r(\acute{S}) \rightarrow V_j$
l
V_j
Where the value of ***V*** depends on the x in the input
Output
l
V
Concept Dictionary (x_p)
$x_p(\acute{S}) \xrightarrow{l \in C} V_k$
l
V_k
Tokenize $\acute{S}$
Unigram & Bigram
Query UMLS
Collect Concept list (C) for token
UMLS Local Cache
UMLS

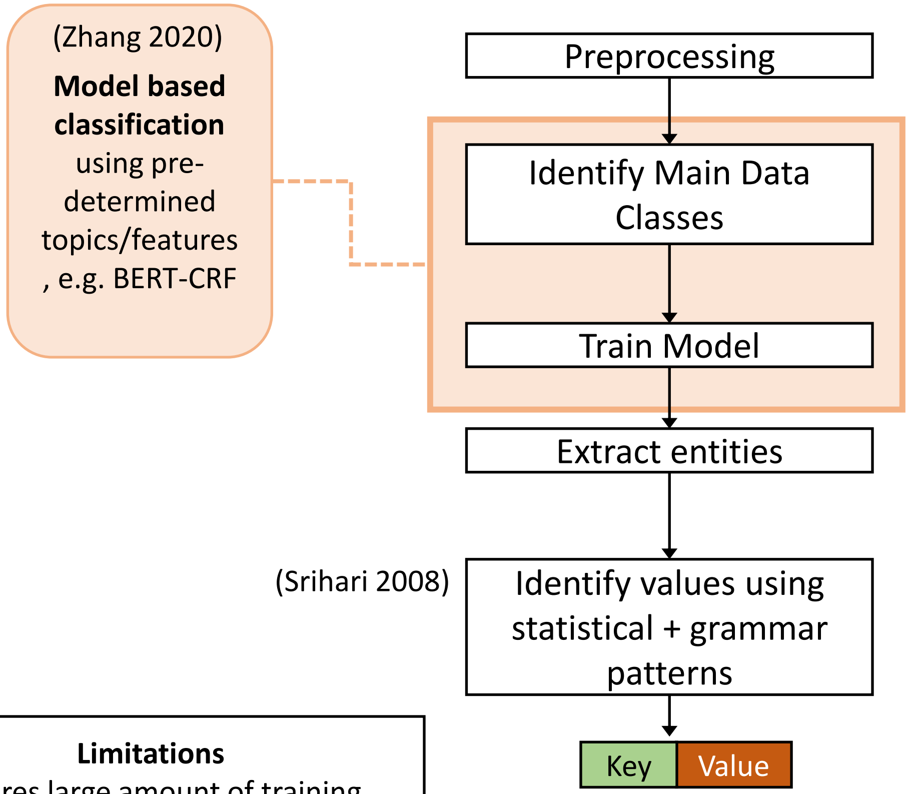


Solution 1: Sequence Contraction

As-is: To-be

Existing Solution

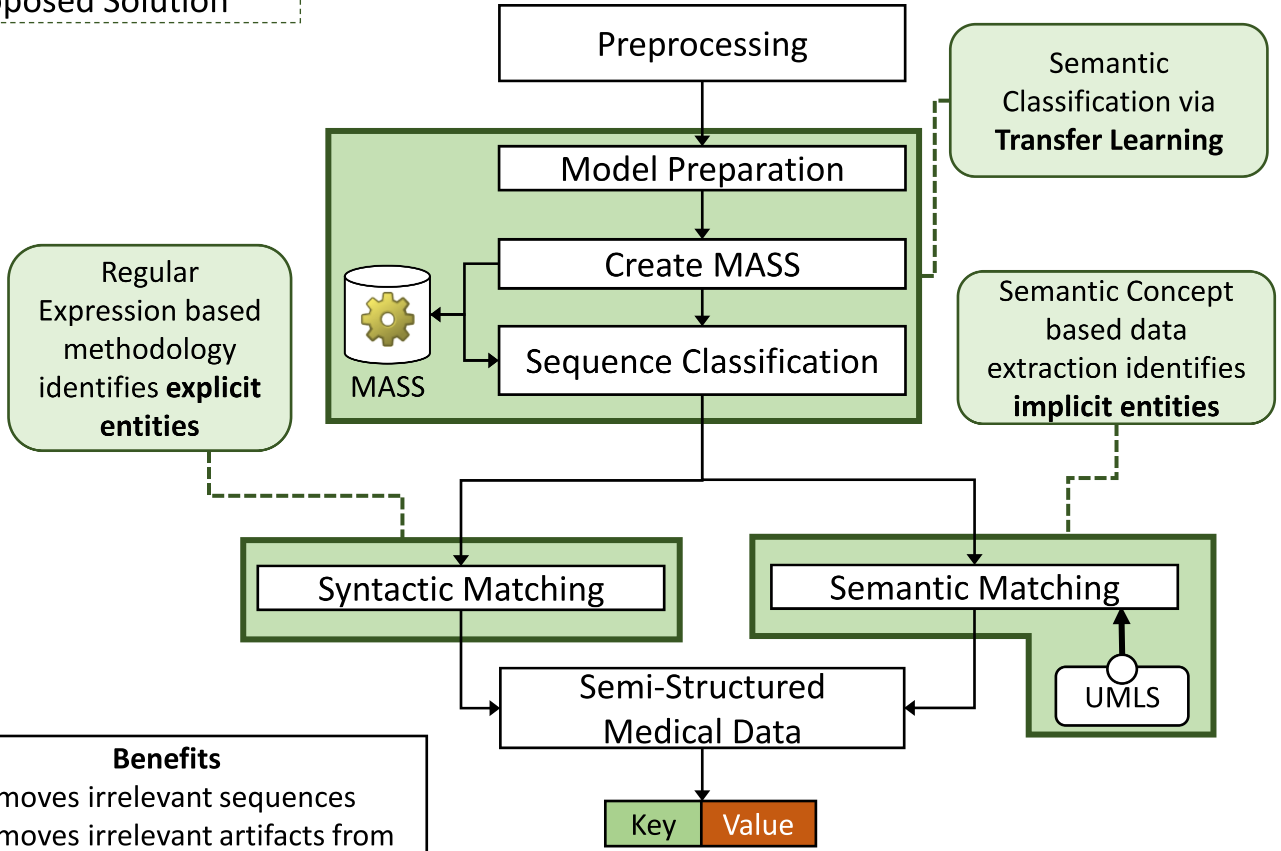
Proposed Solution



(Zhang 2020)
Model based classification using pre-determined topics/features, e.g. BERT-CRF
Preprocessing
Identify Main Data Classes
Train Model
Extract entities
(Srihari 2008)
Identify values using statistical + grammar patterns
Key
Value

Limitations

- requires large amount of training data
- Lacks generalization
- Positional semantics and feature recognition based matching only



Preprocessing
Semantic Classification via **Transfer Learning**
Model Preparation
Create MASS
MASS
Sequence Classification
Regular Expression based methodology identifies **explicit entities**
Semantic Concept based data extraction identifies **implicit entities**
Syntactic Matching
Semantic Matching
UMLS
Semi-Structured Medical Data
Key
Value

Benefits

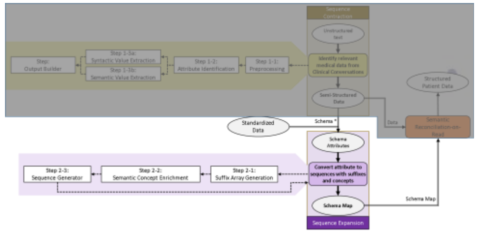
- Removes irrelevant sequences
- Removes irrelevant artifacts from sequences
- Identifies contextualized non-dictionary artifacts
- Model extension is easy to achieve

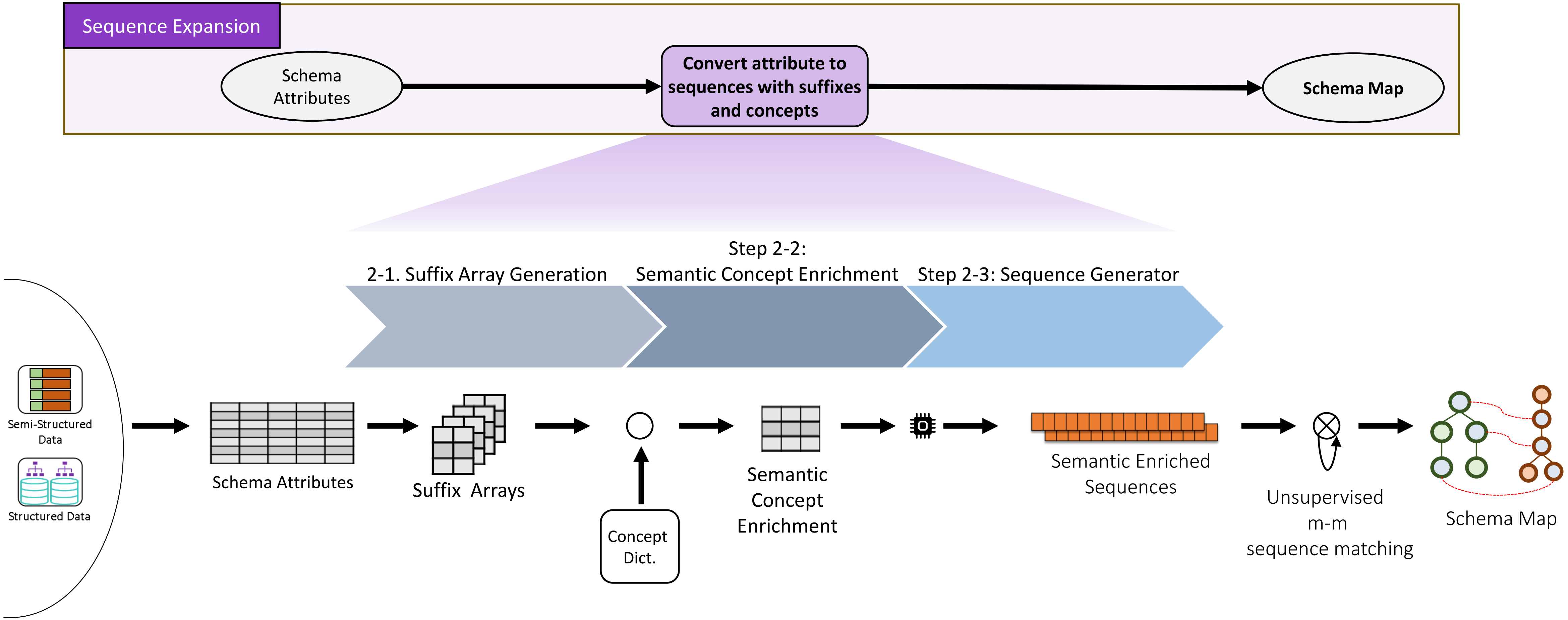
Limitations

- Querying UMLS is time consuming (bypassed by using cache)
- Results are dependent on the quality of instances in MASS

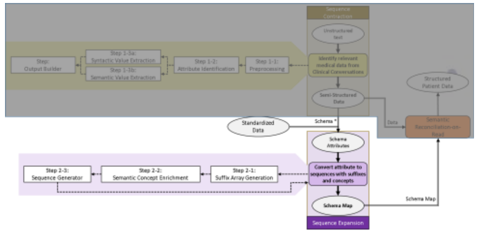
Solution 2: Sequence Expansion

Abstract View




Sequence Expansion
Schema Attributes
Convert attribute to sequences with suffixes and concepts
Schema Map
2-1. Suffix Array Generation
Step 2-2: Semantic Concept Enrichment
Step 2-3: Sequence Generator
Semi-Structured Data
Structured Data
Schema Attributes
Suffix Arrays
Concept Dict.
Semantic Concept Enrichment
Semantic Enriched Sequences
Unsupervised m-m sequence matching
Schema Map

Satti, Fahad Ahmed, et al. "Unsupervised Semantic Mapping for Healthcare Data Storage Schema." *IEEE Access* 9 (2021): 107267-107278.

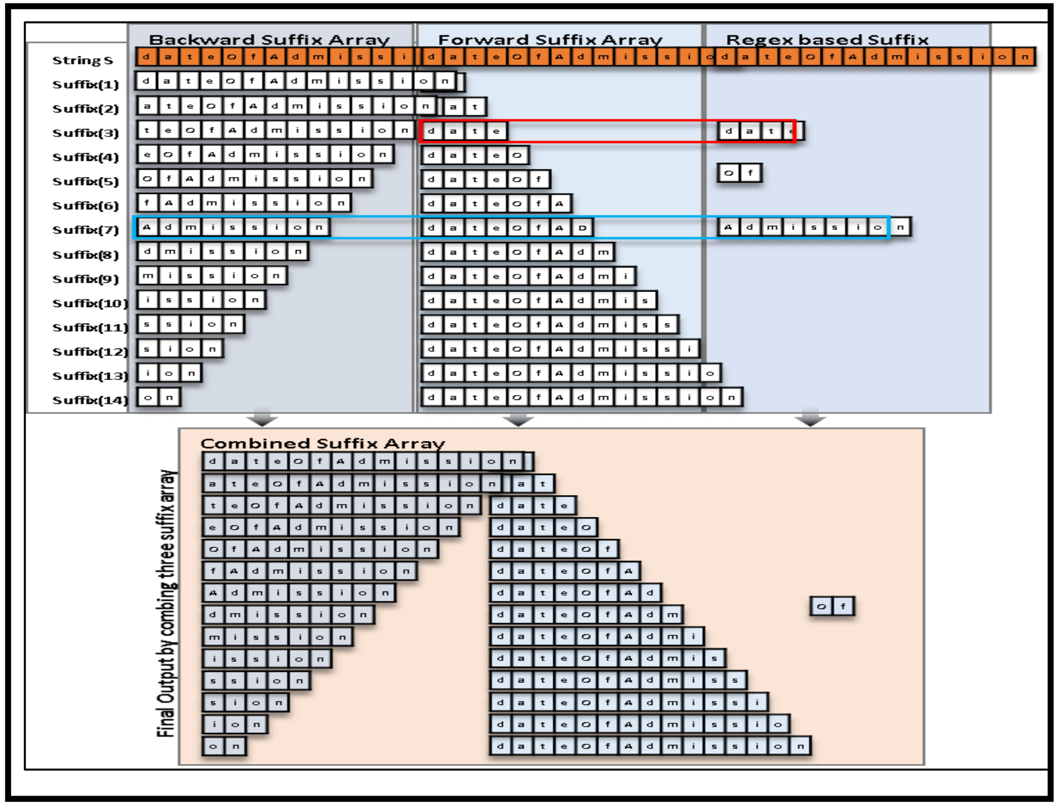

Step 2-3: Sequence Generator
Step 2-2: Semantic Concept Enrichment
Step 2-1: Suffix Array Generation
Sequence Expansion

Solution 2-1: Suffix Array Generation

Sequence Expansion

AIM
Identify the implicit words hidden in the attribute name

Benefits
• Utilizes Generalized Suffix Array; all suffixes for a set of string and is lexicographically sorted • lightweight in space • fast in practice

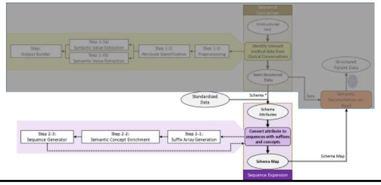


```
Algorithm 3 Suffix Array generation algorithm
Input
    T    token text
Output
    aa   Amplified Attribute
procedure BUILDSUFFIXARRAY($T, aa$)
    $suffixes$: TreeSet = $\{empty\}$
    $N \longleftarrow length(T)$
    $aa : AmplifiedAttribute = \{empty\}$
    for $i \longleftarrow [1, N]$ do
        $suffixes.add\,(token.substring\,([i, N)))$
    end for
    for $j \longleftarrow [1, N)$ do
        $suffixes.add\,(T.substring\,([0, j+1]))$
    end for
    $suffixes.addAll\,(T.split\,(REGEX\_WITH\_CASE))$
    $suffixArray$: HashSet < String > $\leftarrow suffixes$
    if $suffixArray \cdot length \leq 1$ then return
    end if
    $aa.setSuffixes(suffixArray)$
end procedure
```

Forward Suffix Array

Backward Suffix Array

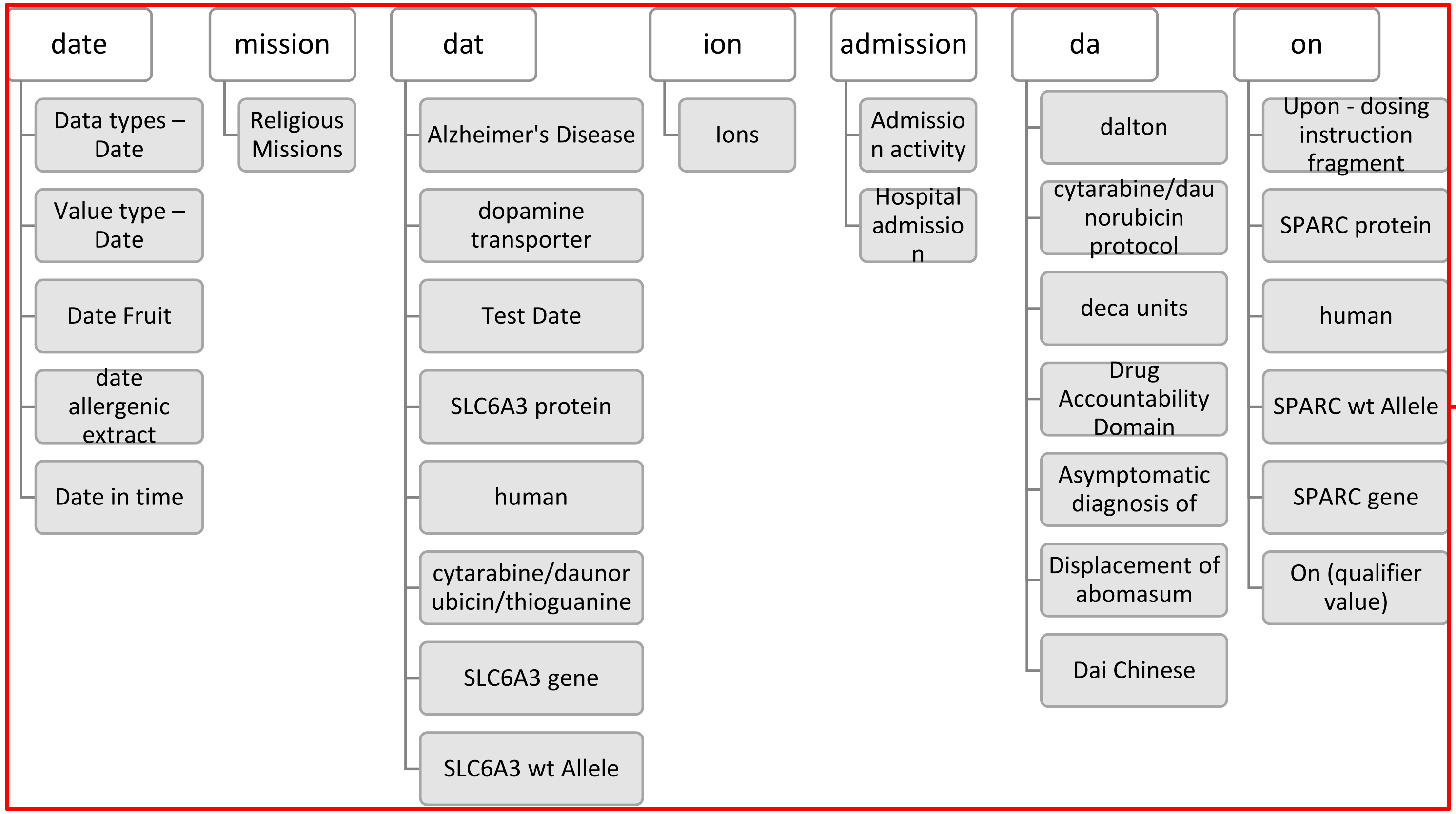
RegEx based Suffix Array



Solution 2-2: Semantic Concept Enrichment

Sequence Expansion

AIM	Benefits
For each suffix, identify the associated concepts	• Semantic matching can now take into account the concepts associated with each suffix



date
- Data types – Date
- Value type – Date
- Date Fruit
- date allergenic extract
- Date in time

mission
- Religious Missions

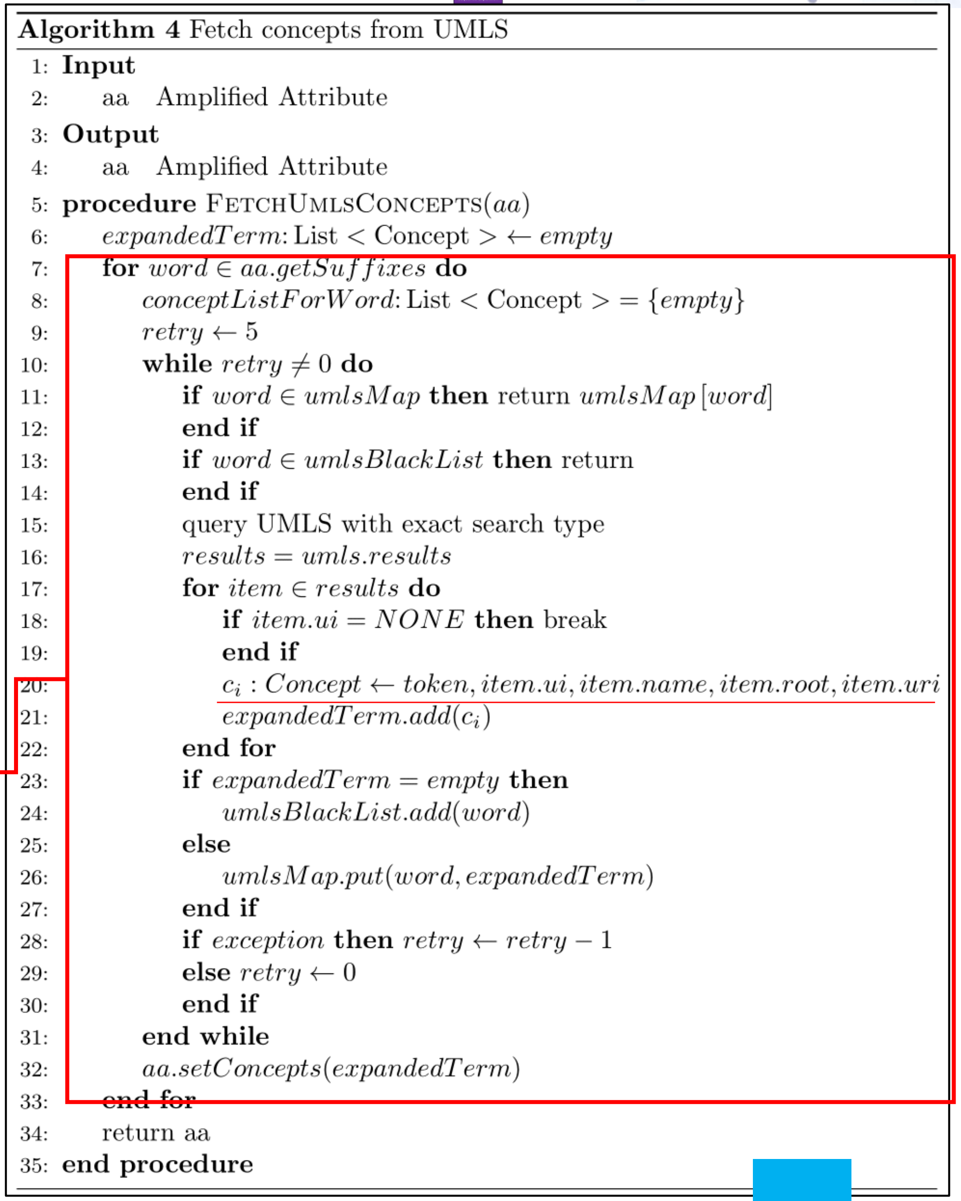
dat
- Alzheimer's Disease
- dopamine transporter
- Test Date
- SLC6A3 protein
- human
- cytarabine/daunorubicin/thioguanine
- SLC6A3 gene
- SLC6A3 wt Allele

ion
- Ions

admission
- Admission activity
- Hospital admission

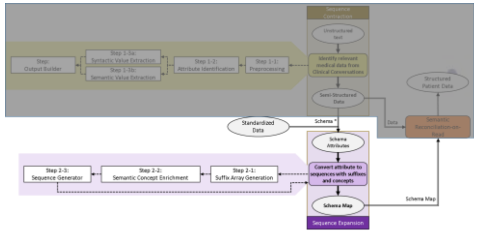
da
- dalton
- cytarabine/daunorubicin protocol
- deca units
- Drug Accountability Domain
- Asymptomatic diagnosis of
- Displacement of abomasum
- Dai Chinese

on
- Upon - dosing instruction fragment
- SPARC protein
- human
- SPARC wt Allele
- SPARC gene
- On (qualifier value)



