Title	臨床物語からの時間情報の抽出
Author(s)	Moharasan, Gandhimathi
Citation	
Issue Date	2019-03
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/15784
Rights	
Description	Supervisor:Dam Hieu Chi, 知識科学研究科, 博士

Doctoral Dissertation

EXTRACTION OF TEMPORAL INFORMATION FROM CLINICAL NARRATIVES

MOHARASAN Gandhimathi

s1460012

submitted to

Japan Advanced Institute of Science and Technology

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Main Supervisor: Professor Tu-Bao Ho

Co-Supervisor: Associate Professor Dam Hieu Chi

School of Knowledge Science

Japan Advanced Institute of Science and Technology

March, 2019

Abstract

Keywords: Temporal information extraction, electronic medical records, clinical narratives, conditional Random fields, semi-supervised learning, Naive Bayes classifier.

The Electronic Medical Records (EMRs), is a digital version of patient medical history written and stored by medical professionals at the hospital. It has many components such clinical narratives, radiology reports, laboratory results and etc. Among these components, text data describing medical observation of patients by doctor's and nurse's known as clinical narratives and it covers the most significant information about patient health status. EMR clinical narratives consists with large unstructured data become promising resources to support advancement and enhancement of clinical studies as it obtained by doctor's and nurse's with their medical expertise and observing on each individual patient for one medical practice during treatment activities. One perspective of the exploiting the potential clinical narratives in EMRs is clinical decision support systems, can vary in the areas include prognosis, disease monitoring, adverse drug effects and drug development, etc.

In spite of having many advantages, EMR clinical narratives remain with many challenges for exploitation. One major problem is structured representation of longitudinal clinical narrative of patients for further utilization in medical care. As all the significant medical information of patient's have noted with the notion of time,

temporal reasoning plays a key role in structured representation. Temporal reasoning is a fundamental, yet vital skill that requires the understanding of natural language text. Therefore temporal reasoning is key task for temporal information extraction. Temporal information in clinical text perform the crucial role in interpretation of the patient clinical information such as progress of disease, frequencies of medication information, to detect treatment pattern and adverse drug events.

The dissertation studies for basic steps on reconstructing clinical narratives to structured representation. In other words, the thesis studies to propose methods that most appropriately extract temporal information from clinical narratives by utilizing annotated and unannotated data. To this end, the thesis systematically approaches the three fundamental problems: *extraction of implicit and explicit temporal expressions*, *extraction of temporal events by using annotated data along with the support of unannotated data* and *detection of temporal relation and classification.* Our main targets are to seek provably stable and reliable models that can effectively extracts information from clinical narratives.

The first contribution comes from temporal expression extraction. A novel feature set has been proposed to address the problem of temporal expressions extraction. Our proposed framework has following key theoretical properties: (1) new proposed feature set is obtained from raw clinical text and (2) adopted HeidelTime features that are appropriate for temporal expressions extraction from clinical narratives. Existing methods are either having the advantage of HeidelTime or developing rules/ machine learning models, but not the integrated components of both. Those properties helps to extract temporal expressions effectively.

The second contribution is stemming from temporal event extraction from clinical narratives. The introduction of a novel semi-supervised framework to exploit abundant unannotated data for extracting temporal events from clinical narratives.

To best of our knowledge and survey from literature, this work is the first to propose semi-supervised method for extracting temporal event from clinical text. This approach innovated the novel idea of gradually extending the training corpus by adding annotated data obtaining from unannotated clinical narratives. When working with very high dimensional medical data, our proposed method effectively extracts temporal events. The main result of this study is a novel semi-supervised method that can reach state-of-art performance with stable improvement than existing methods.

The third contribution is from temporal relation identification and classification. We formulated a new assumption on generating and identifying the potential candidate pairs from list of temporal events or expressions that can appropriately relate events/expressions in clinical narratives based on their attributes. Moreover, to address the problem of temporal relation detection, we exploited Naive Bayesian Classifier to detect the temporal relationship among the identified pair's. The effective candidate pair's generation helps to improve the relation classification performance.