Algorithm 4 Fetch concepts from UMLS

```
Input
    aa   Amplified Attribute
Output
    aa   Amplified Attribute
procedure FetchUmlsConcepts($aa$)
    $expandedTerm$: List < Concept > ← $empty$
    for $word \in aa.getSuffixes$ do
        $conceptListForWord$: List < Concept > = {$empty$}
        $retry$ ← 5
        while $retry \neq 0$ do
            if $word \in umlsMap$ then return $umlsMap[word]$
            end if
            if $word \in umlsBlackList$ then return
            end if
            query UMLS with exact search type
            $results = umls.results$
            for $item \in results$ do
                if $item.ui = NONE$ then break
                end if
                $c_i$ : $Concept$ ← $token, item.ui, item.name, item.root, item.uri$
                $expandedTerm.add(c_i)$
            end for
            if $expandedTerm = empty$ then
                $umlsBlackList.add(word)$
            else
                $umlsMap.put(word, expandedTerm)$
            end if
            if $exception$ then $retry$ ← $retry - 1$
            else $retry$ ← 0
            end if
        end while
        $aa.setConcepts(expandedTerm)$
    end for
    return aa
end procedure
```

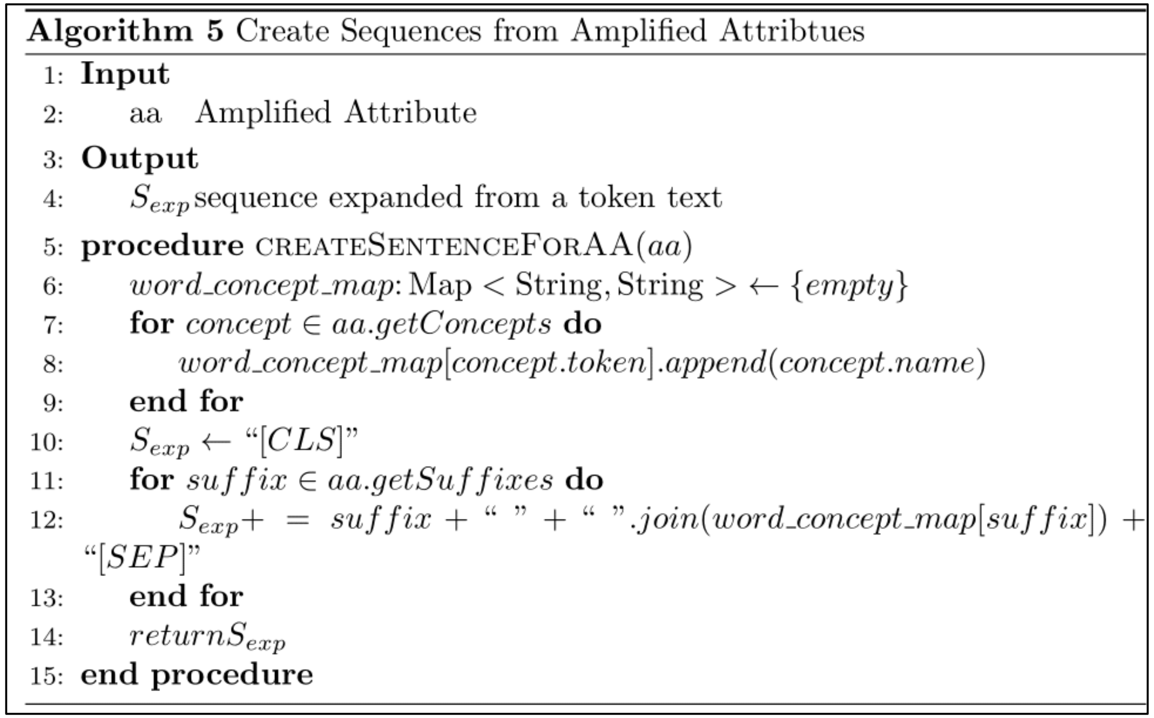

Sequence Contraction
Unstructured text
Identify relevant medical data from Clinical Conversations
Semi-Structured Data
Step: Output Builder
Step 1-3a: Syntactic Value Extraction
Step 1-3b: Semantic Value Extraction
Step 1-2: Attribute Identification
Step 1-1: Preprocessing
Structured Patient Data
Semantic Reconciliation-on-Read
Data
Standardized Data
Schema *
Schema Attributes
Convert attribute to sequences with suffixes and concepts
Schema Map
Sequence Expansion
Step 2-3: Sequence Generator
Step 2-2: Semantic Concept Enrichment
Step 2-1: Suffix Array Generation
Schema Map

Solution 2-3: Sequence Generator

Sequence Expansion

AIM	Benefits
Create sequences from the amplified	• Enriches the sequence of suffixes with their concepts

[CLS] **date** Data types - Date, Value type - Date, Date Fruit, date allergenic extract, Date in time;[SEP] **mission** Religious Missions;[SEP] **dat** Alzheimer's Disease, dopamine transporter, Test Date, SLC6A3 protein, human, cytarabine/daunorubicin/thioguanine, SLC6A3 gene, SLC6A3 wt Allele;[SEP] **ion** Ions;[SEP] **admission** Admission activity, Hospital admission;[SEP] **da** dalton, cytarabine/daunorubicin protocol, deca units, Drug Accountability Domain, Asymptomatic diagnosis of, Displacement of abomasum, Dai Chinese;[SEP] **on** Upon - dosing instruction fragment, SPARC protein, human, SPARC wt Allele, SPARC gene, On (qualifier value); [SEP]

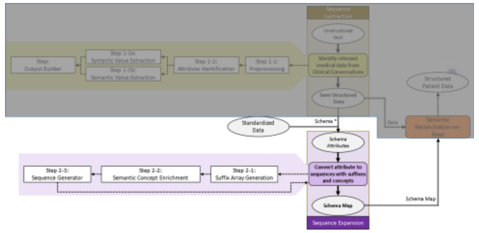


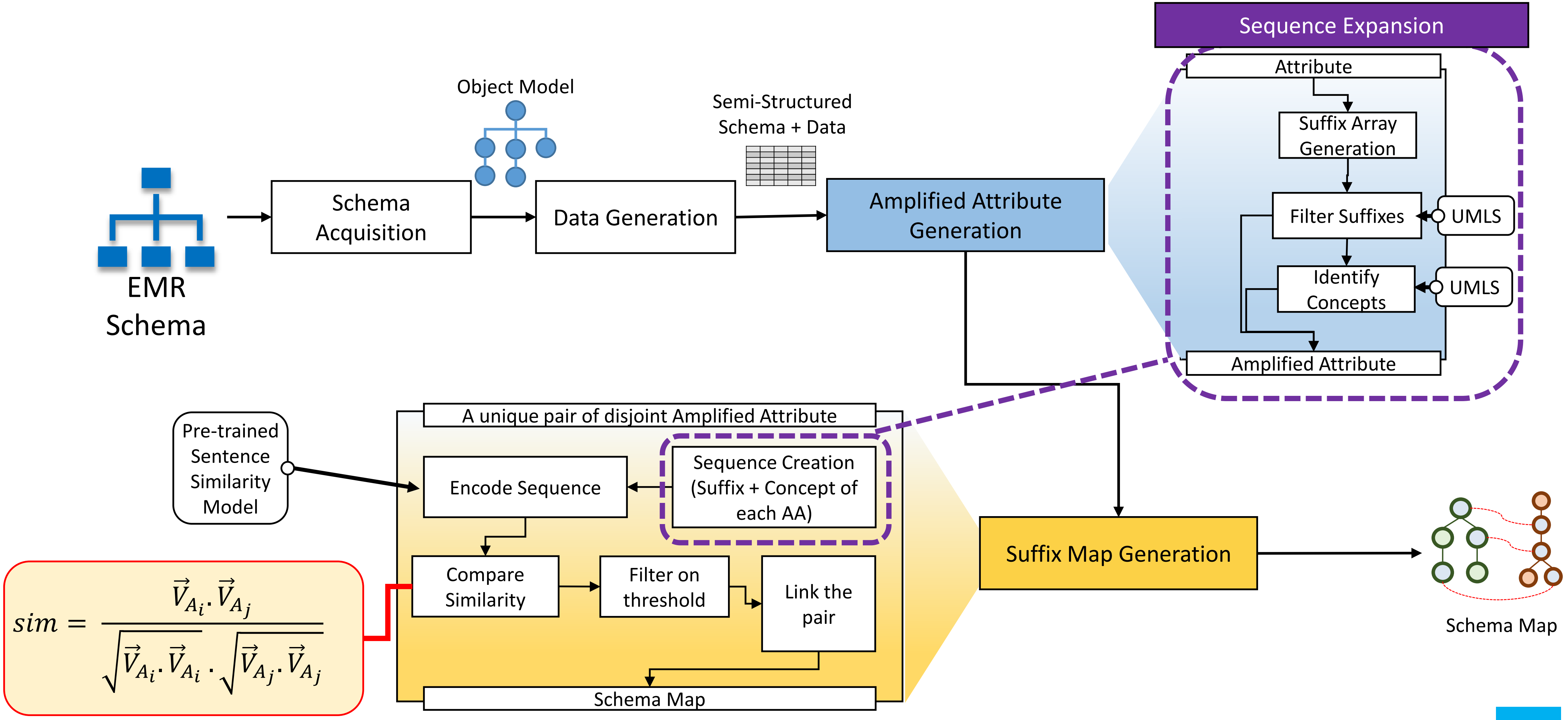
Algorithm 5 Create Sequences from Amplified Attribtues

```
Input
    aa    Amplified Attribute
Output
    $S_{exp}$ sequence expanded from a token text
procedure CREATESENTENCEFORAA($aa$)
    $word\_concept\_map$: Map < String, String > ← {$empty$}
    for $concept \in aa.getConcepts$ do
        $word\_concept\_map[concept.token].append(concept.name)$
    end for
    $S_{exp} \leftarrow$ “$[CLS]$”
    for $suffix \in aa.getSuffixes$ do
        $S_{exp} +\ = suffix +$ “ ” $+$ “ ”$.join(word\_concept\_map[suffix]) +$ “$[SEP]$”
    end for
    $return S_{exp}$
end procedure
```

Solution 2: Sequence Expansion

Detailed Workflow





EMR Schema

Schema Acquisition

Object Model

Data Generation

Semi-Structured Schema + Data

Amplified Attribute Generation

Sequence Expansion

Attribute

Suffix Array Generation

Filter Suffixes

UMLS

Identify Concepts

UMLS

Amplified Attribute

Pre-trained Sentence Similarity Model

A unique pair of disjoint Amplified Attribute

Encode Sequence

Sequence Creation (Suffix + Concept of each AA)

Compare Similarity

Filter on threshold

Link the pair

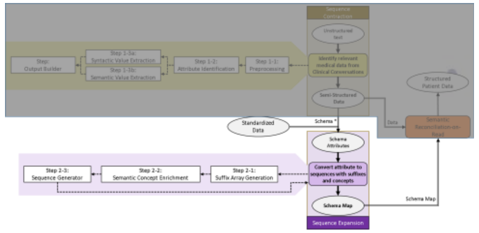
Schema Map

$$sim = \frac{\vec{V}_{A_i} . \vec{V}_{A_j}}{\sqrt{\vec{V}_{A_i} . \vec{V}_{A_i}} . \sqrt{\vec{V}_{A_j} . \vec{V}_{A_j}}}$$

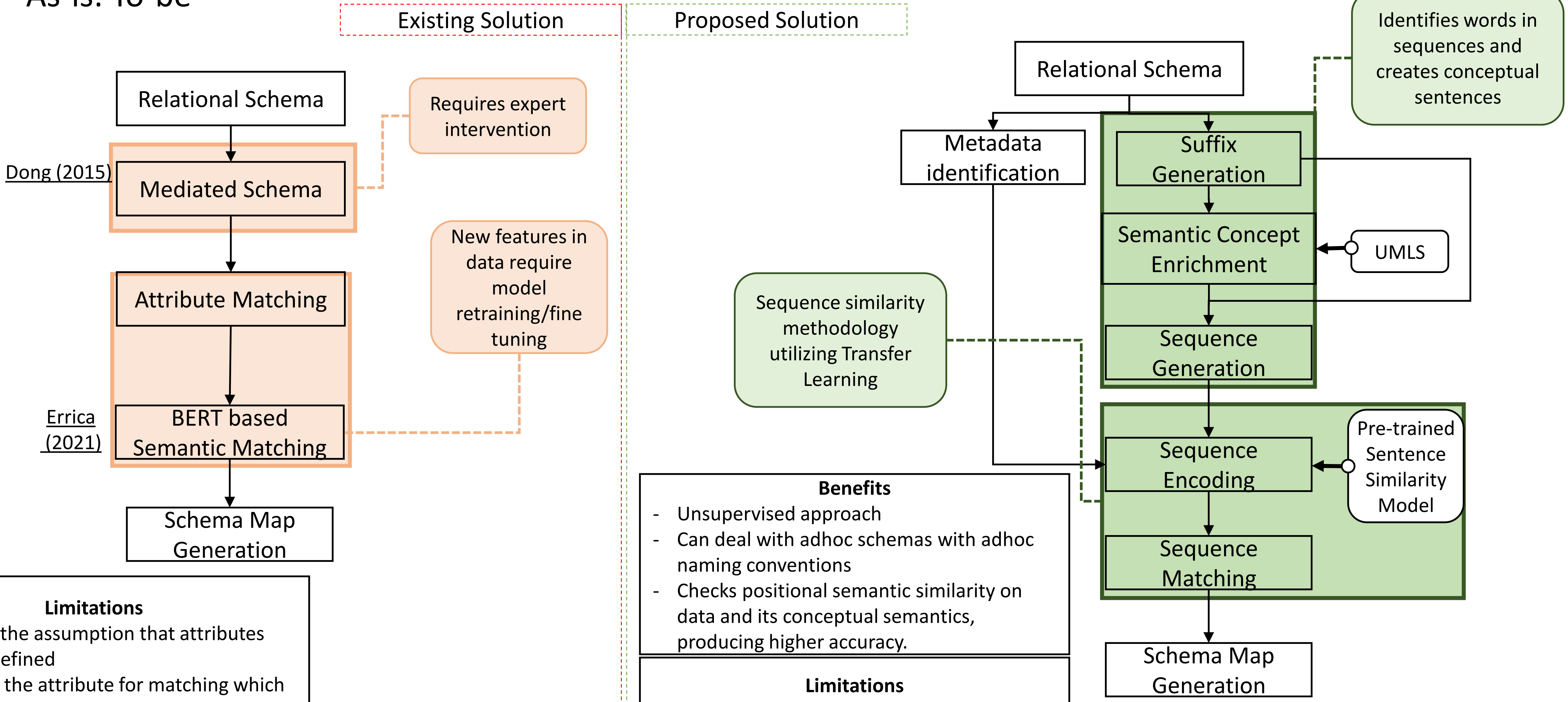
Suffix Map Generation

Schema Map

Solution 2: Sequence Expansion



As-is: To-be


Existing Solution
Relational Schema
Requires expert intervention
Dong (2015)
Mediated Schema
Attribute Matching
New features in data require model retraining/fine tuning
Errica (2021)
BERT based Semantic Matching
Schema Map Generation
Proposed Solution
Relational Schema
Identifies words in sequences and creates conceptual sentences
Metadata identification
Suffix Generation
Semantic Concept Enrichment
UMLS
Sequence Generation
Sequence similarity methodology utilizing Transfer Learning
Sequence Encoding
Pre-trained Sentence Similarity Model
Sequence Matching
Schema Map Generation

Limitations

- Based on the assumption that attributes are well defined
- Only uses the attribute for matching which produces lower accuracy

Benefits

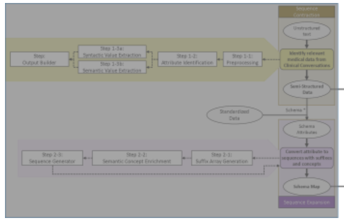
- Unsupervised approach
- Can deal with adhoc schemas with adhoc naming conventions
- Checks positional semantic similarity on data and its conceptual semantics, producing higher accuracy.

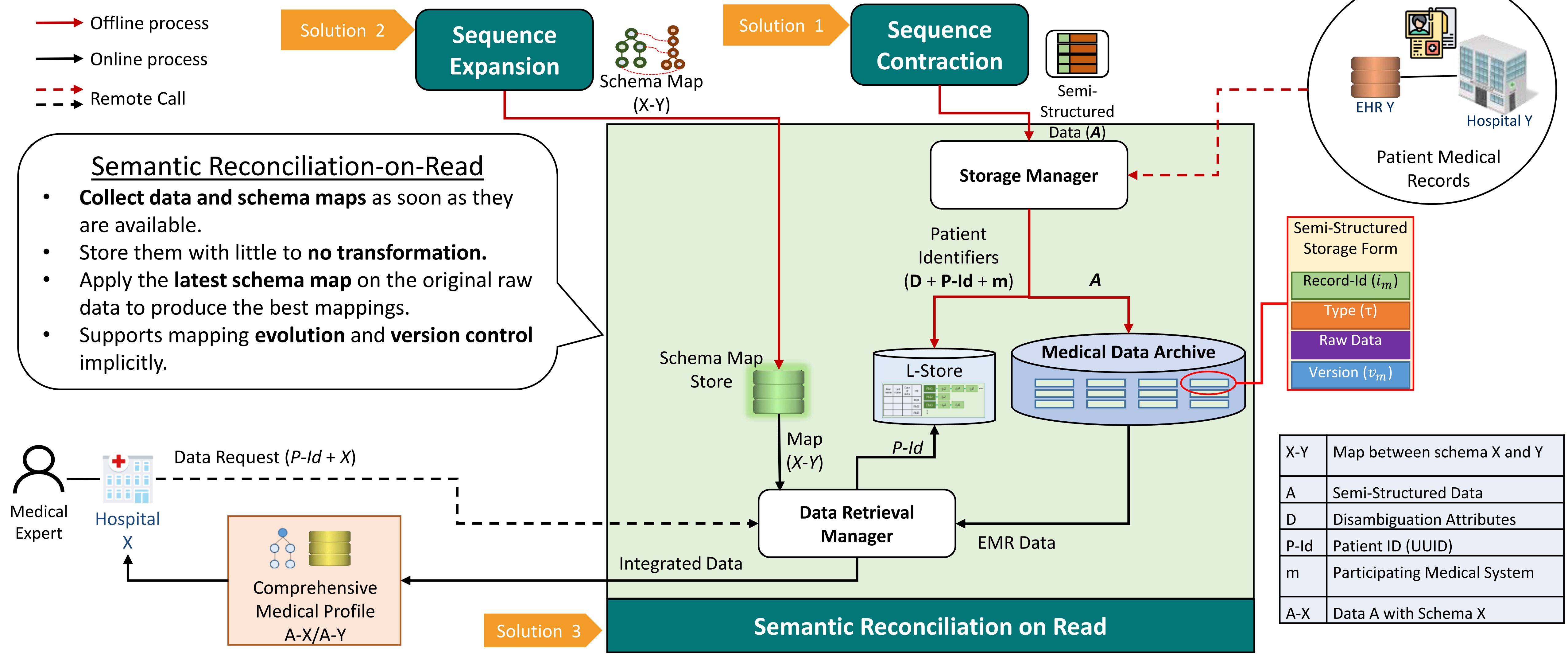
Limitations

- Unable to deal with abbreviations.

Solution 3: Semantic Reconciliation-on-Read

Workflow





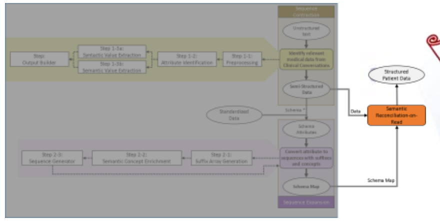
Semantic Reconciliation-on-Read

- **Collect data and schema maps** as soon as they are available.
- Store them with little to **no transformation.**
- Apply the **latest schema map** on the original raw data to produce the best mappings.
- Supports mapping **evolution** and **version control** implicitly.

X-Y	Map between schema X and Y
A	Semi-Structured Data
D	Disambiguation Attributes
P-Id	Patient ID (UUID)
m	Participating Medical System
A-X	Data A with Schema X

Satti, Fahad Ahmed, et al. "Ubiquitous Health Profile (UHPr): a big data curation platform for supporting health data interoperability." Computing 102.11 (2020): 2409-2444.

Solution 3: Semantic Reconciliation-on-Read

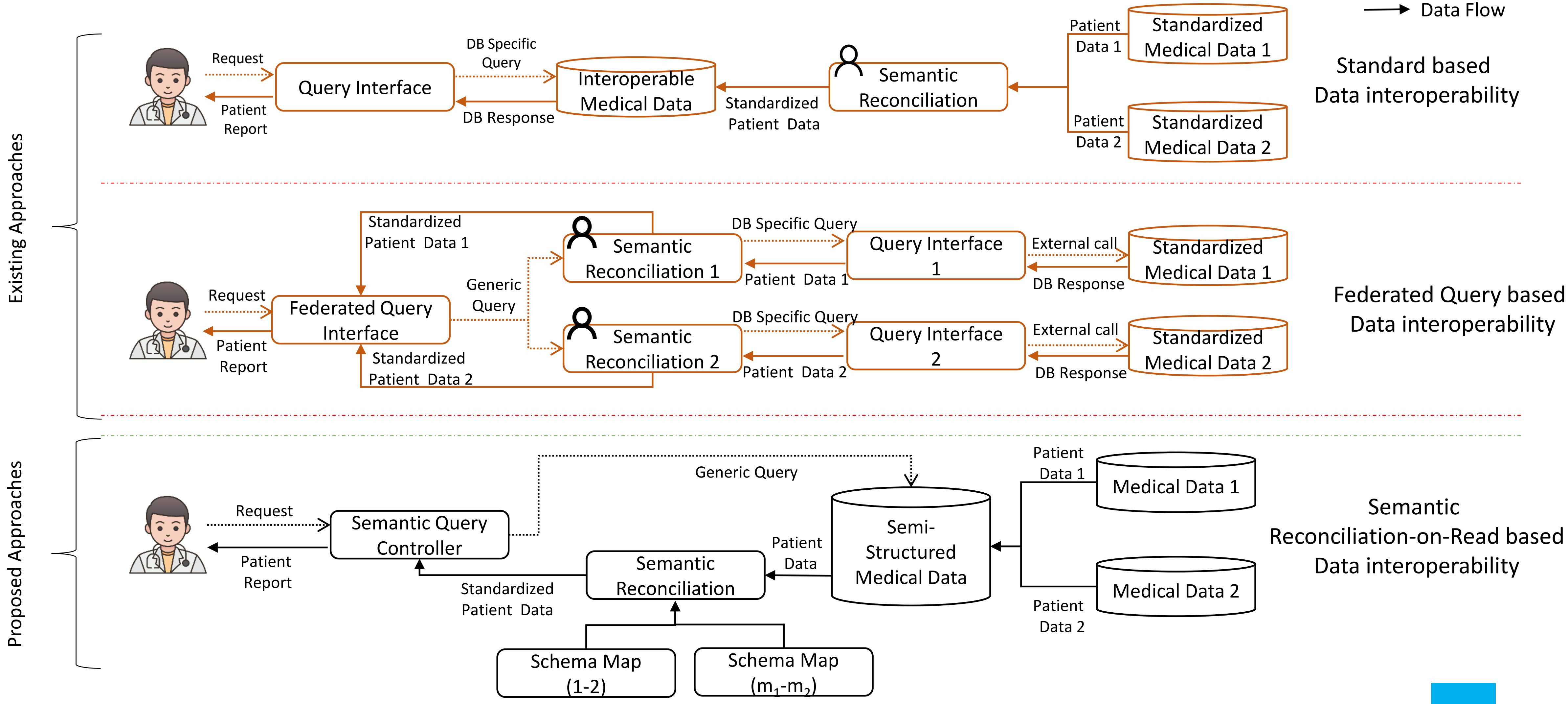


As-is: To-be

Expert Intervention

-----▸ Message Flow

⟶ Data Flow



Existing Approaches

Request

Query Interface

DB Specific Query

Interoperable Medical Data

Patient Report

DB Response

Standardized Patient Data

Semantic Reconciliation

Patient Data 1

Standardized Medical Data 1

Patient Data 2

Standardized Medical Data 2

Standard based Data interoperability

Standardized Patient Data 1

Request

Federated Query Interface

Patient Report

Generic Query

Semantic Reconciliation 1

Semantic Reconciliation 2

Standardized Patient Data 2

DB Specific Query

Query Interface 1

Patient Data 1

External call

DB Response

Standardized Medical Data 1

DB Specific Query

Query Interface 2

Patient Data 2

External call

DB Response

Standardized Medical Data 2

Federated Query based Data interoperability

Proposed Approaches

Request

Semantic Query Controller

Patient Report

Generic Query

Standardized Patient Data

Semantic Reconciliation

Patient Data

Semi-Structured Medical Data

Patient Data 1

Medical Data 1

Patient Data 2

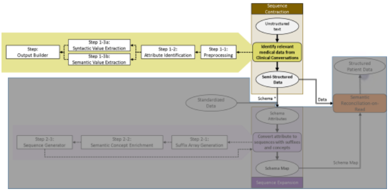
Medical Data 2

Schema Map (1-2)

Schema Map (m_1-m_2)

Semantic Reconciliation-on-Read based Data interoperability

RESULTS AND EVALUATION



Experimental Setup

Solution 1: Sequence Contraction – Performance

A sample of Test dataset

> And how old is he? 4 years;;**age:4 years**

> what happened to him?

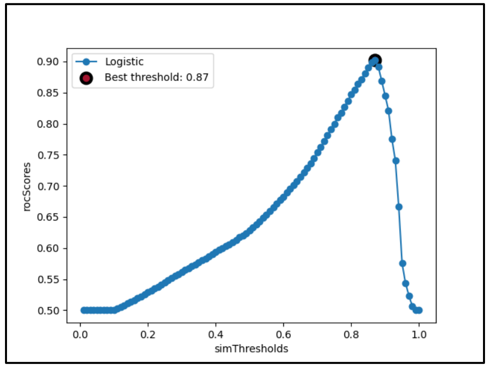
> he has fever;;**Finding:fever**

> also some serious cough;;**Sign or Symptom:cough**

1

Sequence Similarity

m



Threshold Selection

	True Positive	**True Negative**
Predicted Positive	199	88s
Predicted Negative	4769	437

A sample of MASS instances

what is child's name? hammad

***[CLS] what is child's name? [MASK];;name;;(.*)?what(.*)?name(.*?)\? ((his|her|patient)? name is)?(?P<Name>.*)**

how old is he? 5 years

***[CLS] how old is he? [MASK] years;;age;;(old|age)(.*)?\? (he is|she is|shes)?(?P<Age>.*)(years|month)?(.*)?**

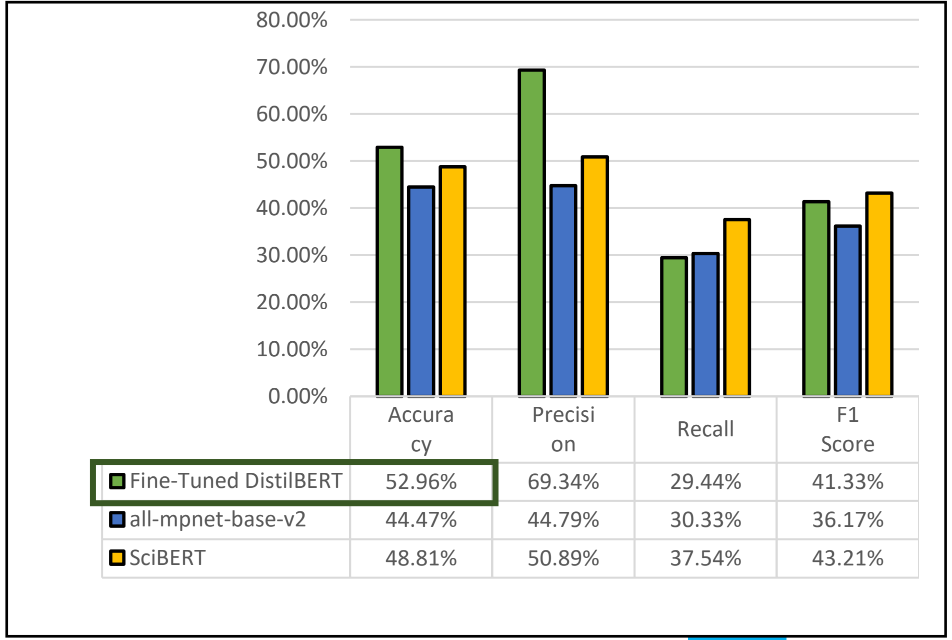
the child has cough

***[CLS] the child has [MASK];;Sign or Symptom;;umls**

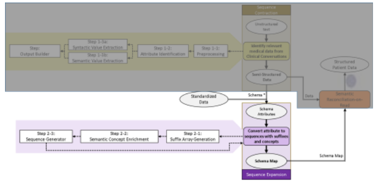
360 Instances

- 79 with 1 attribute
- 281 with multiple attributes

Comparison



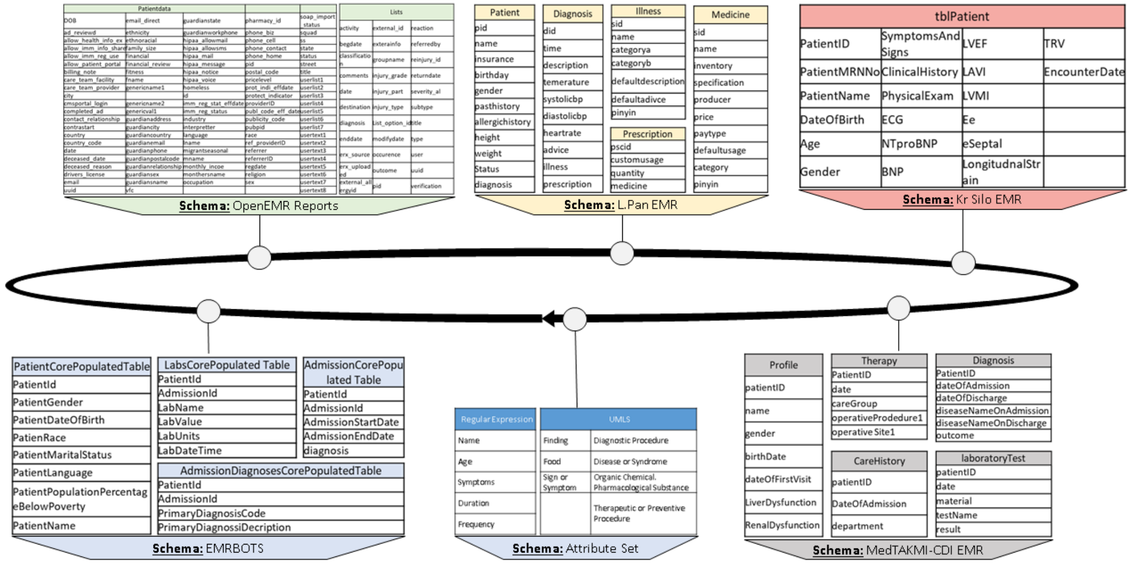
	Accuracy	Precision	Recall	F1 Score
🟩 Fine-Tuned DistilBERT	52.96%	69.34%	29.44%	41.33%
🟦 all-mpnet-base-v2	44.47%	44.79%	30.33%	36.17%
🟨 SciBERT	48.81%	50.89%	37.54%	43.21%



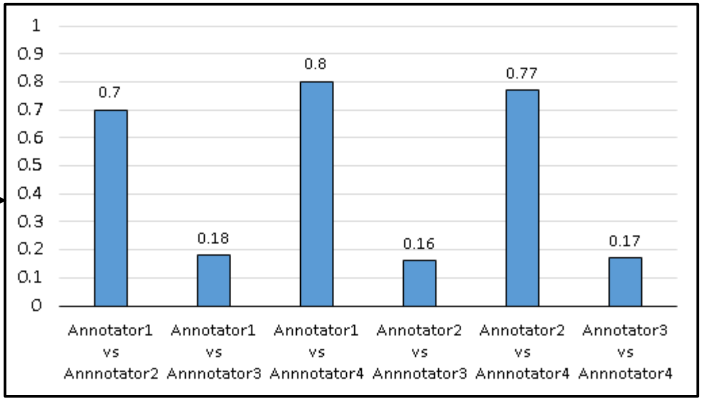
Experimental Setup

Solution 2: Sequence Expansion - Dataset

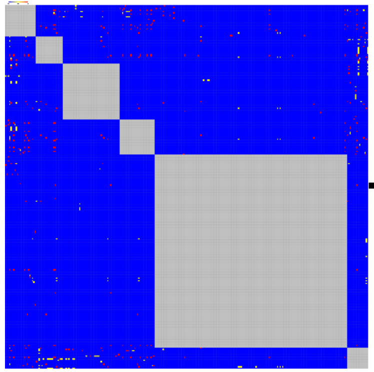
Statistics	
EMR Schema	6
Total Attributes	270
Comparisons	48,826
Annotators	4



Truth Set creation by expert intervention



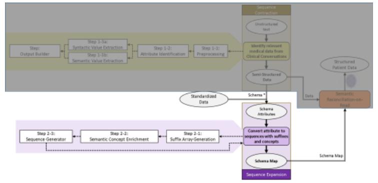
Cohen’s Kappa score among the four annotators



A sample of 2d sheet for annotators

	Emrbots_PatientCorePopulatedTable_PatientRace	**LPanEmr_Diagnosis_heartrate**
Openemr_Patientdata_ethnicity	1	0
unknown_UmlsTypes_DiagnosticProcedure	0	1

Annotator	**Total Matches**	**Marked as**			**Not marked**
		Equal	**Related**	**Unrelated**	
Annotator 1	48,826	326	65+150+10	48275	0
Annotator 2	48,826	329	36+171+25	48265	0
Annotator 3	48,826	348	1179+884+144	46118	153
Annotator 4	48,826	313	46+120+0	48336	11

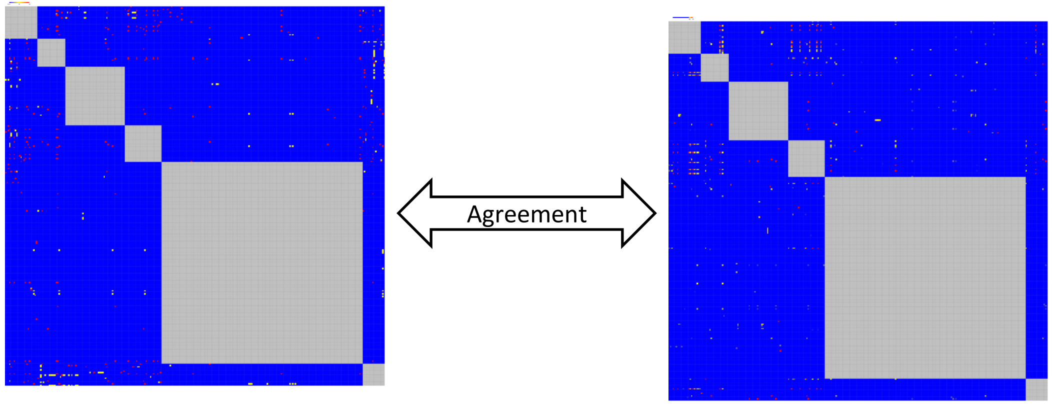


Experimental Setup

Solution 2: Sequence Expansion – Evaluation

Annotated Dataset

Computed Matches



Chicco (2020)

Evaluation Metric : MCC

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

TP = True Positive, TN = True Negative,
FP = False Positive, FN = False Negative

McHugh (2012)

Evaluation Metric : Kappa Score

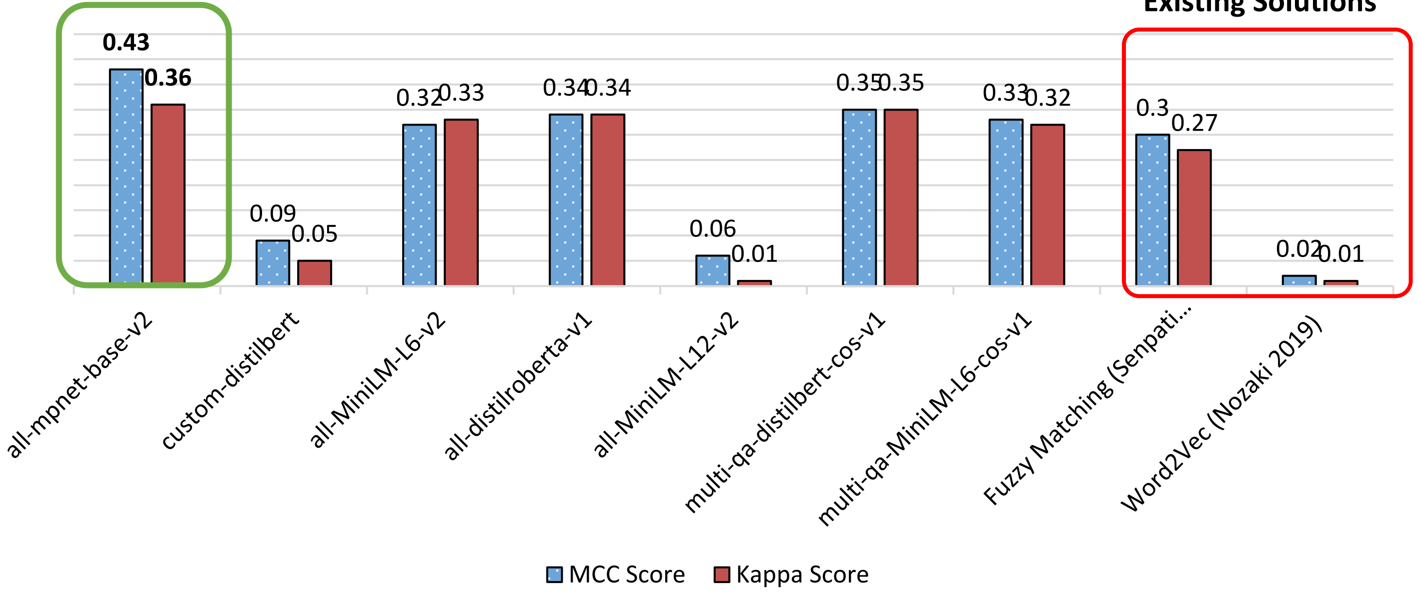
$$K = \frac{(P_o\ -\ P_e)}{(1 - P_e)} \rightarrow [-1,1]$$

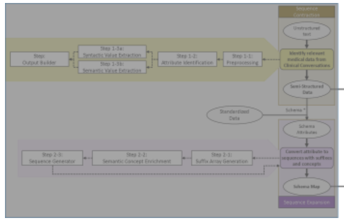
P_o = Empirical probability of agreement on the label assigned to any sample.

P_e = Expected agreement on when annotators assign labels randomly.

Proposed Model for Sequence Similarity

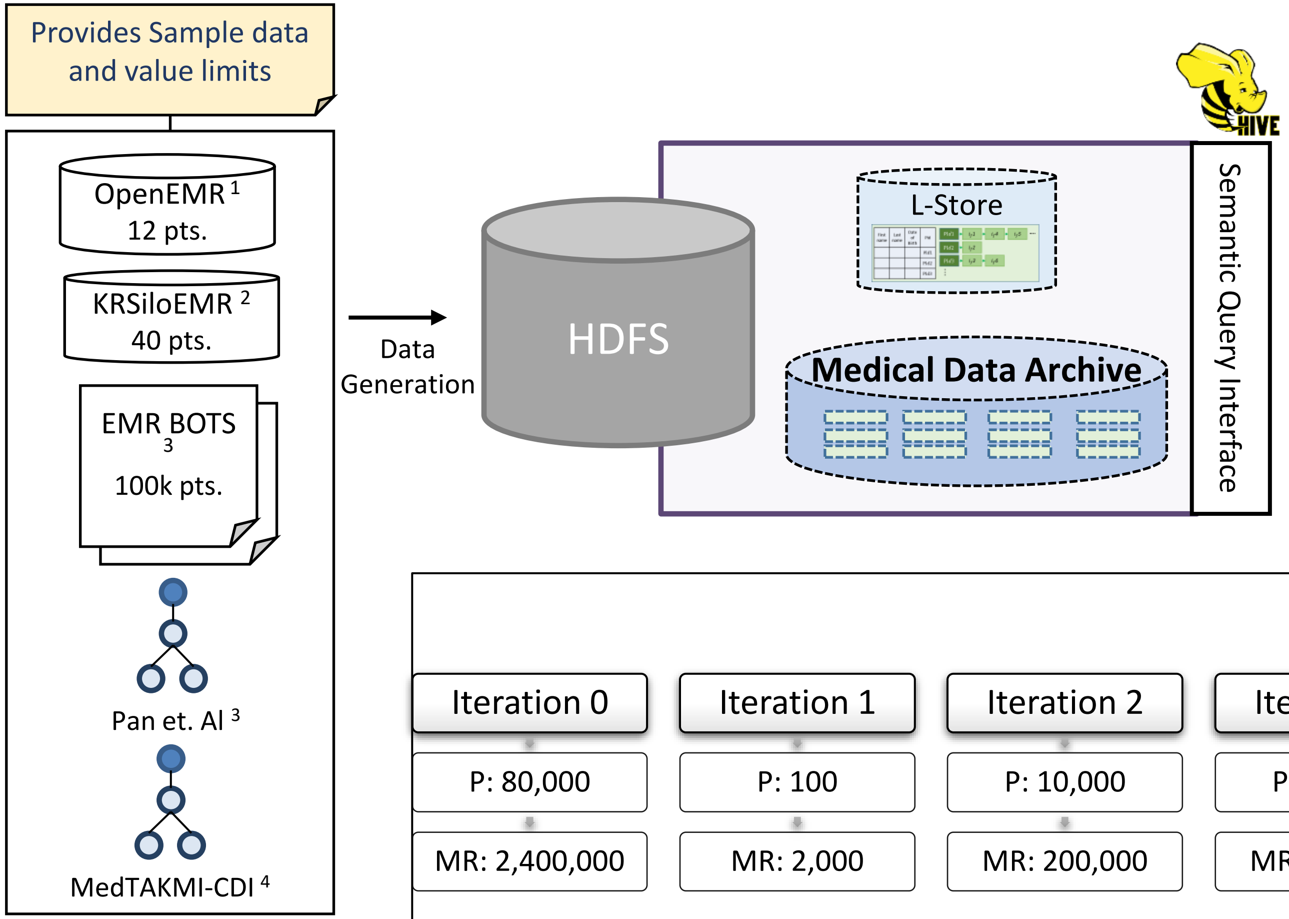
Existing Solutions





Experimental Setup

Solution 3: Semantic Reconciliation-on-Read



Provides Sample data and value limits

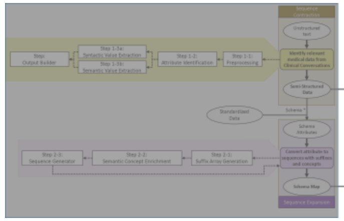
Data is loaded in 7 iterations to evaluate the scalability of the proposed data curation engine

Iteration 0	Iteration 1	Iteration 2	Iteration 3	Iteration 4	Iteration 5	Iteration 6	Iteration 7
P: 80,000	P: 100	P: 10,000	P: 40,000	P: 80,000	P: 80,000	P: 1	P: 100,000
MR: 2,400,000	MR: 2,000	MR: 200,000	MR: 800,000	MR: 2,400,000	MR: 2,400,000	MR: 40	MR: 107,535,388

Final Data Statistics	
Total P	390,101
Total MR	115,737,428

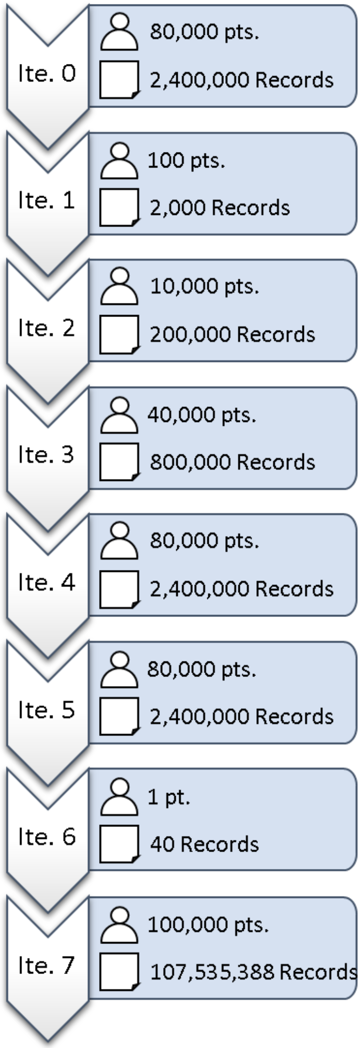
P -> Generated Patients
MR -> Generated Medical Records