In conclusion, our proposed method with novel feature sets can effectively extract temporal expression in clinical narratives. A proposed novel semi-supervised framework for temporal events extraction successfully utilized unannotated clinical narratives along with annotated data and enhanced event extraction performance. In case of temporal relation detection, novel assumption for candidate generation pairs along with adopted dependency parsing approach can improve the quality of candidate pairs and consequently temporal relation classification with Naive Bayesian classification. Finally, this study accomplishes our objectives prosperously as stated.

ACKNOWLEDGMENT

This dissertation would not have been possible without the help, support, guidance and effort of a lot of people. It gives me immense pleasure to express my sincere thanks to whom I am greatly indebted.

First and foremost, I am very grateful to the MEXT (Ministry of Education, Culture, Sports, Science and Technology) Scholarship and Doctoral Research Fellowship of JAIST for its support during the period of study.

I owe my deepest gratitude to my former supervisor, Professor Tu-Bao Ho for his encouragement and continuous support during my study period. His advices and valuable comments helped me to become more mature in both academic and the daily life.

Special thanks to my current supervisor, Associate Professor Dam Hieu Chi for all his generosity and patience. I will always be indebted for his continuous support and kindness.

Also, I would like to express my gratitude to my committee chair, Professor Mitsuru Ikeda (JAIST), Professor Tsutomu Fujinami (JAIST), Associate Professor Hyunh Van-Nam (JAIST) and Professor Satou Kenji (Kanazawa University) who read and give useful comments to improve this dissertation.

I would like to express my thanks to my friends and especially all members in Ho laboratory and Dam laboratory for their support and encouragement in my study period. I have learned a lot through the discussion with them.

Last but not least, I would like to express my gratitude to my family members, who always stay by my side during my hard time and encourage me during my study period at JAIST.

Contents

List of Figures

List of Tables

List of Abbreviations

CDS- Clinical Decision Support

CNLP- Clinical Natural language Processing

CRF - Conditional Random Fields

DSS- Decision Support System

EMR - Electronic Medical Records

NLP - Natural Language processing

NER- Named Entity Recognition

TIE - Temporal Information Extraction

TRC - Temporal Relation Classification

SVM -Support Vector Machine

Chapter 1

Introduction

This chapter first introduces the temporal information extraction, Electronic Medical Records (EMRs) and clinical narratives. Then the various representation of text data are presented. The next part provides a concise view of our research problem, objectives and major contributions. Last part have shown structure of this thesis.

1.1 Introduction to temporal information extraction

Automatic recognition and extraction of information from natural language text has become an active area of research in computational linguistics. Especially Temporal Information Extraction (TIE) has transformed vital area for researchers from the introduction of TimeML corpus [1]. Understanding the temporal information from text is unavoidable for several natural language processing applications such as Information retrieval, Question answering, Temporal intent classification, Query processing and various text mining tasks. To accomplish the goal of temporal information extraction tasks from text, it is important to develop strong annotation standards and corpora for temporal information [2], where the large corpora is available for newswire text[1]. A numerous number of works has been established in TIE systems based on these available corpora [3], [4], [5].

1.1.1 Temporal information extraction from newswire text

Temporal information extraction task initially started with temporal representation in early 1980's. The interval-based algebra has been proposed for representing these temporal information in natural language [6]. Most importantly these temporal representation demanded creation of annotated corpora to advance the Natural Language Processing (NLP) for general text which includes Named Entity Recognition (NER) and temporal information extraction, etc., These advancement along with the Message Understanding Conferences (MUC-6, MUC-7) initiated and introduced the extraction of absolute and relative time informations. This primitive work paves the significant path for temporal information extraction from general text and added the valuable contributions.