1. OpenEMR
2. Ali (2017)
3. Kartoun (2016)
4. Pan (2016)
5. Akihiro (2007)

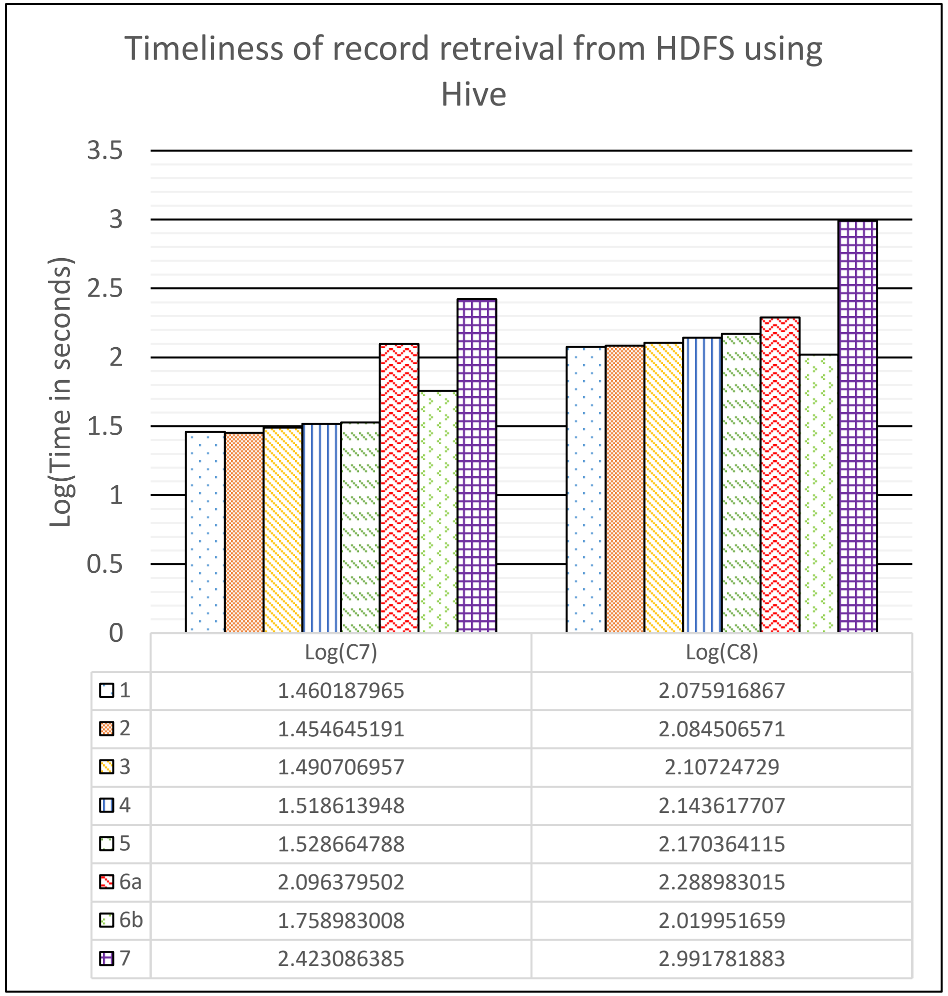


Experimental Setup

Solution 3: Semantic Reconciliation-on-Read

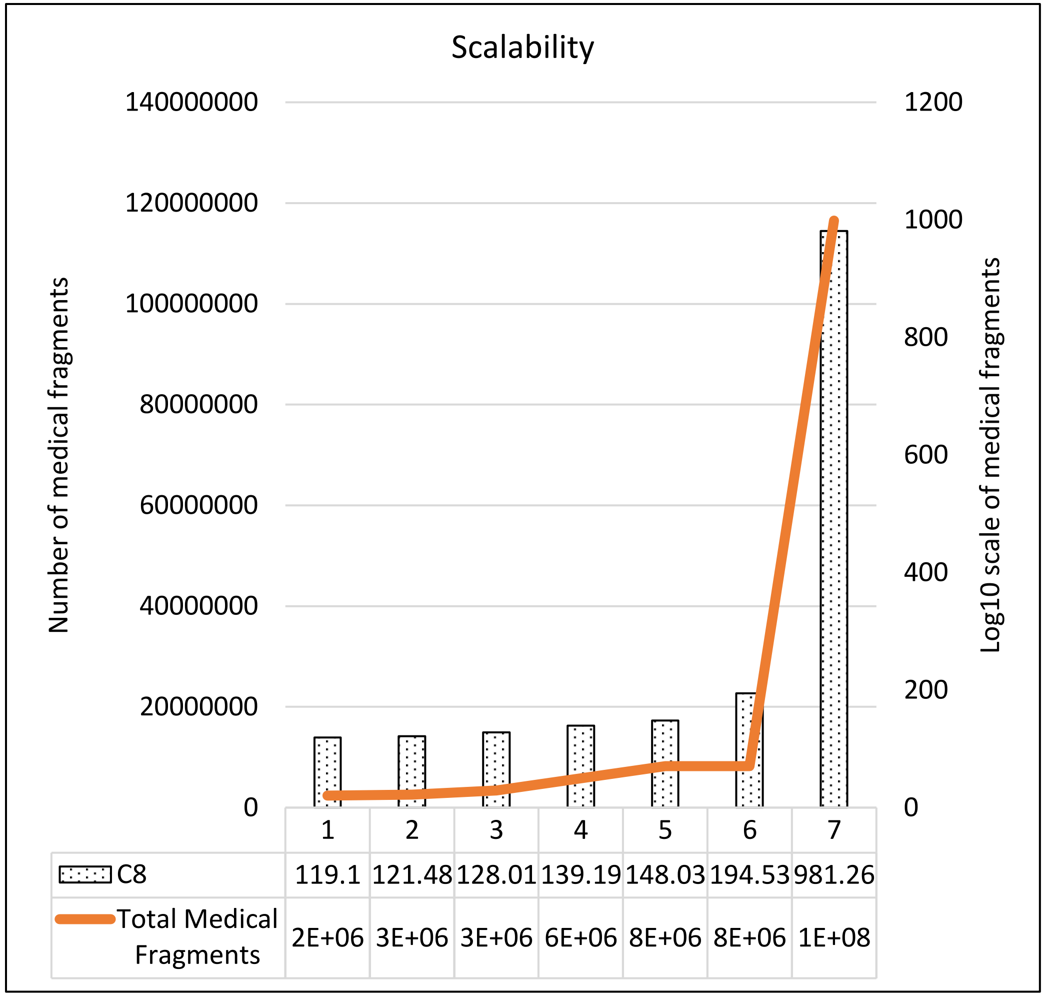


Ite. 0	80,000 pts.	2,400,000 Records
Ite. 1	100 pts.	2,000 Records
Ite. 2	10,000 pts.	200,000 Records
Ite. 3	40,000 pts.	800,000 Records
Ite. 4	80,000 pts.	2,400,000 Records
Ite. 5	80,000 pts.	2,400,000 Records
Ite. 6	1 pt.	40 Records
Ite. 7	100,000 pts.	107,535,388 Records



Timeliness of record retreival from HDFS using Hive

	Log(C7)	Log(C8)
□ 1	1.460187965	2.075916867
▨ 2	1.454645191	2.084506571
▨ 3	1.490706957	2.10724729
▥ 4	1.518613948	2.143617707
□ 5	1.528664788	2.170364115
▨ 6a	2.096379502	2.288983015
□ 6b	1.758983008	2.019951659
▦ 7	2.423086385	2.991781883



Scalability

	1	2	3	4	5	6	7
▒ C8	119.1	121.48	128.01	139.19	148.03	194.53	981.26
▬ Total Medical Fragments	2E+06	3E+06	3E+06	6E+06	8E+06	8E+06	1E+08

Conclusions and Future Works

Sequence Contraction

- Proposed an automatic, semantic similarity based mechanism to extract attribute-value pairs from unstructured data

Sequence Expansion

- Proposed a suffix array, and conceptual semantics based approach to identify the relevant parts of attribute names and used semantic similarity to align heterogeneous schemas.

Semantic Reconciliation-on-Read

- A Big Healthare Data curation engine to archive medical data and supports schema evolution to ensure original data remains available for a longer duration

Future Works

- The presented sequence contraction methodology can be further enhanced by increasing the sample instances in MASS.

Publications

- International Journals (8)
  - First Author: 1 (Minor Revision)
  - First Author: 2 Published
  - Co-author: 5 Published
- Local Journals (1)
  - Co-Author: 1 Published
- Conferences (8)
  - First Author International: 5
  - Local Conferences: 3
- Domestic Patents (1)
  - Registered: 1

Publication

Total Publications: 18

First Author Publications (International): 13

References

Oliver 2016	Müller, Oliver, et al. "Using text analytics to derive customer service management benefits from unstructured data." MIS Quarterly Executive 15.4 (2016): 243-258.
Nguyen 2019	Nguyen, D. Q., & Verspoor, K. (2019, April). End-to-end neural relation extraction using deep biaffine attention. In European conference on information retrieval (pp. 729-738). Springer, Cham.
Srihari 2008	Srihari, R. K., Li, W., Cornell, T., & Niu, C. (2008). Infoxtract: A customizable intermediate level information extraction engine. Natural Language Engineering, 14(1), 33-69.
Hara 2005	Hara, K., & Matsumoto, Y. (2005). Information Extraction and Sentence Classification applied to Clinical Trial MEDLINE Abstracts.
Candel 2022	Candel, C. J. F., Ruiz, D. S., & García-Molina, J. J. (2022). A unified metamodel for NoSQL and relational databases. Information Systems, 104, 101898.
Kartoun (2016)	Kartoun, U. (2016). A methodology to generate virtual patient repositories. arXiv preprint arXiv:1608.00570.
Pan (2016)	Pan, L., Fu, X., Cai, F., Meng, Y., & Zhang, C. (2016, October). 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) (pp. 38-41). IEEE.
Inokuchi (2007)	Inokuchi, A., Takeda, K., Inaoka, N., & Wakao, F. (2007). MedTAKMI-CDI: interactive knowledge discovery for clinical decision intelligence. IBM Systems Journal, 46(1), 115-133.
Ali (2017)	Ali, T., & Lee, S. (2017, July). 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) (pp. 2654-2657). IEEE.
Dong (2015)	Dong, X. L., & Srivastava, D. (2015). Schema Alignment. In Big Data Integration (pp. 31-61). Springer, Cham.
Shvaiko (2005)	Shvaiko, P., & Euzenat, J. (2005). A survey of schema-based matching approaches. In Journal on data semantics IV (pp. 146-171). Springer, Berlin, Heidelberg.
Errica (2021)	Errica, F., Silvestri, F., Edizel, B., Denoyer, L., Petroni, F., Plachouras, V., & Riedel, S. (2021, July). Concept matching for low-resource classification. In 2021 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.
OpenEMR	OpenEMR https://www.open-emr.org/
Ali (2017)	Ali, Taqdir, and Sungyoung Lee. "Reconciliation of SNOMED CT and domain clinical model for interoperable medical knowledge creation." 2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE, 2017.
Kartoun (2016)	Kartoun, Uri. "A methodology to generate virtual patient repositories." arXiv preprint arXiv:1608.00570 (2016).
Pan (2016)	Pan, Lijun, et al. "Design a novel electronic medical record system for regional clinics and health centers in China." 2016 2nd IEEE International Conference on Computer and Communications (ICCC). IEEE, 2016.
Akihiro (2007)	Inokuchi, Akihiro, et al. "MedTAKMI-CDI: interactive knowledge discovery for clinical decision intelligence." IBM Systems Journal 46.1 (2007): 115-133.
Chicco (2020)	Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC genomics*, *21*(1), 1-13.
McHugh (2012)	Mary L McHugh. Interrater reliability: the kappa statistic. Biochemia medica, 22(3):276–282, 2012.
Zhang (2020)	Zhang, Yuanzhe, et al. "MIE: A medical information extractor towards medical dialogues." *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020.

References

Lin (2019) Lin, Xinzhu, et al. "Enhancing dialogue symptom diagnosis with global attention and symptom graph." *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

Du (2019) Du, Nan, et al. "Extracting symptoms and their status from clinical conversations." *arXiv preprint arXiv:1906.02239* (2019).

Du (2020) Du, Ming, et al. "A unified framework for attribute extraction in electronic medical records." *2020 3rd International Conference on Algorithms, Computing and Artificial Intelligence*. 2020.

Nozaki (2019) Nozaki, Kenji, Teruhisa Hochin, and Hiroki Nomiya. "Semantic schema matching for string attribute with word vectors." *2019 6th International Conference on Computational Science/Intelligence and Applied Informatics (CSII)*. IEEE, 2019.

Yousfi (2020) Yousfi, A., El Yazidi, M. H., & Zellou, A. (2020). xmatcher: Matching extensible markup language schemas using semantic-based techniques. *International Journal of Advanced Computer Science and Applications*, *11*(8).

Bulygin (2018) Bulygin, L. (2018, October). Combining lexical and semantic similarity measures with machine learning approach for ontology and schema matching problem. In *Proc. Int. Conf. Data Anal. Manage. Data Intensive Domains (DAMDID/RCDL)* (pp. 245-249).

Kersloot (2020) Kersloot, M. G., van Putten, F. J., Abu-Hanna, A., Cornet, R., & Arts, D. L. (2020). Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies. *Journal of biomedical semantics*, *11*(1), 1-21.

HSB (2015) Meridou, Despina, et al. "An event-driven health service bus." Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare. 2015.

OBDA (2018) Zhang, Hansi, et al. "An ontology-guided semantic data integration framework to support integrative data analysis of cancer survival." BMC medical informatics and decision making 18.2 (2018): 129-147.

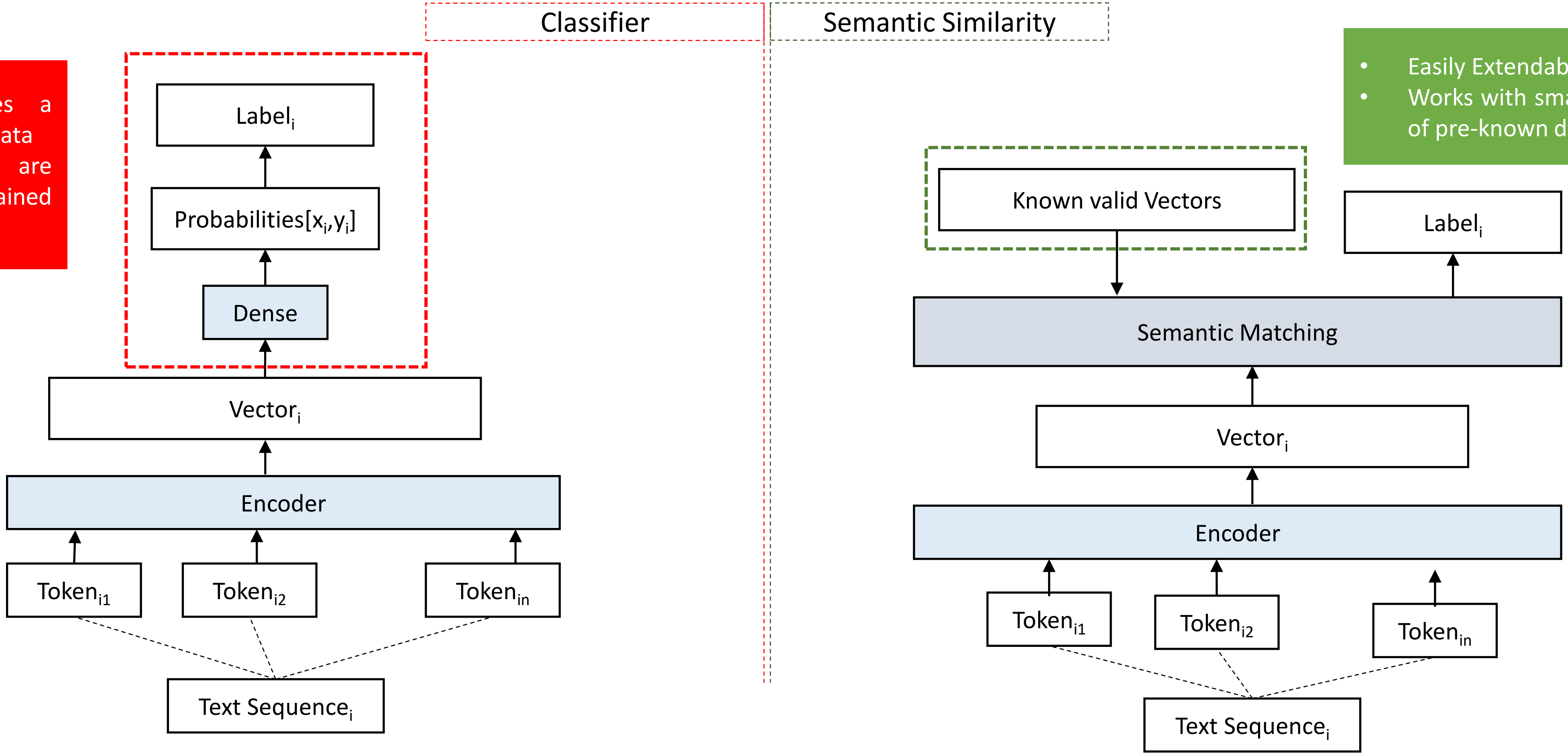
Senpati (2022) Senapati, A., Mondal, A., & Maji, S. (2022). A Fuzzy String Matching-Based Reduplication with Morphological Attributes. In *Pattern Recognition and Data Analysis with Applications* (pp. 183-190). Springer, Singapore.

Thank you

Comments & Suggestions

Prof. Seong Bae Park Comments

- **Why did you not use a classifier?**



- Training Requires a large amount of data
- Output labels are limited to the trained model

- Easily Extendable
- Works with small amount of pre-known data

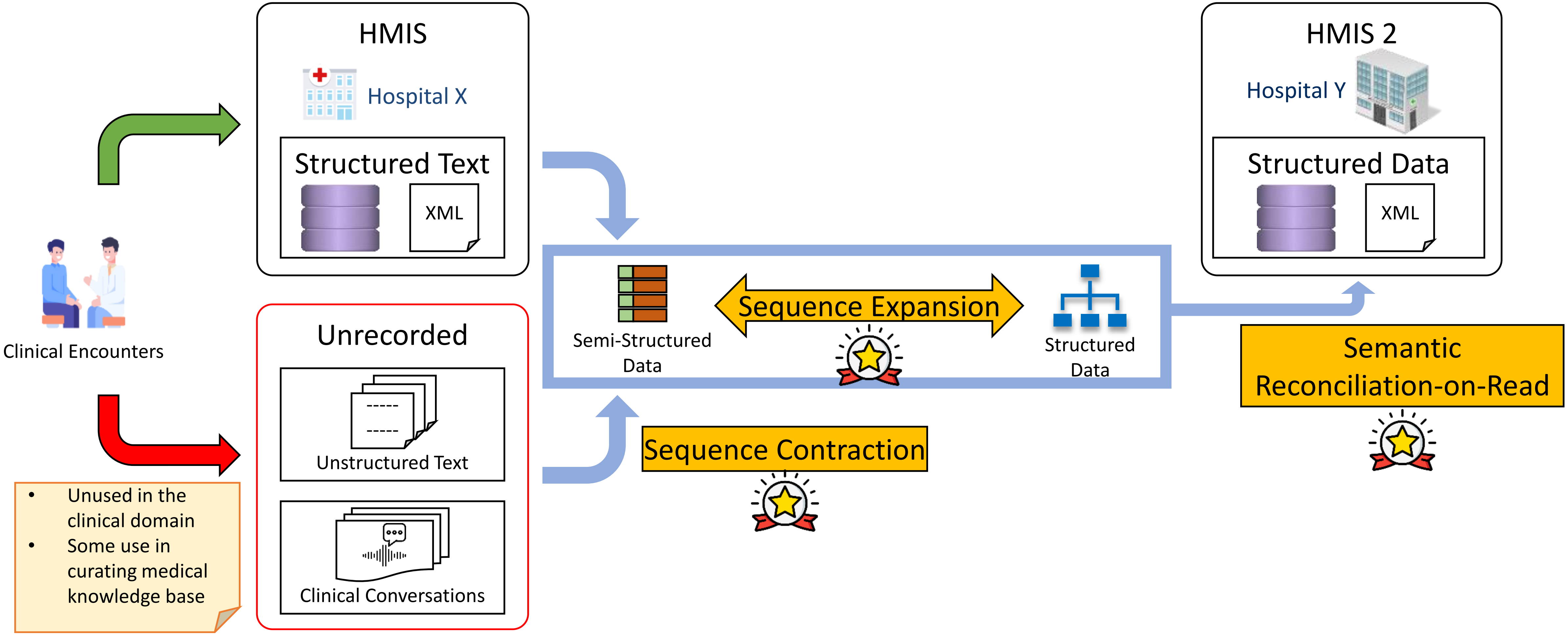
https://ai.googleblog.com/2018/05/advances-in-semantic-textual-similarity.html
https://www.kaggle.com/code/thanish/bert-for-token-classification-ner-tutorial/notebook

Prof. Eui Nam Huh Comments

Key Idea

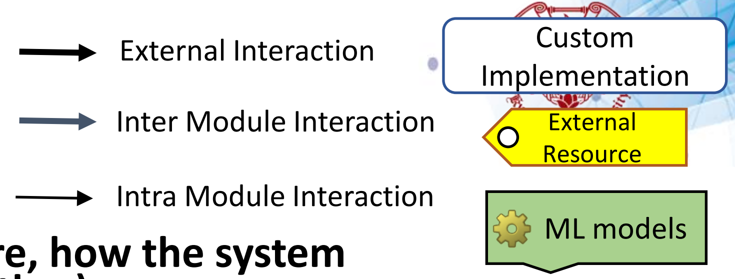
- Create Sequences from Unstructured text and attribute names
- Define a set of true sequences, enriched with semantic concepts
- Apply semantic similarity to classify unseen data
- Transform the classified instances into required results

- **Present an overall abstract architecture of the three solutions**

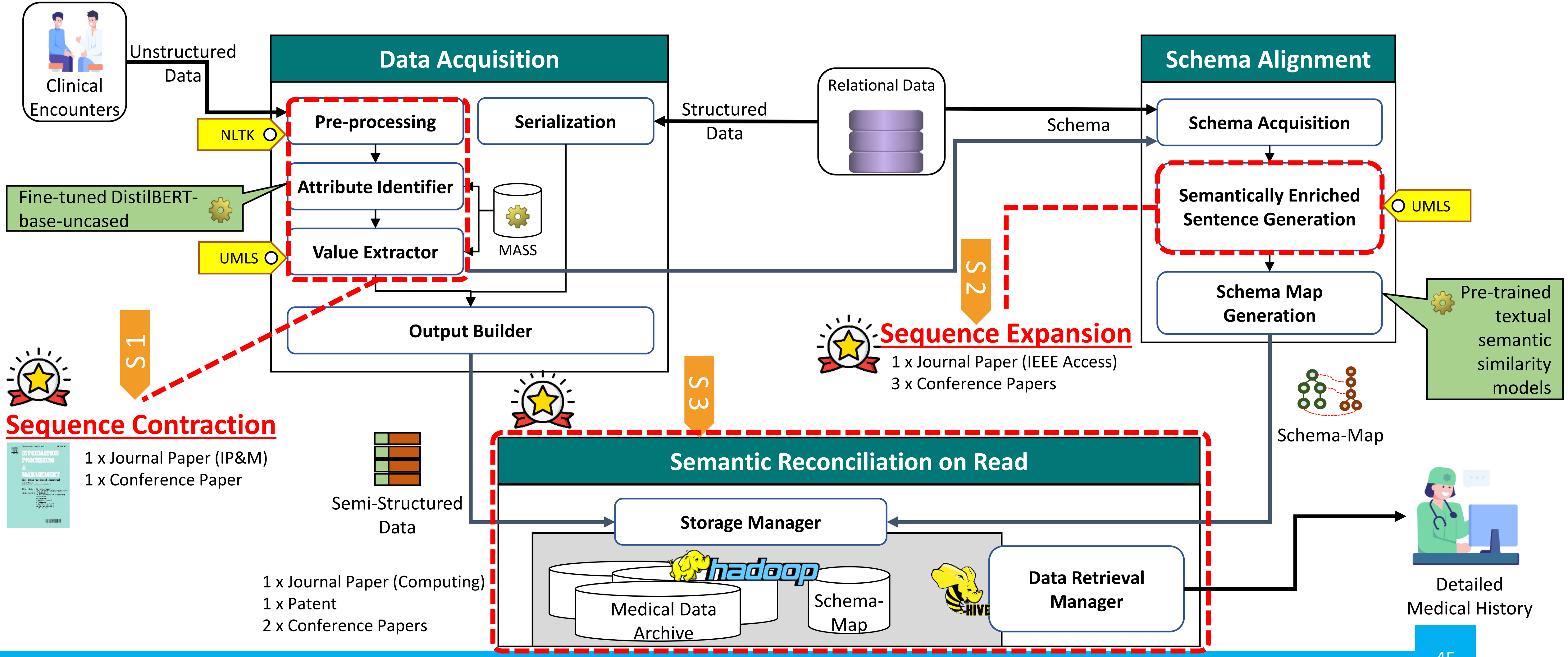


- Unused in the clinical domain
- Some use in curating medical knowledge base

Prof. Eui Nam Huh Comments



- **What kind of technologies are you using? (Focusing on high level architecture, how the system works, how data is collected, and how the components interact with each other)**



Sequence Contraction

1 x Journal Paper (IP&M)
1 x Conference Paper

Sequence Expansion

1 x Journal Paper (IEEE Access)
3 x Conference Papers

1 x Journal Paper (Computing)
1 x Patent
2 x Conference Papers

Prof. Eui Nam Huh Comments

- **Why did you use Generalized Suffix Array?**

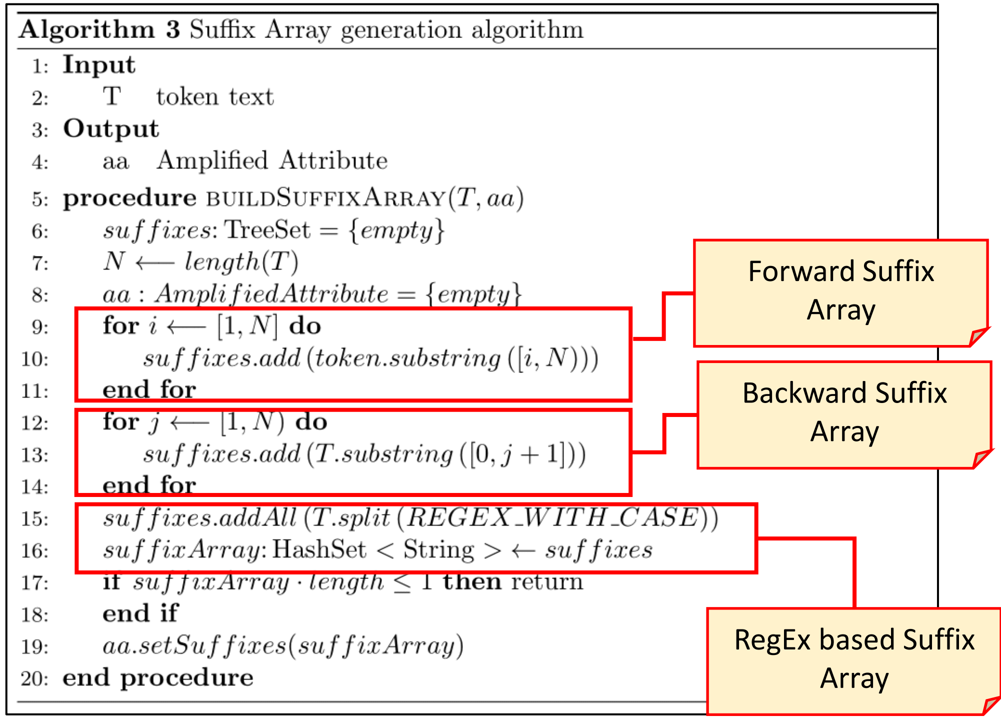
Generalized Suffix Array:
A generalized suffix array (or GSA), is a suffix array that contains all suffixes for a set of strings (for example, $S = S_1, S_2, S_3, \ldots$) and is lexicographically sorted with all suffixes of each string.

The method to produce suffix array here is equivalent to the following:
S1 = dateOfAdmission
S2 = reverseElements(suffixArray(S1)) = [’on’,’ion’,’sion’…..]
S3 = [‘date’,’Of’,’Admission’]

S = [Admission, Of, OfAdmission, ateOfAdmission, da, dat, date, dateO, dateOf, dateOfA, dateOfAd, dateOfAdm, dateOfAdmi, dateOfAdmis, dateOfAdmiss, dateOfAdmissi, dateOfAdmissio, dateOfAdmission, dmission, eOfAdmission, fAdmission, ion, ission, mission, on, sion, ssion, teOfAdmission]

Typical GSA Usecases: pattern matching, longest common subsequence problem, longest previous factor array (for text compression and detection of motifs and repeats)

Significance in Solution 2: Use of suffix array built using only one methodology, such as Forward pass, backward pass, or regex based one, would not capture all possible suffixes which may pertain to a medical concept. Thus we use a combination of these three algorithms.



Algorithm 3 Suffix Array generation algorithm

```
Input
    T    token text
Output
    aa   Amplified Attribute
procedure BUILDSUFFIXARRAY($T, aa$)
    $suffixes$: TreeSet = {$empty$}
    $N \longleftarrow length(T)$
    $aa : AmplifiedAttribute = \{empty\}$
    for $i \longleftarrow [1, N]$ do
        $suffixes.add(token.substring([i, N)))$
    end for
    for $j \longleftarrow [1, N)$ do
        $suffixes.add(T.substring([0, j + 1]))$
    end for
    $suffixes.addAll(T.split(REGEX\_WITH\_CASE))$
    $suffixArray$: HashSet < String > $\leftarrow suffixes$
    if $suffixArray \cdot length \leq 1$ then return
    end if
    $aa.setSuffixes(suffixArray)$
end procedure
```

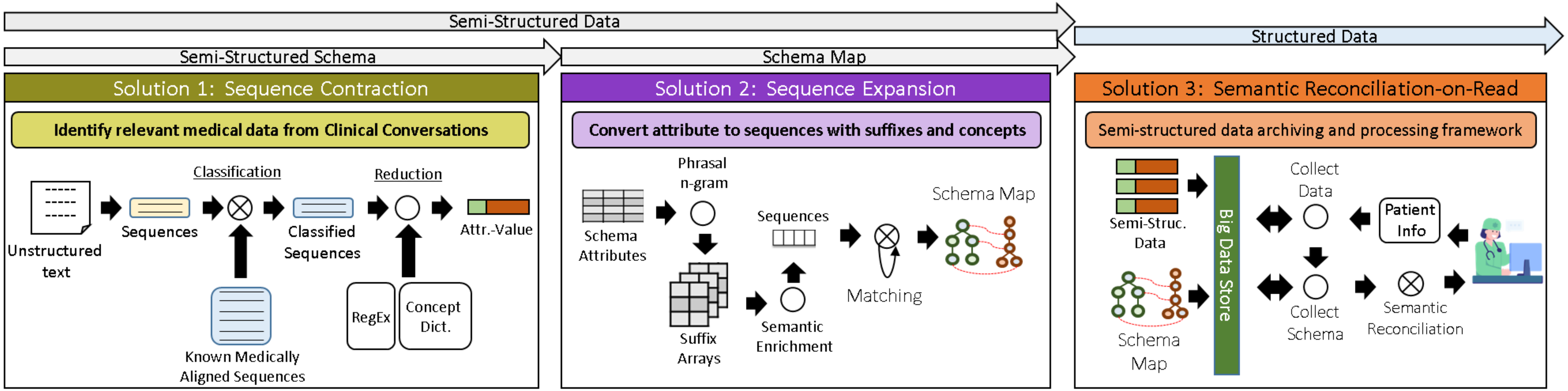
Forward Suffix Array

Backward Suffix Array

RegEx based Suffix Array

Prof. Eui Nam Huh Comments

- **What are your significant contributions? Focusing on strong argument for each motivation and how it is different from others.**



Semi-Structured Data

Structured Data

Semi-Structured Schema

Schema Map

Solution 1: Sequence Contraction

Identify relevant medical data from Clinical Conversations

Unstructured text

Sequences

Classification

Classified Sequences

Known Medically Aligned Sequences

Reduction

RegEx

Concept Dict.

Attr.-Value

Solution 2: Sequence Expansion

Convert attribute to sequences with suffixes and concepts

Schema Attributes

Phrasal n-gram

Suffix Arrays

Sequences

Semantic Enrichment

Matching

Schema Map

Solution 3: Semantic Reconciliation-on-Read

Semi-structured data archiving and processing framework

Semi-Struc. Data

Schema Map

Big Data Store

Collect Data

Collect Schema

Patient Info

Semantic Reconciliation

Novelty

- Semantic Similarity based Classification
  - Transfer Learning to classify sequences
  - Easily extendable
  - Less training requirements
- Reduction
  - syntactic and semantic extractors for creating attribute-value pairs.
  - Captures syntactic artifacts like name, age, etc.
  - Utilizes conceptual semantics to enrich reduction

Limitation

- Requires expert intervention to build the set of known medically aligned sequences and Regular expressions

Novelty

- Phrasal n-gram
  - Identify hidden words within attribute names
  - Handles adhoc naming conventions
- Semantic Enrichment
  - Utilizes semantic concepts for enriched matching
- Matching
  - Unsupervised
  - m-m matching between enriched sequences

Limitation

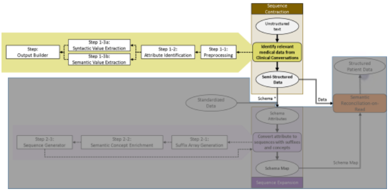
- No simple pathway for attributes with atomic names

Novelty

- Big Data Store
  - Low resource requirements at the end-nodes (hospitals/clinics)
  - Data archiving to prevent data loss
  - Evolvable schema-maps
- On demand transformation
  - Conversion to any standard
  - When required, utilize latest schema-map
  - Supports 1-1 and 1-m mappings

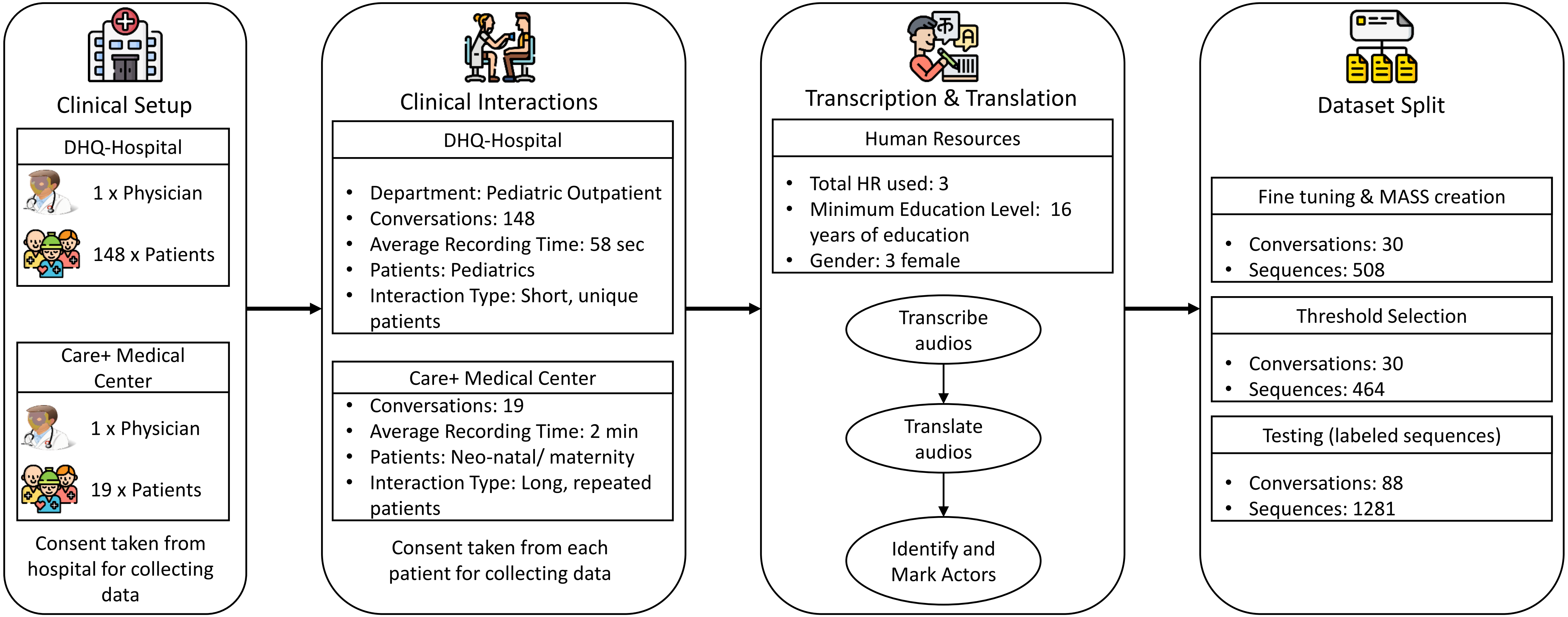
Limitation

- Resource requirements for the Big Data Store will be high



Experimental Setup

Solution 1: Sequence Contraction - Dataset



Clinical Setup

DHQ-Hospital
- 1 x Physician
- 148 x Patients

Care+ Medical Center
- 1 x Physician
- 19 x Patients

Consent taken from hospital for collecting data

Clinical Interactions

DHQ-Hospital
- Department: Pediatric Outpatient
- Conversations: 148
- Average Recording Time: 58 sec
- Patients: Pediatrics
- Interaction Type: Short, unique patients

Care+ Medical Center
- Conversations: 19
- Average Recording Time: 2 min
- Patients: Neo-natal/ maternity
- Interaction Type: Long, repeated patients

Consent taken from each patient for collecting data

Transcription & Translation

Human Resources
- Total HR used: 3
- Minimum Education Level: 16 years of education
- Gender: 3 female

Transcribe audios → Translate audios → Identify and Mark Actors

Dataset Split

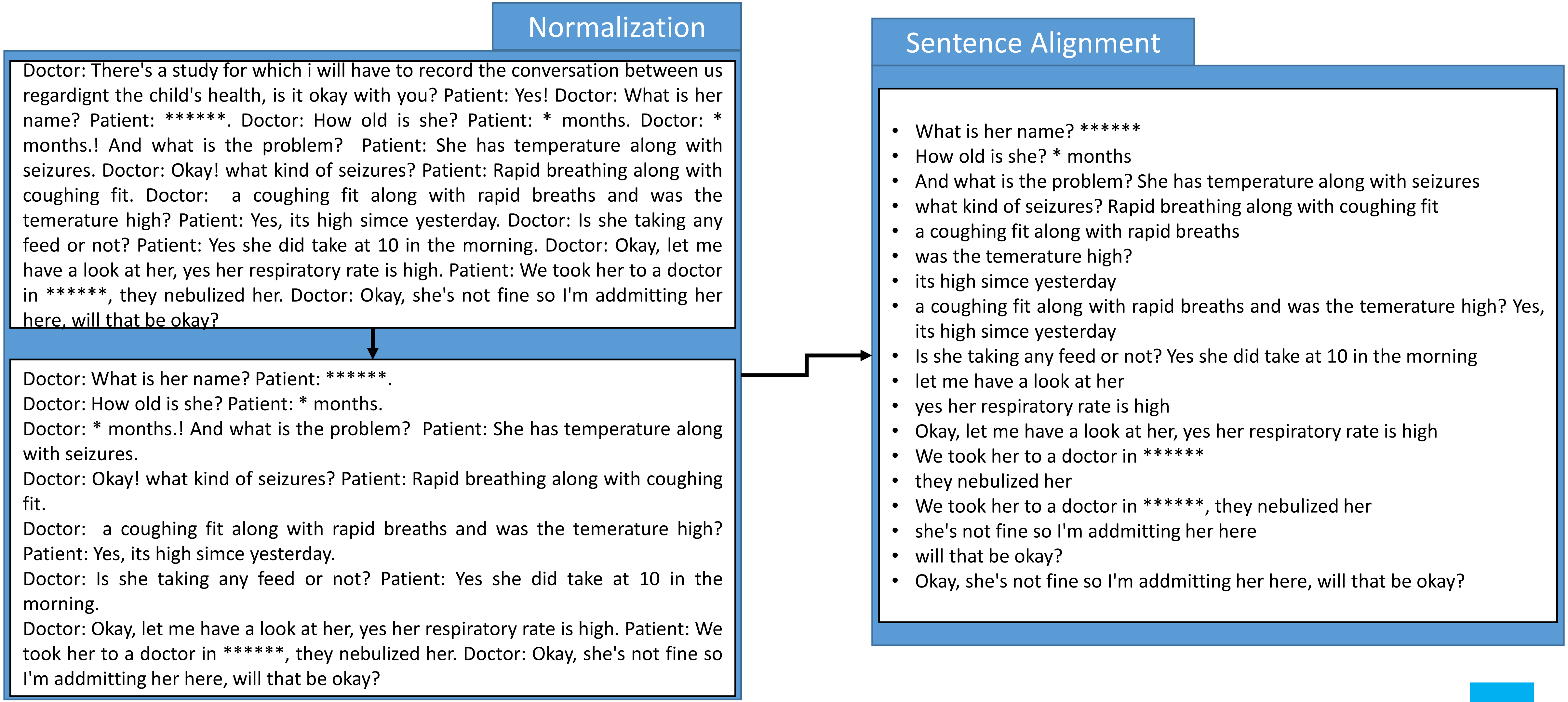
Fine tuning & MASS creation
- Conversations: 30
- Sequences: 508

Threshold Selection
- Conversations: 30
- Sequences: 464

Testing (labeled sequences)
- Conversations: 88
- Sequences: 1281

Solution 1: Sequence Classification

Conversation Example



Normalization

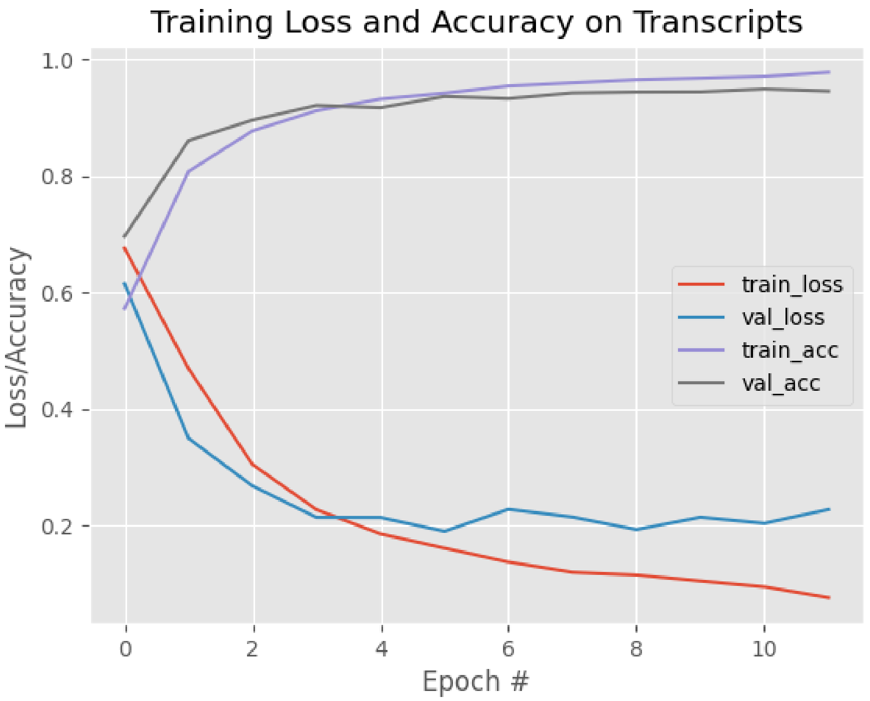
Doctor: There's a study for which i will have to record the conversation between us regardignt the child's health, is it okay with you? Patient: Yes! Doctor: What is her name? Patient: ******. Doctor: How old is she? Patient: * months. Doctor: * months.! And what is the problem? Patient: She has temperature along with seizures. Doctor: Okay! what kind of seizures? Patient: Rapid breathing along with coughing fit. Doctor: a coughing fit along with rapid breaths and was the temerature high? Patient: Yes, its high simce yesterday. Doctor: Is she taking any feed or not? Patient: Yes she did take at 10 in the morning. Doctor: Okay, let me have a look at her, yes her respiratory rate is high. Patient: We took her to a doctor in ******, they nebulized her. Doctor: Okay, she's not fine so I'm addmitting her here, will that be okay?

Doctor: What is her name? Patient: ******.
Doctor: How old is she? Patient: * months.
Doctor: * months.! And what is the problem? Patient: She has temperature along with seizures.
Doctor: Okay! what kind of seizures? Patient: Rapid breathing along with coughing fit.
Doctor: a coughing fit along with rapid breaths and was the temerature high? Patient: Yes, its high simce yesterday.
Doctor: Is she taking any feed or not? Patient: Yes she did take at 10 in the morning.
Doctor: Okay, let me have a look at her, yes her respiratory rate is high. Patient: We took her to a doctor in ******, they nebulized her. Doctor: Okay, she's not fine so I'm addmitting her here, will that be okay?