In early 2000's rapid development of temporal reasoning and temporal information extraction (TIE) methods accelerated with the creation of TimeBank corpus for general newswire text [1], [7]. TimeBank corpus annotated a temporal reasoning elements by using TimeML specifications and guidelines[2]. This corpora have three types of temporal information: temporal expressions, events and temporal relations. It is a typical and most popular corpus for temporal reasoning and relation learning in the natural language text.

To advance the temporal information extraction research works in general domain, series of TempEval Challenge proposed and added the significant contribu-

[1] http://timeml.org/

[2] $https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf$

tions. The state-of-art contributions for temporal information extraction in newswire text from various research groups are summarized [3]. Later in 2000's the availability of massive quantity of clinical text from Electronic Med- ical Records (EMRs) has created significant demand for clinical text processing and temporal information extraction in the field of health care and medical research.

1.1.2 Temporal information extraction from clinical text

In recent days, the interest of temporal information extraction from domain-specific clinical text has received great attention due to richness of temporality in clinical text, significance of temporal information exploitation in medical care and availability of clinical text. Medical doctors and nurse notes their diagnosis about the patient symptoms, disease, disease progress, occurrence and treatments details in the form on unstructured clinical text with implicit temporal information. The produced huge collection of clinical text from EMRs provides a vast but still underutilized rich source of patient medical information in the real world. Exploiting clinical text in EMR is very challenging due to implicit expression of temporal information, heterogeneity nature, lack of structure and writing quality [8], [9]. Clinical text is special kind of text, rich in temporality and ambiguous in nature compared to general newswire text. Natural Language Processing (NLP) and machine learning perform a significant role in converting unstructured clinical text into structured representation which is inevitable. This structured illustration of text is easily accessible and understandable by machines.

Many researchers worked on various topics of temporal representation and reasoning with clincal data [10], [11], [12] [13]. A few researches in NLP over clinical text have been opened to address different exploitation strategies of extracted tempo-

[3] $https : //aclweb.org/aclwiki/Temporal_Information_Extraction_(State_of_the_art)$

ral information in medical care and research (develop applications by using temporal informations) [14], [15], [16]. Some of the applications, such as clinical decision support systems are exploiting temporal information [16].

To accelerate the temporal information extraction research works in clinical domain, I2B2[4] (Informatics for Integrating Biology and the Bedside) challenge [17], [18] provided an annotated corpora on temporal relations in the year 2012. The I2B2 workshop was the first pioneer along the collaboration of Mayo-Clinic generated the set of annotated and unannotated, deidentified patient discharge summaries for publicly available clinical dataset for research, which was the biggest barrier for NLP in clinical data [19]. This process made the preprocessed dataset available to interested research groups and submit their research findings and contributions in the I2B2-NLP challenge. Further they extended the dataset and challenges for interested researchers such as Obesity Challenge, Medication Challenge, Relations Challenge, Coreference Challenge, temporal Relations Challenge, and De-identification and Heart Disease Risk Factors Challenge. Consequently various research groups such as SemEval 2015 [20] and SemEval 2016 [21] contributed an annotated corpora on temporal relations and entities to develop the temporal information extraction methods. Bethard *et al.* adopted the temporal annotation guidelines from ISO-TimeML [22] and developed domain specific guidelines for the clinical domain, which is called THYME-Guidelines to ISO-TimeML (THYME-TimeML)[5], where THYME stands for Temporal Histories of Your Medical Events.

On the other hand, following with I2B2, ShARe CLEF 2013/2014 eHealth challenges from Physionet, Clinical TempEval 2015 [20] and Clincial TempEval 2016 [21] also provided annotated data for temporal information from clinical text. In japan, NII Testbeds and Community for Information access Research (NTCIR-MedNLP)

[4]https://www.i2b2.org/NLP/TemporalRelations/

[5]http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

provided the text for various topics of medical data exploitation [23], [24]. With available of all these resources, much effort has been done to develop different methods to extract temporal information from clinical narratives [17], [18].