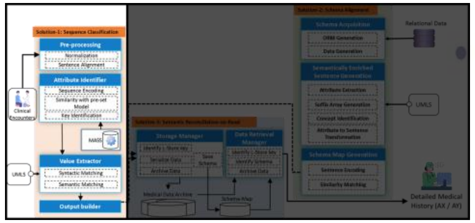
Sentence Alignment

- What is her name? ******
- How old is she? * months
- And what is the problem? She has temperature along with seizures
- what kind of seizures? Rapid breathing along with coughing fit
- a coughing fit along with rapid breaths
- was the temerature high?
- its high simce yesterday
- a coughing fit along with rapid breaths and was the temerature high? Yes, its high simce yesterday
- Is she taking any feed or not? Yes she did take at 10 in the morning
- let me have a look at her
- yes her respiratory rate is high
- Okay, let me have a look at her, yes her respiratory rate is high
- We took her to a doctor in ******
- they nebulized her
- We took her to a doctor in ******, they nebulized her
- she's not fine so I'm addmitting her here
- will that be okay?
- Okay, she's not fine so I'm addmitting her here, will that be okay?

Solution 1: Fine Tuning DistilBERT

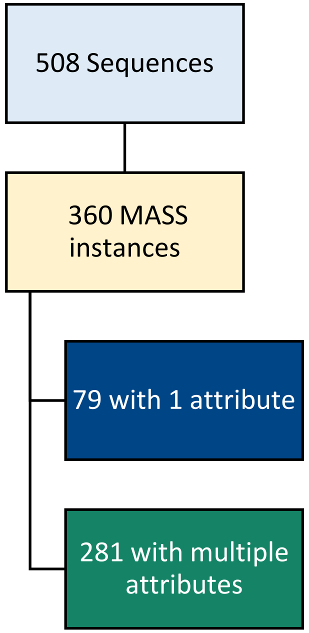


Hyper parameter	Value
Batch Size	32
Loss Function	Cross Entropy
Evaluation Metric	Sparse Categorical Accuracy
Optimizer	AdamW
Initial Learning Rate	1e-4
Warmup steps	10%
Test Accuracy	95%



Experimental Setup

Solution 1: Sequence Contraction - MASS development



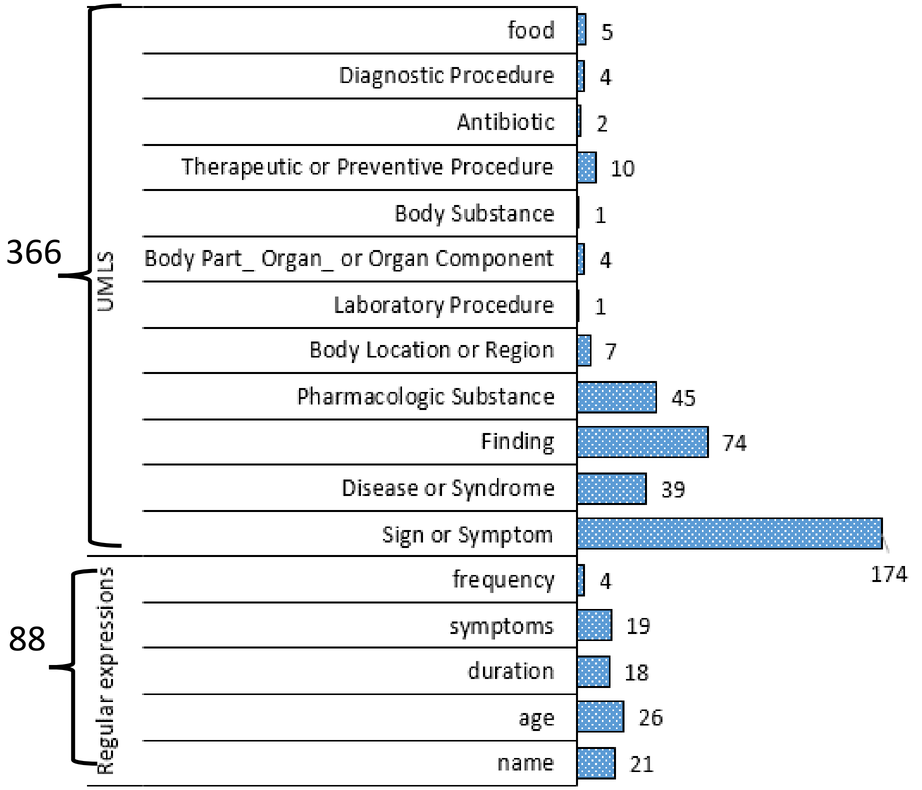
508 Sequences

360 MASS instances

79 with 1 attribute

281 with multiple attributes

Data Labeling



Label Stats

what is child's name? hammad

[CLS] what is child's name? [MASK];;name;;(.)?what(.*)?name(.*?)\? ((his|her|patient)? name is)?(?P<Name>.*)

how old is he? 5 years

[CLS] how old is he? [MASK] years;;age;;(old|age)(.)?\? (he is|she is|shes)?(?P<Age>.*)(years|month)?(.*)?

the child has cough

*[CLS] the child has [MASK];;Sign or Symptom;;umls

what's wrong with the baby? the child has cough and cold

*[CLS]what's wrong with the baby? [SEP] the child has [MASK] and [MASK];;Sign or Symptom,Sign or Symptom;;umls

how long? it's been 3 days

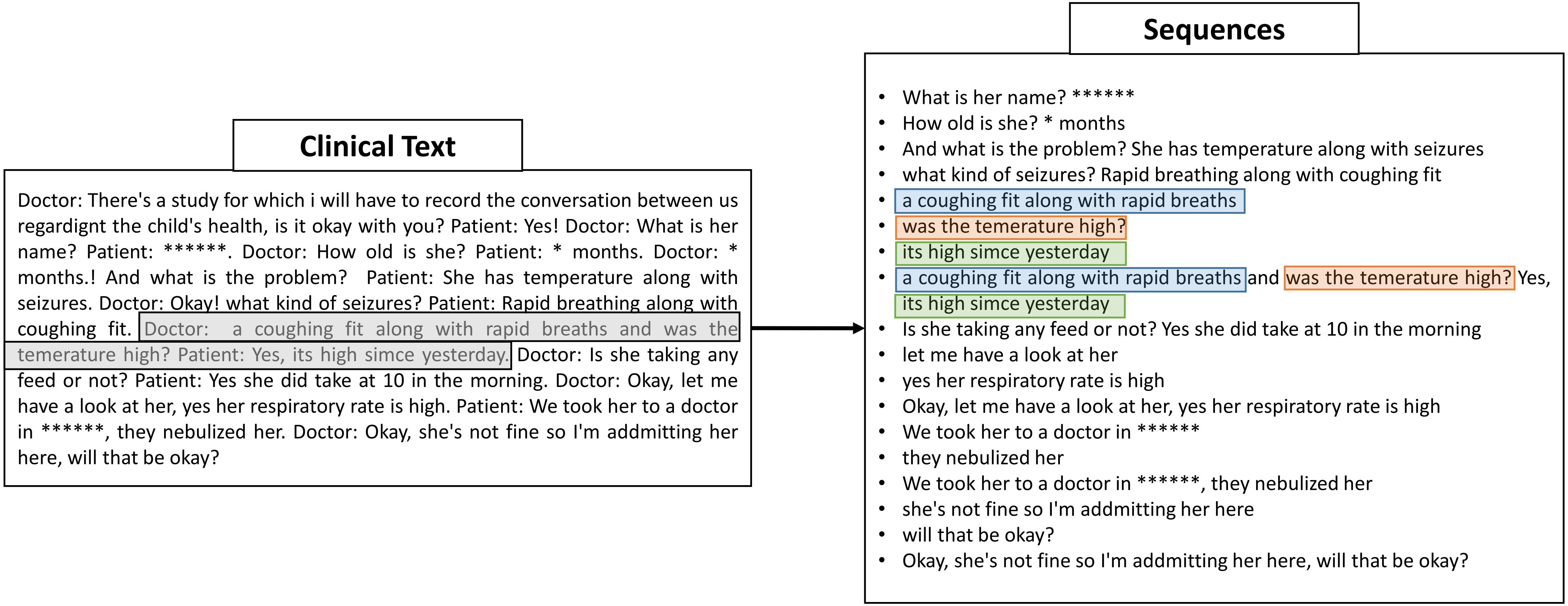
[CLS] how long? [SEP] it's been [MASK] [days];;duration;;(.)? it's been (\s+the last|\s+previous)?\s+(?P<Duration>.*(day|week|month|year)?(s)?)

A sample of MASS instances

- Due to the presence of multiple attributes in MASS instances, the number of labels are more than the number of sequences
- UMLS labels correspond to the UMLS semantic concept types

Solution 1: Sequence Classification

Pre-processing Example



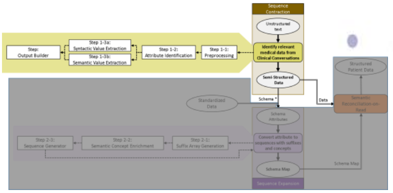
Clinical Text

Doctor: There's a study for which i will have to record the conversation between us regardignt the child's health, is it okay with you? Patient: Yes! Doctor: What is her name? Patient: ******. Doctor: How old is she? Patient: * months. Doctor: * months.! And what is the problem? Patient: She has temperature along with seizures. Doctor: Okay! what kind of seizures? Patient: Rapid breathing along with coughing fit. Doctor: a coughing fit along with rapid breaths and was the temerature high? Patient: Yes, its high simce yesterday. Doctor: Is she taking any feed or not? Patient: Yes she did take at 10 in the morning. Doctor: Okay, let me have a look at her, yes her respiratory rate is high. Patient: We took her to a doctor in ******, they nebulized her. Doctor: Okay, she's not fine so I'm addmitting her here, will that be okay?

Sequences

- What is her name? ******
- How old is she? * months
- And what is the problem? She has temperature along with seizures
- what kind of seizures? Rapid breathing along with coughing fit
- a coughing fit along with rapid breaths
- was the temerature high?
- its high simce yesterday
- a coughing fit along with rapid breaths and was the temerature high? Yes, its high simce yesterday
- Is she taking any feed or not? Yes she did take at 10 in the morning
- let me have a look at her
- yes her respiratory rate is high
- Okay, let me have a look at her, yes her respiratory rate is high
- We took her to a doctor in ******
- they nebulized her
- We took her to a doctor in ******, they nebulized her
- she's not fine so I'm addmitting her here
- will that be okay?
- Okay, she's not fine so I'm addmitting her here, will that be okay?

Experimental Setup

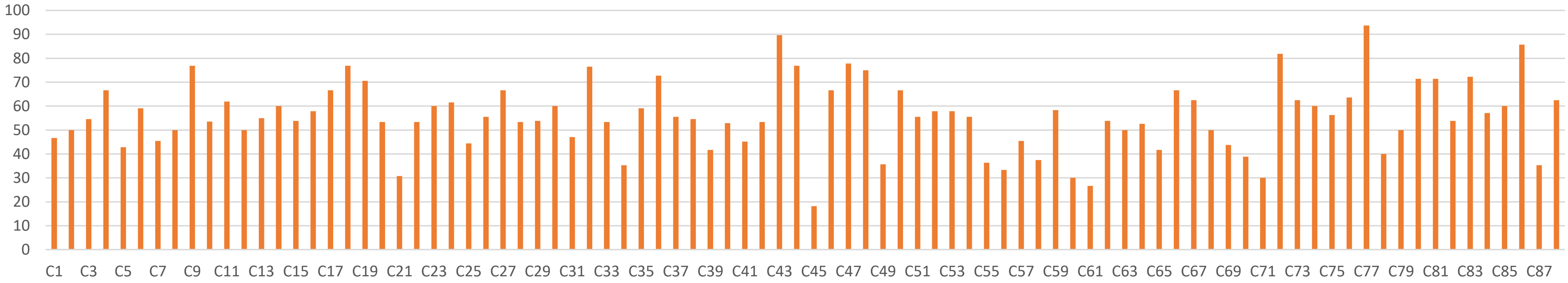


Solution 1: Sequence Contraction – Individual conversation Performance

(a) **Baseline Methodology**: all-mpnet-base-v2

Max: 93.75

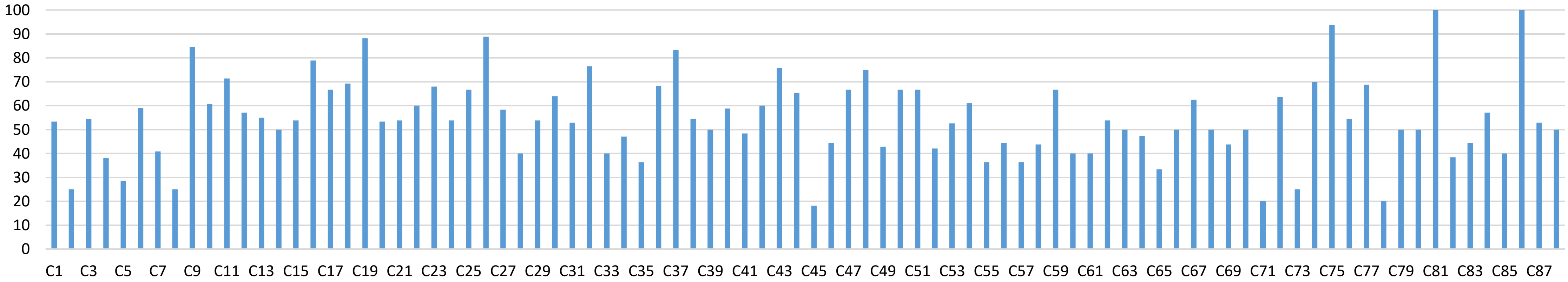
Min: 18.18182



(b) **Proposed Methodology**: Fine-Tuned DistilBERT

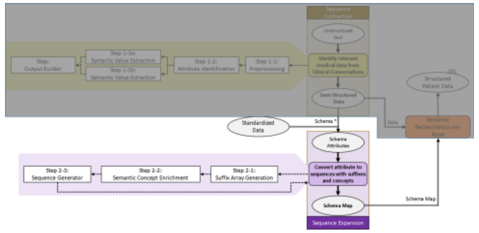
Max: 100

Min: 18.18182

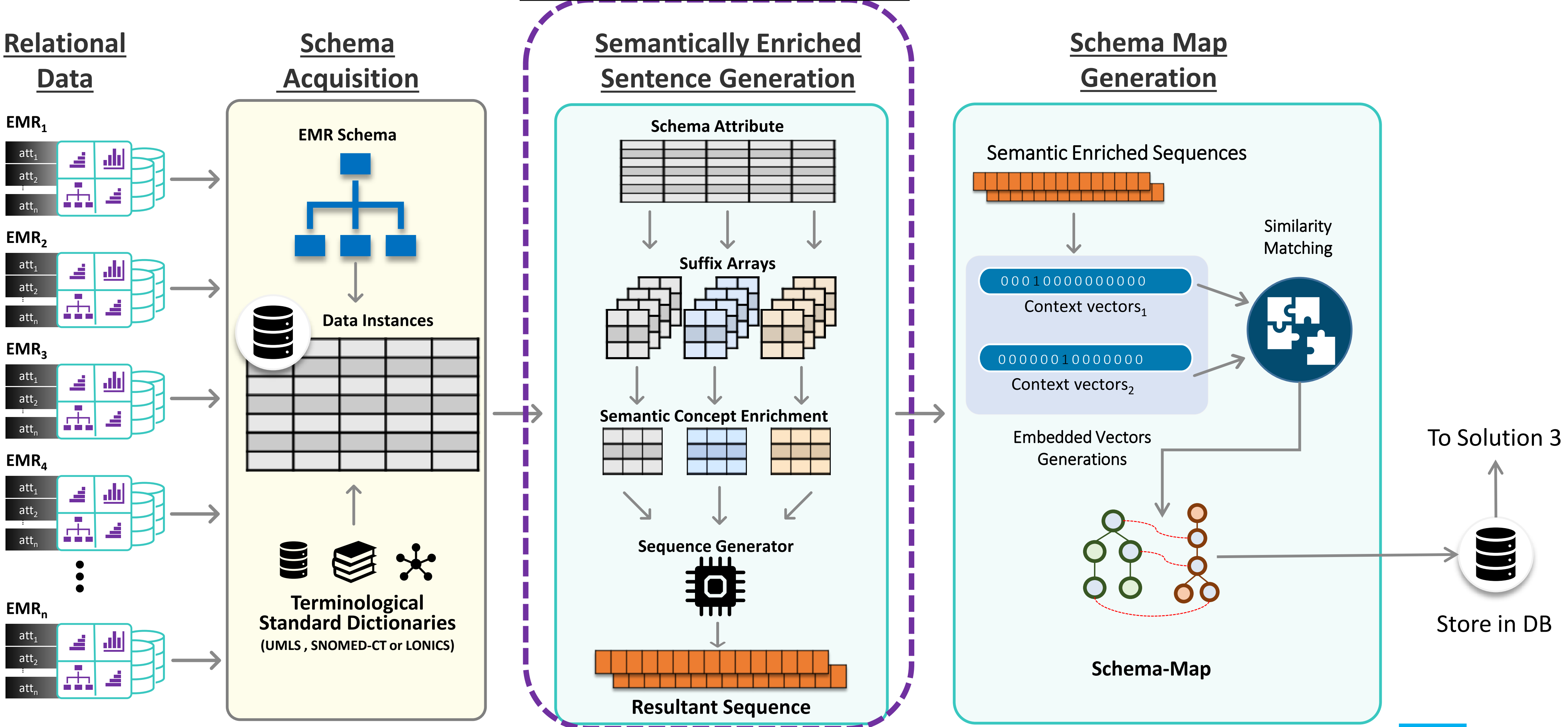


Solution 2: Sequence Expansion

Abstract View

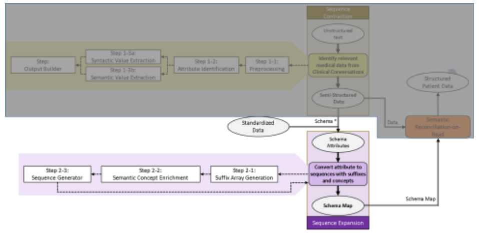


Sequence Expansion



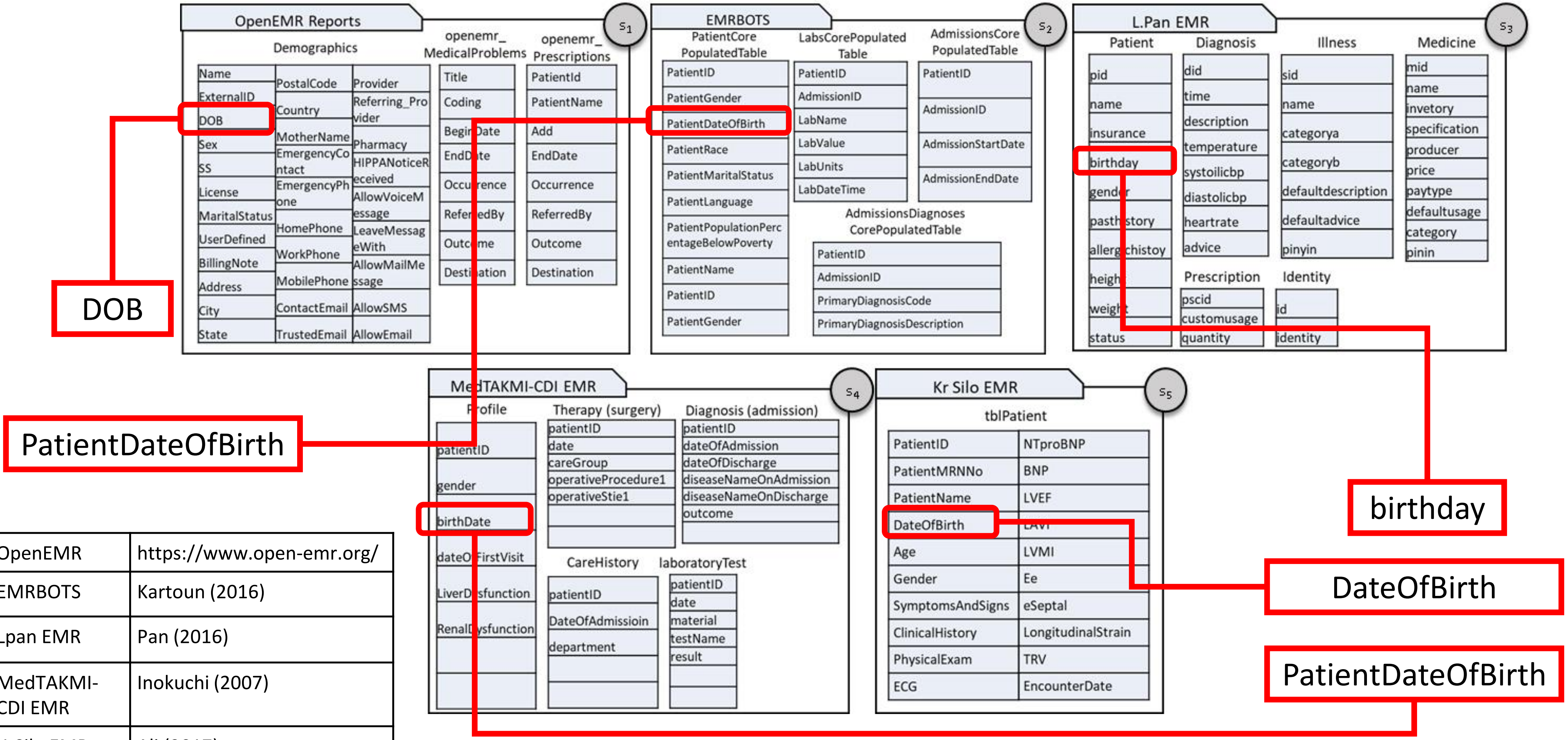
Satti, Fahad Ahmed, et al. "Unsupervised Semantic Mapping for Healthcare Data Storage Schema." IEEE Access 9 (2021): 107267-107278.

Solution 2: Sequence Expansion



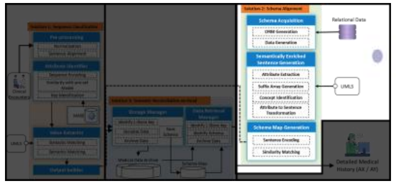
Step 2-3: Sequence Generator
Step 2-2: Semantic Concept Enrichment
Step 2-1: Suffix Array Generation
Schema Map
Sequence Expansion

Challenge: Heterogeneous Schema



S1	OpenEMR	https://www.open-emr.org/
S2	EMRBOTS	Kartoun (2016)
S3	Lpan EMR	Pan (2016)
S4	MedTAKMI-CDI EMR	Inokuchi (2007)
S5	KrSilo EMR	Ali (2017)

Experimental Setup



Solution 2: Sequence Expansion – Threshold Selection

Evaluation Metric : MCC

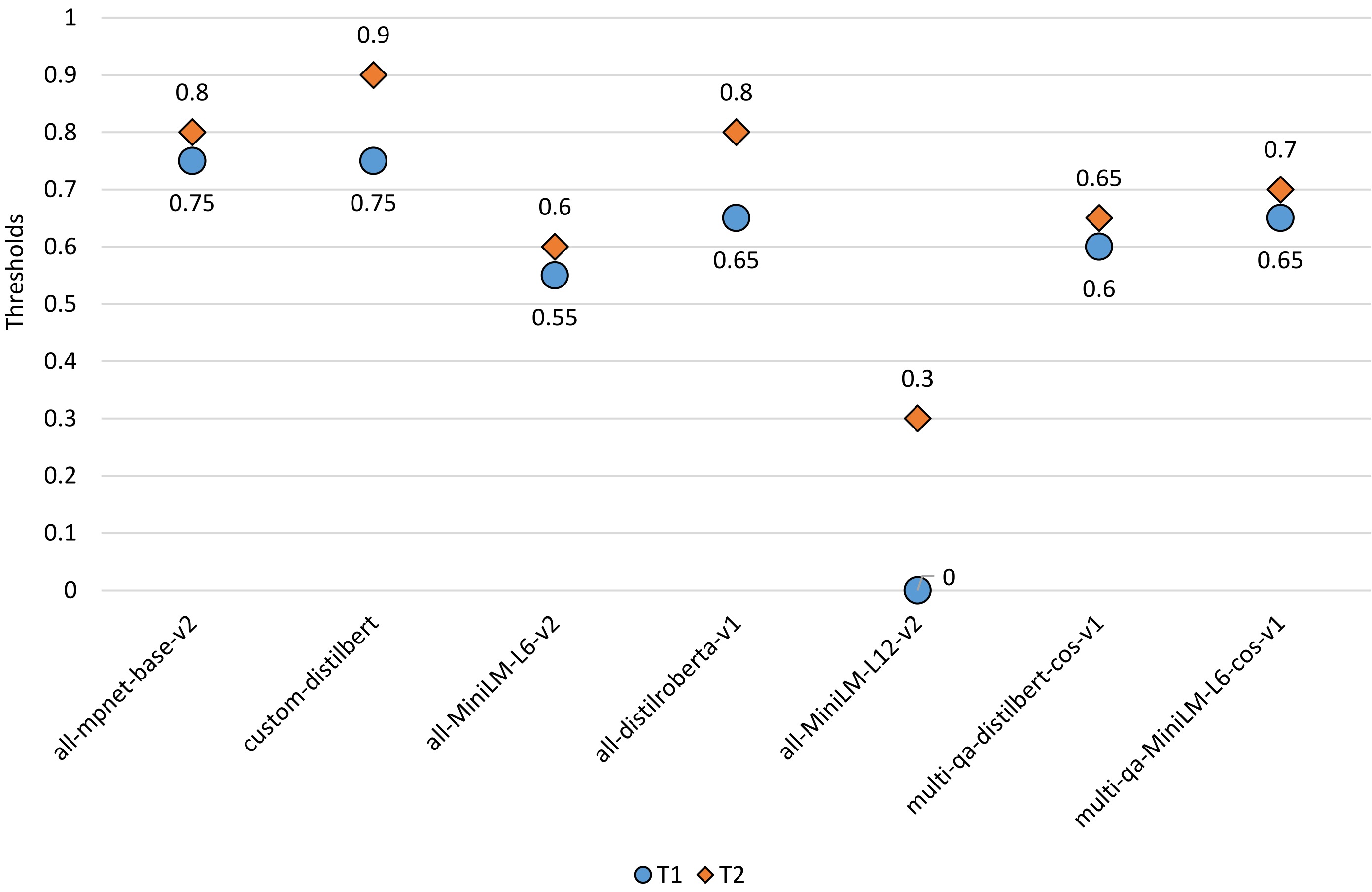
$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \rightarrow [-1,1]$$

TP = True Positive, TN = True Negative,
FP = False Positive, FN = False Negative

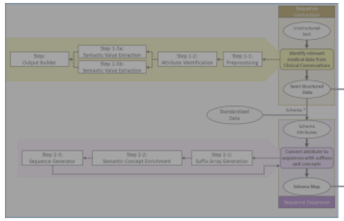
- Accuracy fails to account for imbalanced datasets
- F1 measure is not affected by the true negative scores.
- MCC provides an acceptable alternate in our current scenario comprising of imbalanced dataset (largely in favour of class "unrelated")

Chicco (2020)

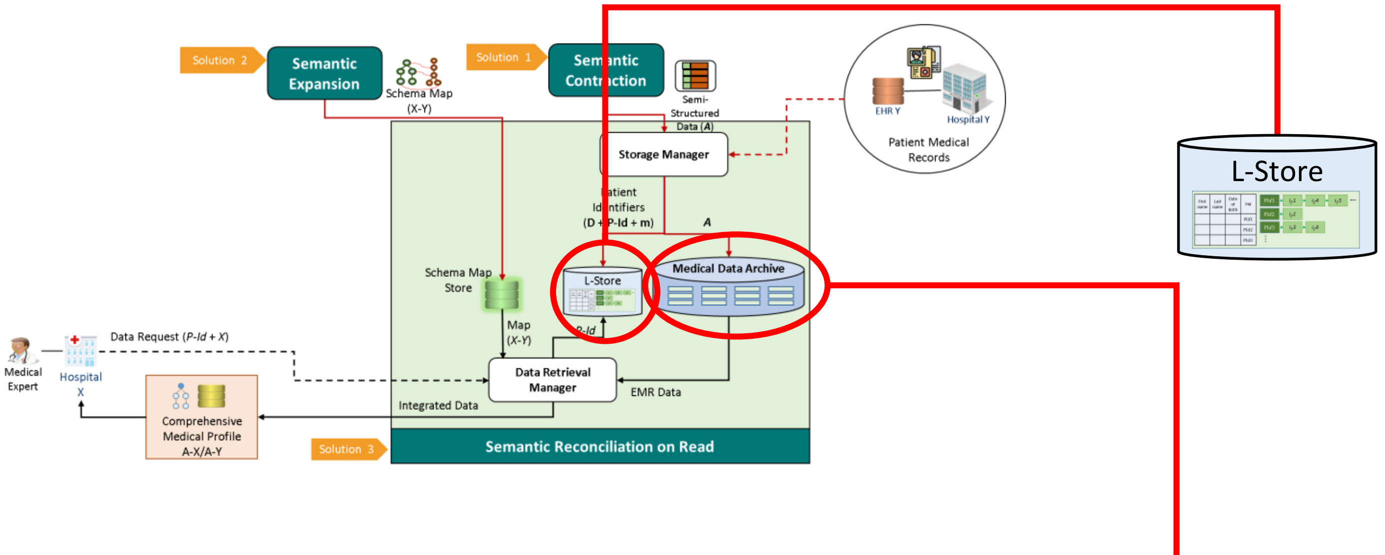
	Threshold	
Method	Unequal and similar	Similar and equal
all-mpnet-base-v2	0.75	0.8
custom-distilbert	0.75	0.9
all-MiniLM-L6-v2	0.55	0.6
all-distilroberta-v1	0.65	0.8
all-MiniLM-L12-v2	0	0.3
multi-qa-distilbert-cos-v1	0.6	0.65
multi-qa-MiniLM-L6-cos-v1	0.65	0.7

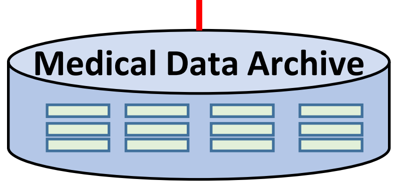


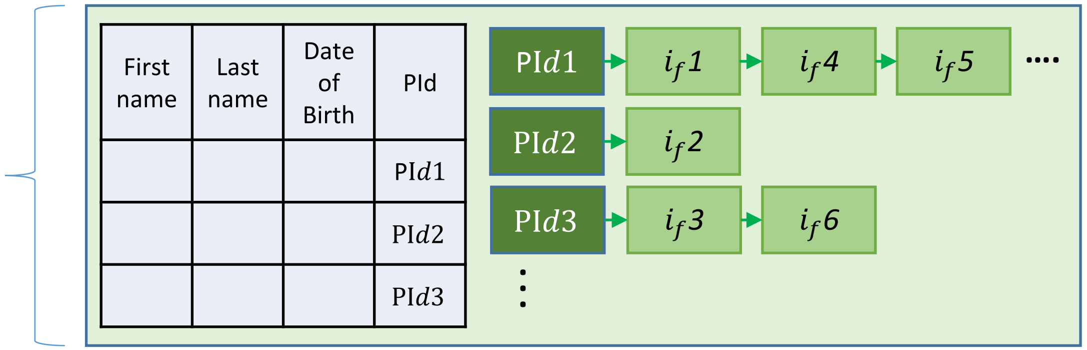
Solution 3: Semantic Reconciliation-on-Read



Data Storage View

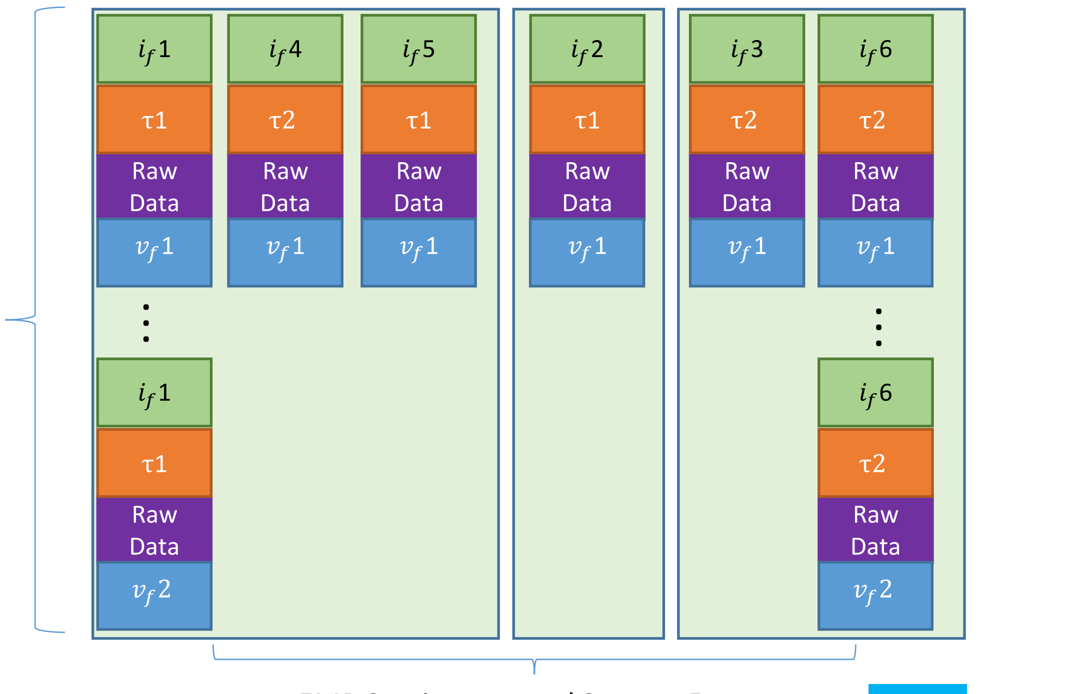






First name	Last name	Date of Birth	PId
			$\mathrm{PI}d1$
			$\mathrm{PI}d2$
			$\mathrm{PI}d3$
⋮			

$\mathrm{PI}d1$ → i_f1 → i_f4 → i_f5 ....
$\mathrm{PI}d2$ → i_f2
$\mathrm{PI}d3$ → i_f3 → i_f6
⋮

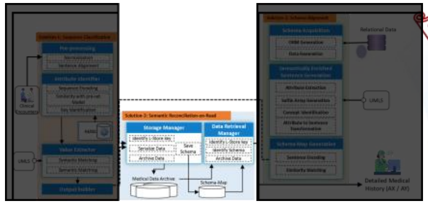


i_f1 $\tau1$ Raw Data v_f1
⋮
i_f1 $\tau1$ Raw Data v_f2
i_f4 $\tau2$ Raw Data v_f1
i_f5 $\tau1$ Raw Data v_f1
i_f2 $\tau1$ Raw Data v_f1
i_f3 $\tau2$ Raw Data v_f1
i_f6 $\tau2$ Raw Data v_f1
⋮
i_f6 $\tau2$ Raw Data v_f2

EMR Semi-structured Storage Form

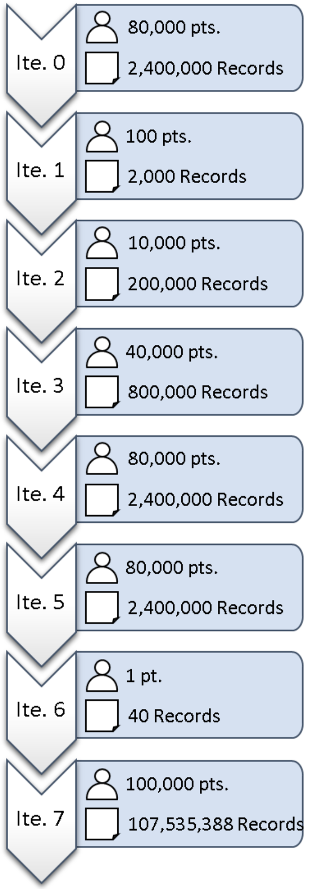
- EMR data is **serialized** into Semi-Structured form with “Raw Data” containing “**key:value**” pairs.
- **Disambiguation attributes** (such as firstname, lastname, dateofbirth) can be used to **identify patients** across organizational boundaries.

Experimental Setup



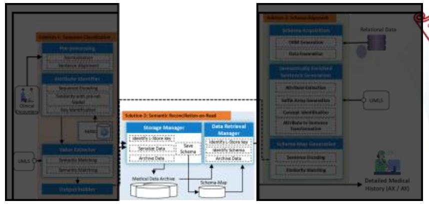
Solution 3: Evaluation Critera

Timeliness Evaluation for all 7 iterations shows the performance of the proposed approach in the presence of Big Data.


Ite. 0
80,000 pts.
2,400,000 Records
Ite. 1
100 pts.
2,000 Records
Ite. 2
10,000 pts.
200,000 Records
Ite. 3
40,000 pts.
800,000 Records
Ite. 4
80,000 pts.
2,400,000 Records
Ite. 5
80,000 pts.
2,400,000 Records
Ite. 6
1 pt.
40 Records
Ite. 7
100,000 pts.
107,535,388 Records

Iteration	Total Fragments	File size for C1 (Kb)	File size for C2 (Kb)	File size for C3 (Kb)
Initial	2,400,000	-	-	-
1	2000	659	6	181
2	200,000	66,260	580	18,059
3	800,000	264,923	2320	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	107,535,388	25,752,400	7,263	11,118,380
Total	**115,737,428**	**27,594,967**	**19,448**	**11,642,091**

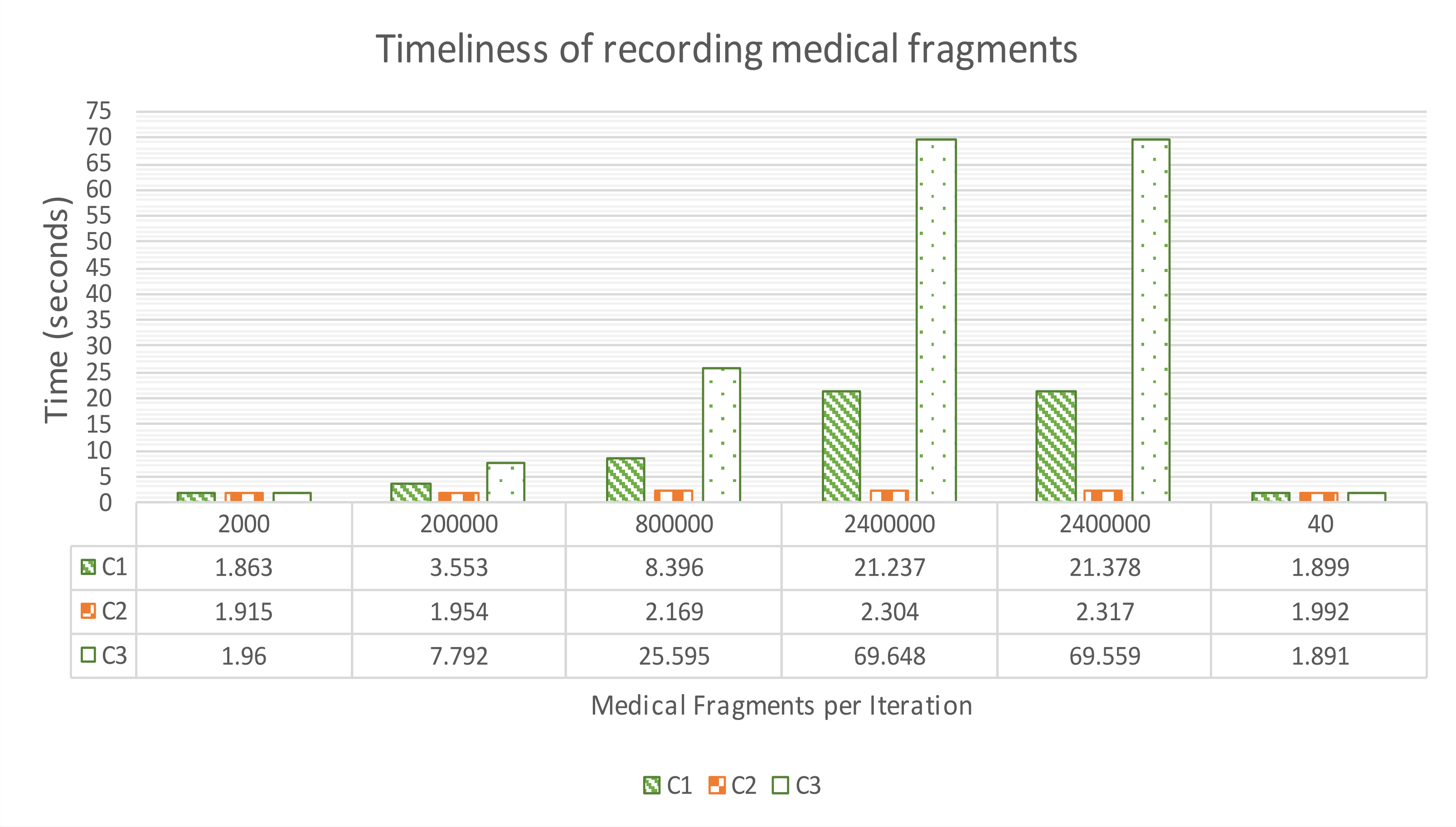
Criteria	Description	Metric
C1	Time taken to insert medical fragment file into HDFS	Time
C2	Time taken to insert medical fragment bridging information, linking Patient Id with fragment id into HDFS	Time
C3	Time taken to insert patient index part of L-Store into HDFS	Time
C4	Time taken to create table schema in Hive	Time
C5	Time taken to create medical fragment bridging table schema in Hive	Time
C6	Time taken to create 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



Experimental Setup

Solution 3: Semantic Reconciliation-on-Read

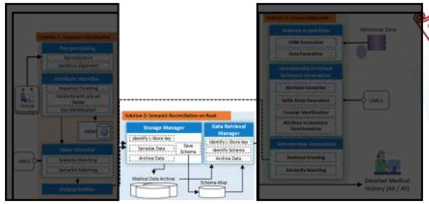
Criteria	Description
C1	Time taken to insert patient index part of L-Store into HDFS
C2	Time taken to insert medical fragment bridging information, linking Patient Id with fragment id into HDFS
C3	Time taken to insert medical fragment file into HDFS



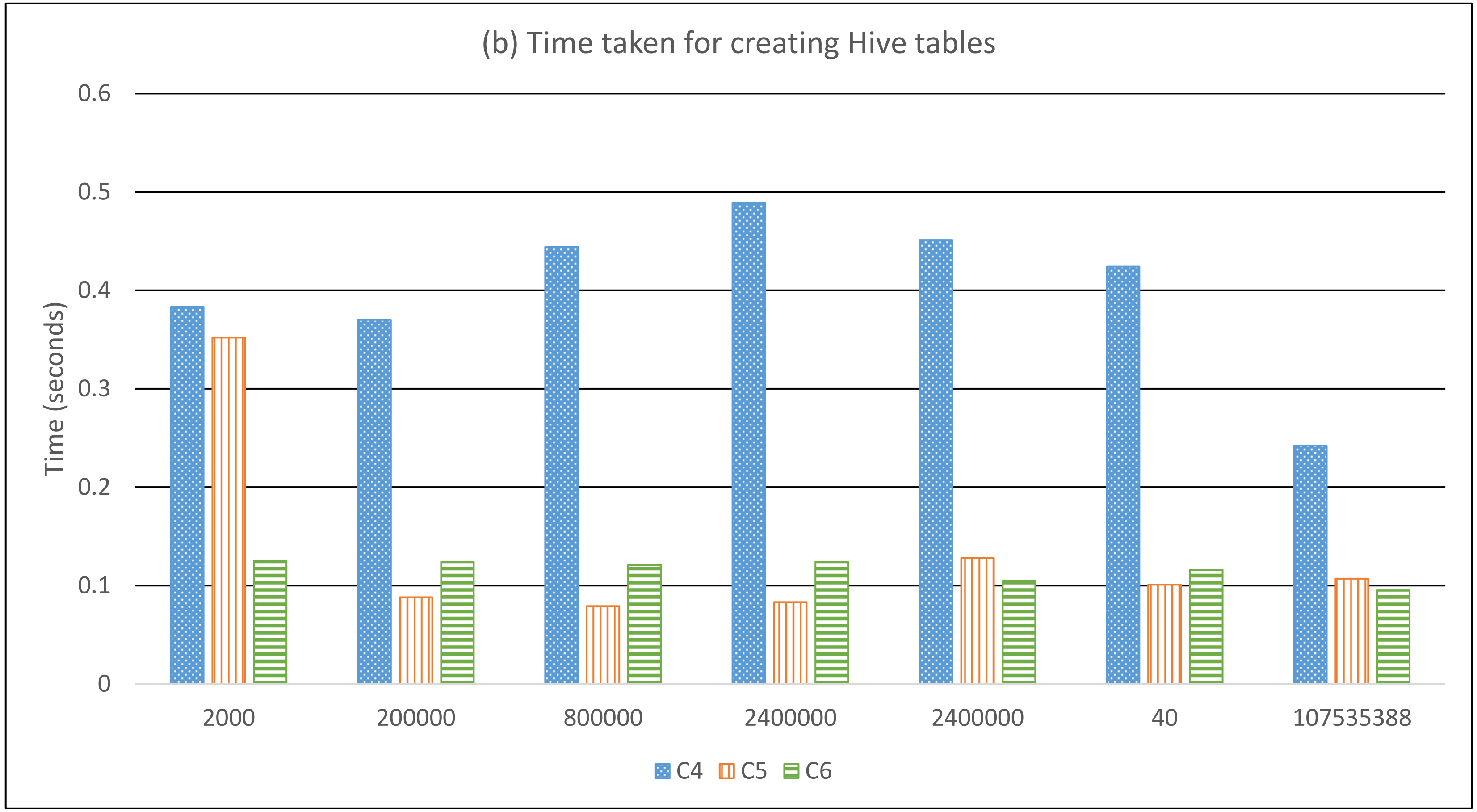
	2000	200000	800000	2400000	2400000	40
▨ C1	1.863	3.553	8.396	21.237	21.378	1.899
▨ C2	1.915	1.954	2.169	2.304	2.317	1.992
□ C3	1.96	7.792	25.595	69.648	69.559	1.891