1.2 Introduction to electronic medical records

The use of information and communication technologies is fundamental to accessing and helping health problems and challenges faced by patients. One of such recent emerged technology in health care is Electronic Medical Records (EMRs). EMR includes both hardware and software solutions and services, electronic devices such as mobile phones and applications, text messages, and clinic or remote monitoring sensors to note about the observation of patients.

Over the past decade, the implementation of Electronic Medical Records (EMRs) in hospitals has made the job easier for doctors to store and manage the patient medical history in one place. EMRs are digital version of paper documents that contains the clinical observation of a patient described by doctors and nurses. This evolution of EMRs has brought an abundant amount of digitalized longitudinal patient clinical records, which is considered one of the valuable asset affecting the success and progress of medical care and patient treatments in health care [8]. EMRs contains various informations such as clinical text, test reports, digitized images, X-Ray, and etc., among these information, the clinical text in EMRs are believed to be rich resource for major advancement to provide better treatment for patients in hospitals. Figure 1.1 shows an example of clinical observation of a patient in EMR noted by various departments. Especially, EMR contains the longitudinal information that describes about symptoms, disease, treatment and medication of an individual patient. Basically, this information composed of five main data components, which is explained below.

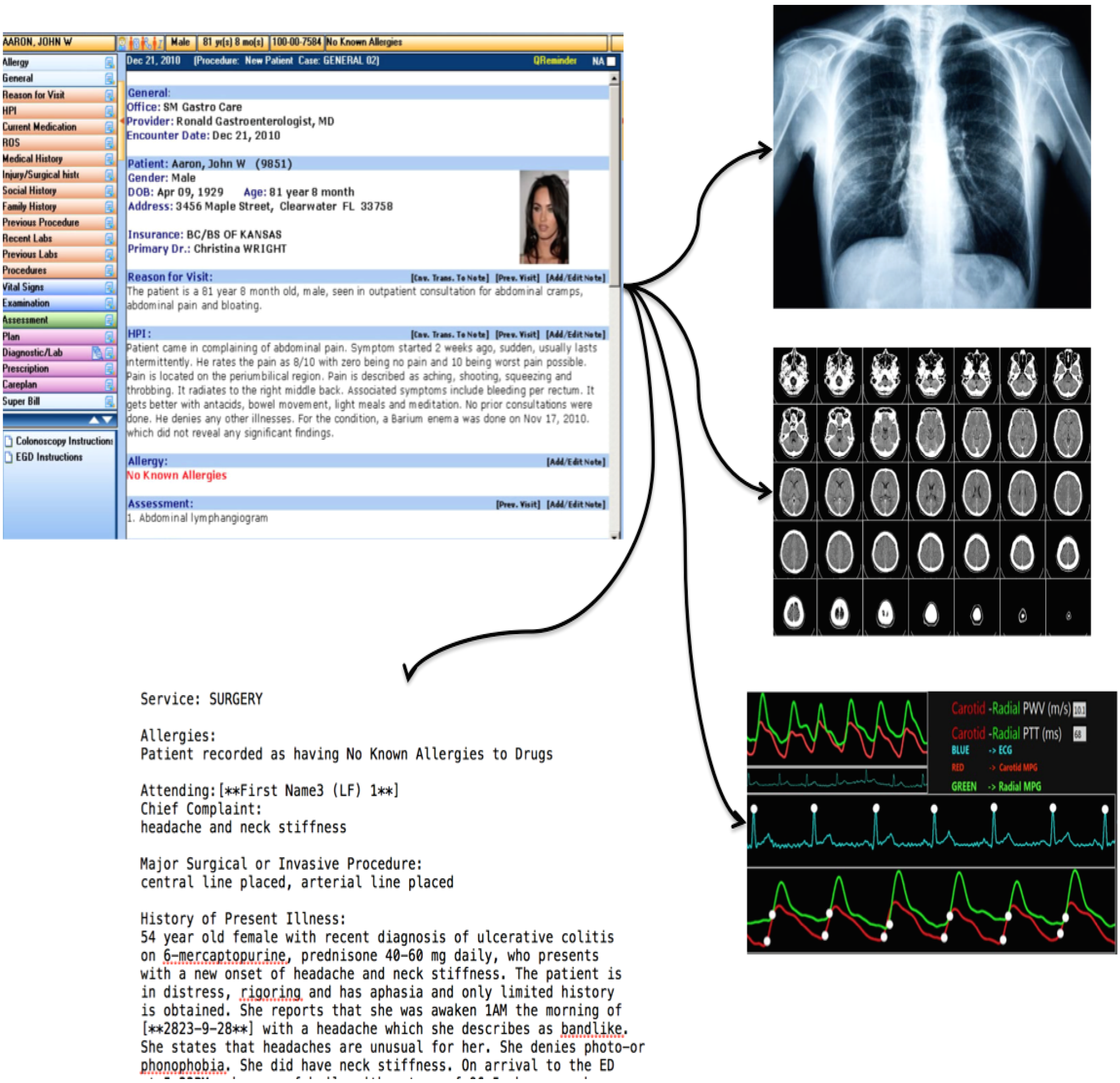

Figure 1.1: Electronic Medical Records: An Example

1. Demographic Information (or) Patient personal information such as patient name, gender, martial status and etc.

2. Physiological measurement (or) body function measurements. This includes two types of data

   (a) Numerical measurements such as body temperature and blood pressure

   (b) Signal measurements such as pulse rate, respiration rate and etc.

3. Laboratory test results such as blood test, urine analysis, etc.

4. Radiology images such as X-Ray, CT-Scan and MRI-Scan reports etc.

5. Clinical notes observed and noted by doctors and nurse which is known as clinical narratives

Although the development of EMRs has already provided an refined solution for efficient storage and data management, it just only reflects the primary usage which is data saving. The stored data from EMRs open new opportunities to support clinical studies for enhancing patient treatment methods and health care innovation. However, the exploitation is still in its infancy stage due to characteristics of clinical text in EMR.

1.2.1 Clinical narratives

Clinical narratives (or) clinical text in EMRs is a special kind of text that contains a lot of medical knowledge. This clinical narratives in EMR almost comes from notes of nurse and narratives of doctor daily assessments during their patient treatment. This clinical text is written in unstructured format using natural language which is consists of abbreviations, short forms and poor writing quality[]. One of the perspectives in clinical text exploitation is to analyze and evaluate patient health status through symptoms, disease, treat offered and conditions observed by a doctor analysis view which are collected and noted in the clinical text during the course of patient admission at hospital. Clinical text analysis demands to develop domain dependent natural language processing techniques to process due to several special characteristic of such text. Figure 1.2 shows clinical observation of a patient noted in EMR by doctor and nurse.

EMR data exploitation, especially clinical text exploitation is important to reduce medical errors through improving the patient treatment patterns. Clinical text is a large, rich and esteemed resource that contains information of patient's symptoms, disease, treatment observations known as phenotype informations. Doctors and nurses who have medical knowledge and experience in patient treatment note this clinical narratives. Therefore, credibility of the information in clinical narratives appeared EMR is more abundant and believable as well. Thus, these clinical narratives serve greatly promising resources to improve the patient health status and reduce cost through the effective treatments.

Discharge summary

Generally, discharge summary contains all the details such as patient previous history, current problems, medications used when he stayed at the hospital, what are the medications should be continued after discharging from the hospital and etc. But these details are written in concise (not explanatory) in contrast to the doctor daily notes and nurse narratives. Hence the step-by-step improvement of treatment effects and the disease progressions will not explained in detail as shown in figure 1.2. The content of a patient discharge summary is usually consist of multiple sections that describe the patient history of illness, chief complaint, social and family history, hospital course, allergies, and plan after discharge. Moreover the discharge summaries

Admission Date :
2016-08-08
Discharge Date :
2016-08-15
Discharge Date :
2016-08-15
HISTORY OF PRESENT ILLNESS :
The patient is a 37 year old lady with type 1 diabetes mellitus , who is four months post cadaveric kidney transplantation and now has good graft function . She presents for cadaveric pancreas transplantation . Her diabetes mellitus has been complicated by retinopathy and nephropathy as well as peripheral neuropathy . She takes 14 units of NPH insulin twice a day supplementing with a sliding scale .