Experimental Setup

Solution 3: Semantic Reconciliation-on-Read

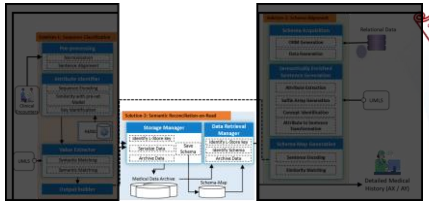


Criteria	Description
C4	Time taken to create table schema in Hive
C5	Time taken to create medical fragment bridging table schema in Hive
C6	Time taken to create patient index table schema in Hive



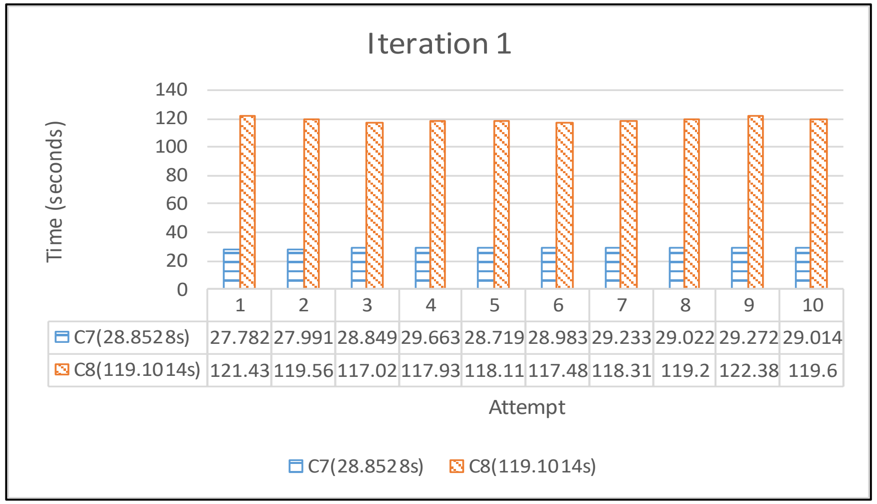
Experimental Setup

Solution 3: Semantic Reconciliation-on-Read



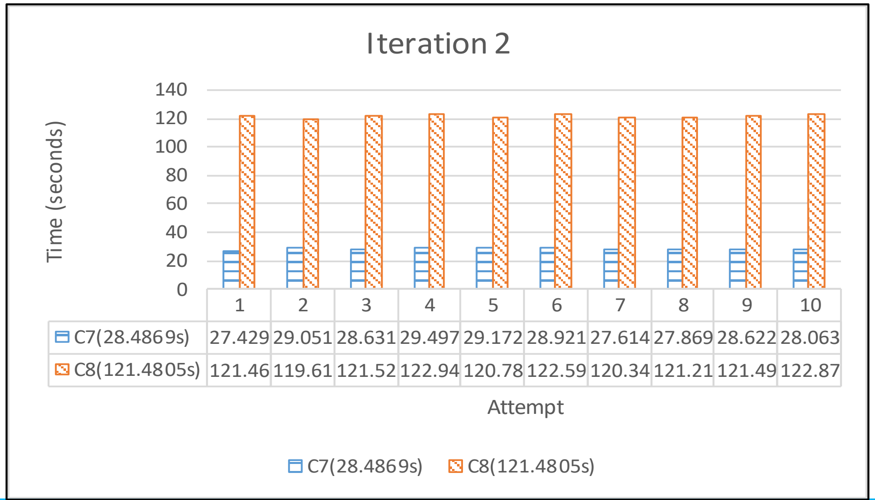
Criteria	Description	Metric
C7	Time taken to retrieve all fragment ids for 1 user	Time
C8	Time taken to retrieve all medical fragments for 1 user	Time

Iteration 1



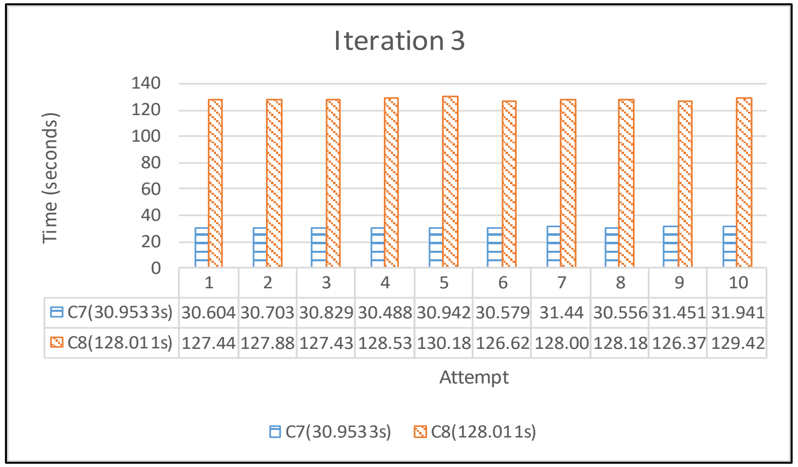
	1	2	3	4	5	6	7	8	9	10
C7(28.8528s)	27.782	27.991	28.849	29.663	28.719	28.983	29.233	29.022	29.272	29.014
C8(119.1014s)	121.43	119.56	117.02	117.93	118.11	117.48	118.31	119.2	122.38	119.6

Iteration 2



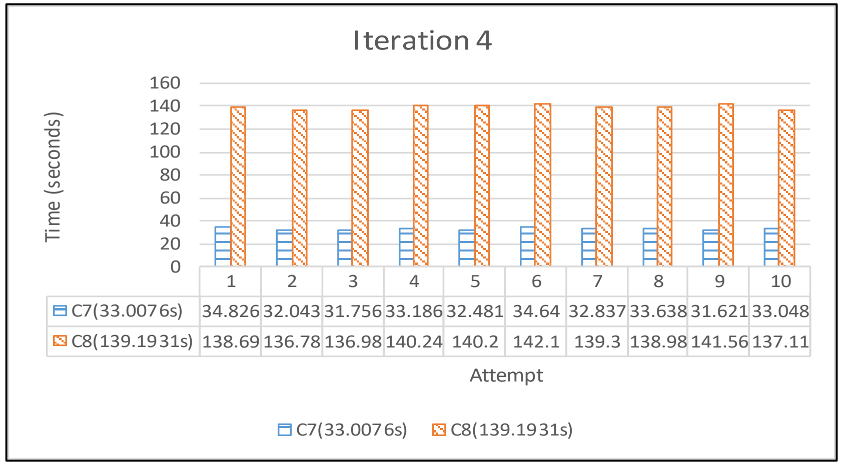
	1	2	3	4	5	6	7	8	9	10
C7(28.4869s)	27.429	29.051	28.631	29.497	29.172	28.921	27.614	27.869	28.622	28.063
C8(121.4805s)	121.46	119.61	121.52	122.94	120.78	122.59	120.34	121.21	121.49	122.87

Iteration 3



	1	2	3	4	5	6	7	8	9	10
C7(30.9533s)	30.604	30.703	30.829	30.488	30.942	30.579	31.44	30.556	31.451	31.941
C8(128.011s)	127.44	127.88	127.43	128.53	130.18	126.62	128.00	128.18	126.37	129.42

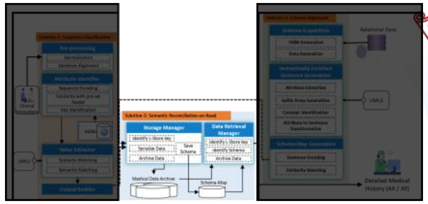
Iteration 4



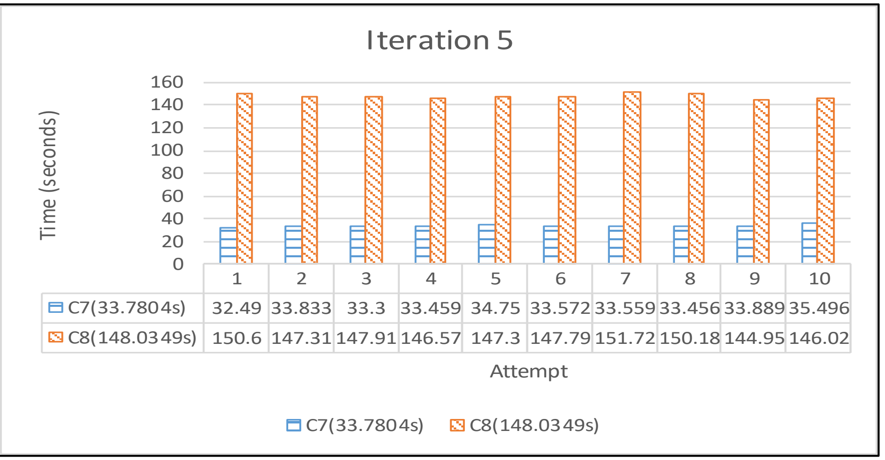
	1	2	3	4	5	6	7	8	9	10
C7(33.0076s)	34.826	32.043	31.756	33.186	32.481	34.64	32.837	33.638	31.621	33.048
C8(139.1931s)	138.69	136.78	136.98	140.24	140.2	142.1	139.3	138.98	141.56	137.11

Experimental Setup

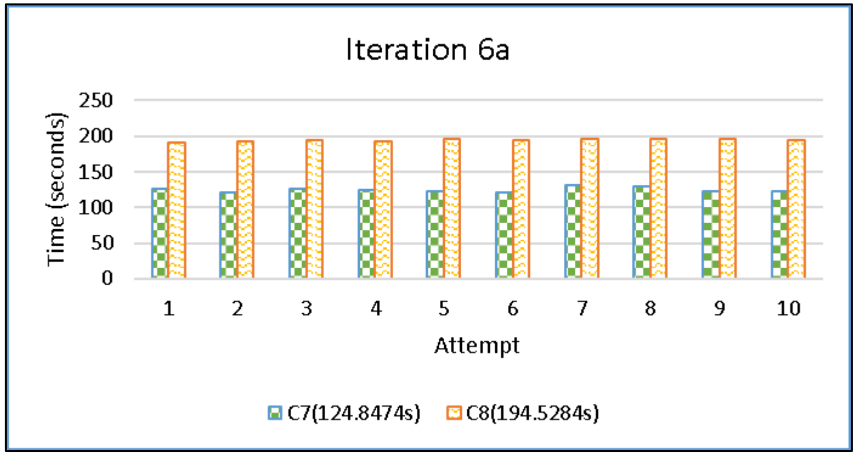
Solution 3: Semantic Reconciliation-on-Read

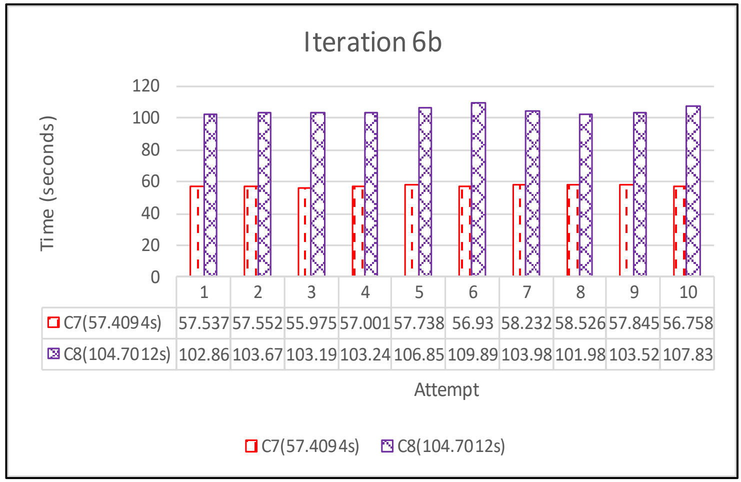


Criteria	Description	Metric
C7	Time taken to retrieve all fragment ids for 1 user	Time
C8	Time taken to retrieve all medical fragments for 1 user	Time

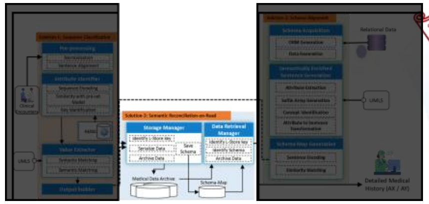

Iteration 5

	1	2	3	4	5	6	7	8	9	10
C7(33.7804s)	32.49	33.833	33.3	33.459	34.75	33.572	33.559	33.456	33.889	35.496
C8(148.0349s)	150.6	147.31	147.91	146.57	147.3	147.79	151.72	150.18	144.95	146.02


Iteration 6a


Iteration 6b

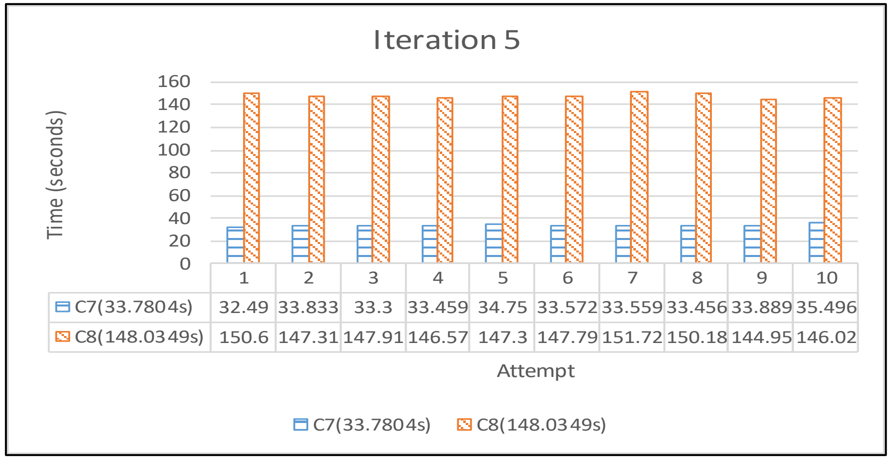
	1	2	3	4	5	6	7	8	9	10
C7(57.4094s)	57.537	57.552	55.975	57.001	57.738	56.93	58.232	58.526	57.845	56.758
C8(104.7012s)	102.86	103.67	103.19	103.24	106.85	109.89	103.98	101.98	103.52	107.83



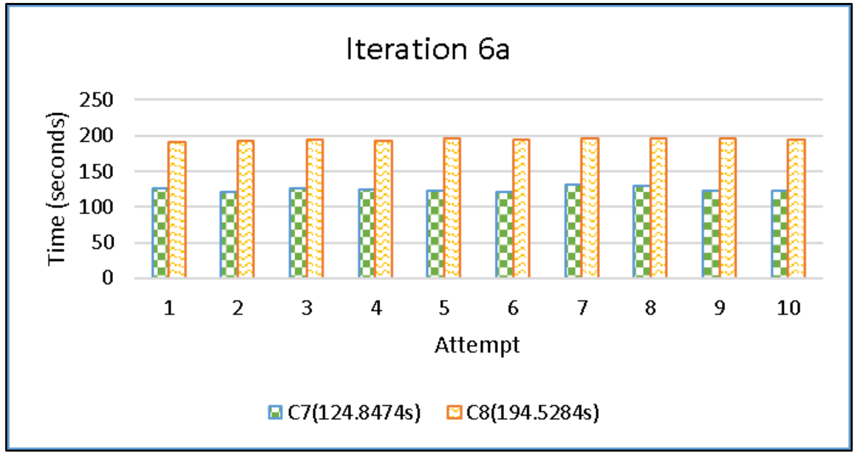
Experimental Setup

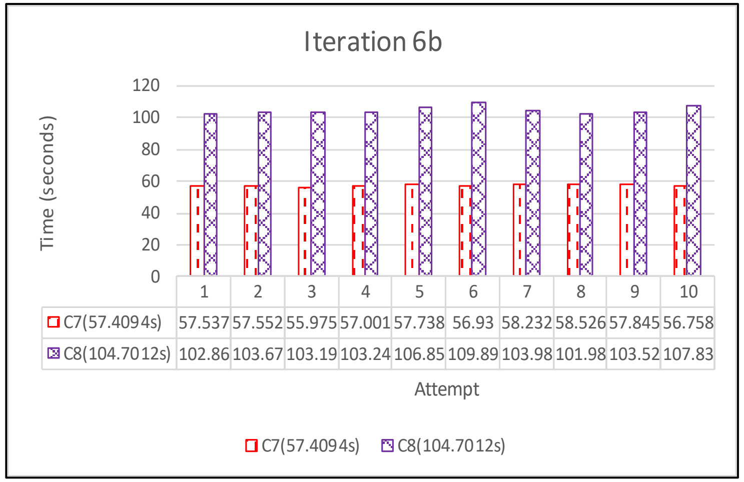
Solution 3: Semantic Reconciliation-on-Read

Criteria	Description	Metric
C7	Time taken to retrieve all fragment ids for 1 user	Time
C8	Time taken to retrieve all medical fragments for 1 user	Time

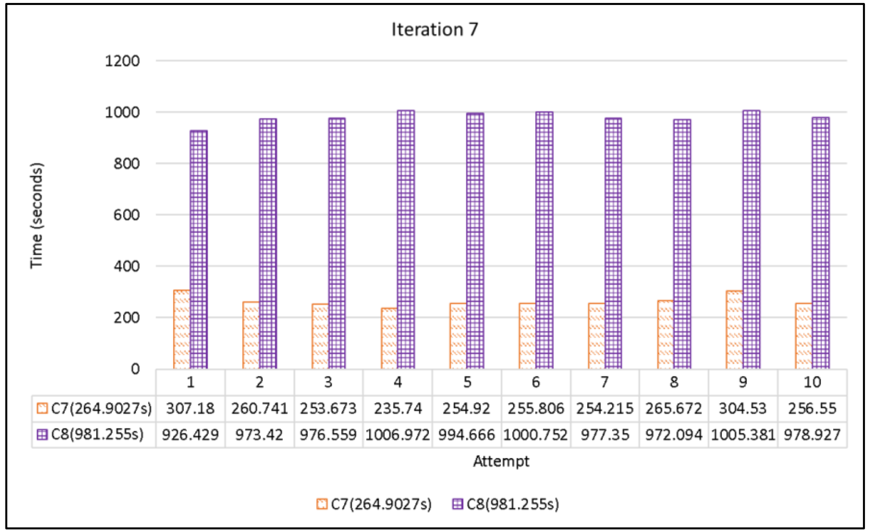


	1	2	3	4	5	6	7	8	9	10
C7(33.7804s)	32.49	33.833	33.3	33.459	34.75	33.572	33.559	33.456	33.889	35.496
C8(148.0349s)	150.6	147.31	147.91	146.57	147.3	147.79	151.72	150.18	144.95	146.02





	1	2	3	4	5	6	7	8	9	10
C7(57.4094s)	57.537	57.552	55.975	57.001	57.738	56.93	58.232	58.526	57.845	56.758
C8(104.7012s)	102.86	103.67	103.19	103.24	106.85	109.89	103.98	101.98	103.52	107.83



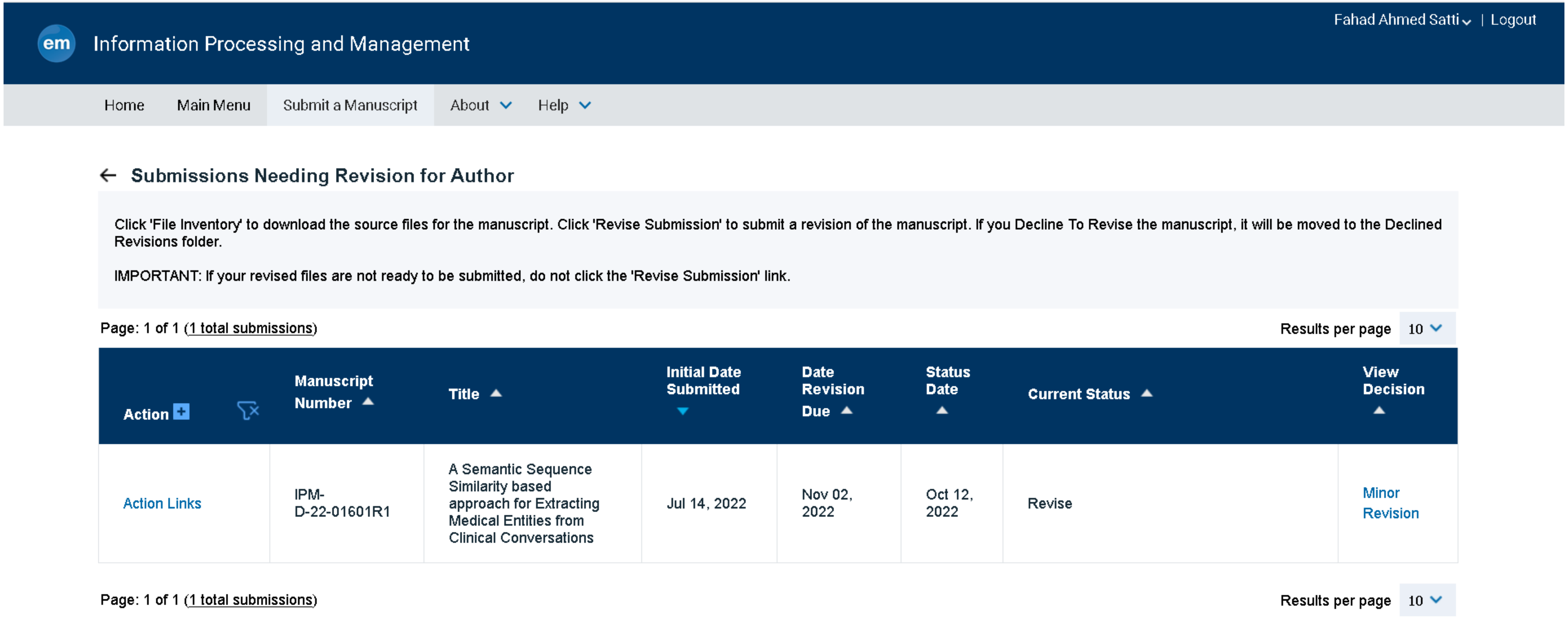
	1	2	3	4	5	6	7	8	9	10
C7(264.9027s)	307.18	260.741	253.673	235.74	254.92	255.806	254.215	265.672	304.53	256.55
C8(981.255s)	926.429	973.42	976.559	1006.972	994.666	1000.752	977.35	972.094	1005.381	978.927

Solution 1: Paper Status

Information Processing & Management

Supports *open access*

11 CiteScore | 7.466 Impact Factor



em
Information Processing and Management
Fahad Ahmed Satti | Logout
Home
Main Menu
Submit a Manuscript
About
Help

Submissions Needing Revision for Author

Click 'File Inventory' to download the source files for the manuscript. Click 'Revise Submission' to submit a revision of the manuscript. If you Decline To Revise the manuscript, it will be moved to the Declined Revisions folder.

IMPORTANT: If your revised files are not ready to be submitted, do not click the 'Revise Submission' link.

Page: 1 of 1 (1 total submissions) Results per page 10

Action	Manuscript Number	Title	Initial Date Submitted	Date Revision Due	Status Date	Current Status	View Decision
Action Links	IPM-D-22-01601R1	A Semantic Sequence Similarity based approach for Extracting Medical Entities from Clinical Conversations	Jul 14, 2022	Nov 02, 2022	Oct 12, 2022	Revise	Minor Revision

Page: 1 of 1 (1 total submissions) Results per page 10