HOSPITAL COURSE :
She underwent cadaveric pancreas transplantation without complication . She received induction therapy with thymoglobulin intraoperative and postoperatively for five days . She was kept on a similar immunosuppressive regimen as with her kidney transplant . She had excellent pancreas graft function immediately . Her renal function also remained stable in the perioperative period . She was quickly placed on a diet and advanced to regular diet . She was discharged home in stable condition on postoperative day six .

Figure 1.2: A clinical narrative from EMR

tend to be little bit structured narratives with mostly complete sentences, rich in medical terminologies and multiple temporal expressions.

Doctor notes and nurse narratives

Diagnosis of patients in digital medicine by doctor's based on detection of physiological positions and monitoring is the most important event. These positional and monitoring information sometimes gathered with the help of sensor or sensory device.

One of the key features of these sensor device is that patients can wear any part of the body and it regularly monitors the physiology of the organs in that area. Doctor's analyze the collected information and describes the health status as clinical narratives from their medical knowledge and experience. Moreover patients are treated based on collected and compiled information. For example electrolytes (or) sensors are used to monitor the health status of patients who are in the tumor or intensive care unit. Doctor's summarize the current status and decide the right treatment procedure for patient's from recorded electronic signals.

In other hand, It is the nurses who regularly care for the prescriptions given by the doctors and take care of their patients at the right time intervals. A nurse observes the intensive care unit patient with a intermittent (occurring at random or scheduled intervals of time) or constant time interval and write their observed information in EMR, which is called nurse narratives.

Doctor's noted and nurse narratives contains the text rich in temporal information yet to exploit in medical care for diagnose a disease, predict drug side effects and etc. Nurse will note down all the events like improvements of the treatments and progression of disease effects in constant time interval. Hence it has more valuable and significant temporal information, which is related to development of disease and symptoms. More importantly nurse will note down the treatment effect in patients, what kind of effect the patient has after using the particular drug. Detecting these information is unavoidable to exploit the EMR data in various medical care research and very useful for medical professionals to provide the treatment to the patients. Also disease progressions and treatment improvement details will help to predict the adverse drug reactions and post market surveillances. Since the drug are perfected based on the clinical trials.

At the same time, doctors conclude their diagnosis from their own observation, analysis from various tests and test reports, finally provide the appropriate treatment (prescribe medications and necessary test). Hence the clinical narrative contains the both frequent observation of nurse narrative and deepest analysis of doctor notes.

1.2.2 Opportunities and challenges

The opportunities and challenges on exploiting EMR clinical narratives are diverse. Some of the significant opportunists are described in below.

Opportunities

Clinical narratives in EMRs stored as a longitudinal observation of each patients that contains all medical details and treatment informations. The clinical narrative is considering as a promising sources that can help health providers as well as medical care researchers to understand more about nature of diseases, medical treatment effects and other significant insights to support health care development and innovation.

In one perspective of exploiting clinical narratives for the purpose on treatment improvements of patients. The key opportunity of utilizing the clinical narratives from EMRs is to analyses the effectiveness of treatment effect and discover drug side effects. Moreover this reliable clinical narratives has the potential informations that helps to discover the new usage indication of existing drugs or drug repositioning [25].

Another important opportunity of exploiting clinical narratives from EMRs is sentiment analysis and opinion mining. To answer the following questions *"What is the patinet health status now? or How effective of an ongoing treatment? "*, sentiment analysis is important for the clinical domain.

Sentiment analysis in clinical narratives has the purposes to support the assessment of patient status as well as assessment of treatment activities and courses. Thus, sentiment analysis play the rules in identify whether the treatment gives negative effect, positive effect, neutral outcomes or determine if patient status is improved or worsen or neutral conditions [26]. Sentiment analysis task is also remained with key challenges because of sudden sentiment shifter expressions. As mentioned in [27] clinical narratives contains plenty of negation statements such as "The patient did not report back pain", or "No evidence of edema or erythema", uncertain statements, example that "She is not sure if she really ready for chemotheraphy at this time ", etc. Thus natural language processing techniques found difficulty in analyzing these kind of sentiment expressions in clinical narratives.

One another significant chance of utilizing the clinical narratives from EMRs in temporal reasoning and temporal information extraction. The clinical narratives embedded with critical, precise and massive amount of temporal information described by doctors and nurses. Also it has a potential time orientation in through out the description. Moreover, this temporal information is crucial for constructing the patient similarity based on relation-based information from the medical database. In the work of [28] have developed a time-sensitive cancer tissue information retrieval system that helps to construct and answer user questions. Yet, the developed system is totally based on structured data (i.e.) the data consist of time-related information to determine the relation between the records. The usage of unstructured data with temporal information for temporal retrieval system is yet to considered in future research.

Applications of temporal reasoning in medicine

Temporal reasoning helps to nurture a wide range of NLP applications in medicine. For example, information extraction, question answering and information retrieval most widely developed tasks based on temporal reasoning. Basically, information extraction denotes task of identifying crucial information in natural language text, whereas question answering aims to respond of humans questions with appropriate answer. In case of information retrieval searches for similar patient record from stored electronic medical records in database. We will discuss each applications in detail below.

1. **Medical information extraction**

   As previously mentioned, EMRs contain massive amount of valuable informations in the form of unstructured clinical narratives which is distinct from general standard literature. Natural language processing technologies has been developed to extract key information from such clinical narratives such as Named Entity Recognition (NER) [29], medication information extraction [12], [30], relation extraction [31]. As all the information in clinical narratives have temporal boundaries, temporal reasoning emerged one of most challenging task for extraction of temporal information.

2. **Clinical question answering**

   - How the patient should be treated now?
   - what is the best medicine to cure migraines?
   - What does the patient experience after taking Naproxen ?

   Most of the questions have the temporal dimension. To answer such above questions, clinical information retrieval systems allow physician to quickly locate similar patient records that are specific to their information need from the available rich resources of clinical narratives in EMRs. The time-oriented information retrieval system demands new approaches to query and select most similar records for delivering higher quality care with short span of time and reduced cost. Various study has been conducted to meet the need of clinical query answering [32], [33]. Trivedi *et al.,* have developed a visualization tool with various components that helps to analyze the clinical records for domain experts and end users [34]. Also this tool allows to send the feedback if the correction is necessary for analyzed data.

3. **Temporal information retrieval**

   Information retrieval is a task of searching a relevant record from a collection of available documents. Traditional information retrieval used a group of key words to search relevant records, but it do not use time or temporal relation information. In recent times, temporal information based search has emerged one way of information retrieval system that satisfies the user needs more accurately. The central idea of temporal information based information retrieval utilizes temporal expressions along with the keyword indexing method to retrieve the more accurate and relevant documents [35]. Moreover, temporal information in clinical text is significant and important which is useful for temporal relation based information retrieval. Therefore, temporal reasoning is key task for temporal information retrieval applications in clinical domain.

Challenges

Though clinical narratives have many advantages and opportunities, exploiting the clinical narratives in EMR poses many challenges [36], but promising to make an innovation in health care. It is worth noting that the clinical text in EMRs is usually unstructured and written in ungrammatical way with full of shorthand and abbreviations that makes the process of exploiting clinical text is very different, even much more difficult, from standard text from the literature. As mentioned in previous section, EMR exploitation including various tasks:

1. Patient De-identification
2. Medical phrases identification
3. Abbreviation Disambiguation
4. Named Entity Recognition
5. Temporal information extraction
6. Data Representation
7. Sentiment analysis and opinion mining

Several domain-specific characteristics of clinical text makes the exploiting task more challenging. To promote and accelerate the development of techniques to solve the tasks in analyzing clinical text, shared tasks such as I2B2 challenges, ClinicalTempEval-2015 ClinicalTempEval-2016 for english since 2006 and NTCIR challenges for Japanese since 2013 have opened the stage for researchers. The characteristics of the clinical text are analyzed below in detail.

The clinical text is not followed grammatical structure as such general text and it composed of lot of abbreviations, short forms that makes the syntactic parsing and concept recognition more difficult and challenging task. As the clinical text is written quickly by the doctors and nurse, the structure and grammar is not considered carefully during the writing. Additionally, medical professionals note the patient clinical observation with lot of abbreviations due to time constraint. This abbreviations in clinical narratives rise the problem of abbreviation disambiguation and its restoration [37]. Abbreviation disambiguation from clinical narratives have to handle problem of sense inventories, which is considered two types. They are multiple abbreviation with one sense and one abbreviation with multiple sense. For example the abbreviation ARF have three expansions; Acute renal failure, Acute rheumatic fever and Acute respiratory failure. Unlocking the right expansion for the abbreviation is totally depends on the medical dictionary such as MetaMap, Mesh, UMLS and etc., in the existing study, various methods have been developed to deal with this problem [38].

Another challenge of utilizing the clinical narrative in EMRs is sentiment analysis and opinion mining. It is significant to understand the reaction of patients for the given treatment (i.e.) the effect of the treatment is reflected as a patient opinions. Hence, analyzing these opinions and sentiments will help to improve the quality of the treatment patterns.

Yet another crucial challenge in exploiting clinical narratives is detecting and extracting the implicit and vague time-oriented expression of medical concepts and temporal events.

Besides these above challenges, data representation is another challenge due to the short clinical text. This characteristic makes the task representation is difficult with lack of enough information.

Another significant characteristic of clinical text is comes from longitudinal, which is strongly related to time. As the clinical text is created by doctors or health care provider when the patients visits hospitals. Therefore, clinical narratives consists of patient health information over the time. This quality and quantity of time-series clinical information poses the challenge of temporal reasoning and temporal information extraction.

The last challenge in exploiting clinical narratives is its domain-specific nature. As the identified and processed information will be utilized to improve the medical and patient care, developing basic representation method is supposed to bring more accurate and meaningful results from medical domain knowledge than the normal text representation.

1.3 Representation of text data

To reconstruct the clinical narratives for further usage in clinical applications, the effective learning representations are required. Patient-level representation of clinical narratives helps to cohort patient records for wide range of tasks [39]. Some of the well-known representations for text data is discussed precisely in next subsections 1.3.1, 1.3.2 and 1.3.3.

1.3.1 Vector space model

In vector space models (VSM), each word represented by a feature. Therefore, the text is encoded by a vector, is having the equivalent length of the vocabulary count in the corpus of a domain. For clinical domain, as the number of terminologies is up millions, thus the length of vector representation will be tremendous. This limitation leads to the problem of dealing with high dimensional data, a very challenging in text mining. Moreover, representing a piece of short clinical text in vector space models will result in a vector with many zero values. Dealing with sparse vectors is another challenge that we have to encounter when applying vector space models [40].

Let us consider the following example sentences from clinical narrative to understand the traditional vector space representation with binary weights for each term.

1. *Sentence 1: The patient reports nausea and vomiting.*

2. *Sentence 2: The patient complained vomiting and headache.*

The traditional vector space representation learns from all given words in the sentence, and each unique word represents a binary value based on co-occurrence as shown in figure 1.3. In case of sentence level, all the words in each sentence are concatenated and represented with binary weights as shown in figure 1.4.

	The	patient	reports	nausea	and	vomiting	complained	headache
The	[1 ,	0 ,	0 ,	0 ,	0 ,	0 ,	0 ,	0]
patient	[0 ,	1 ,	0 ,	0 ,	0 ,	0 ,	0 ,	0]
reports	[0 ,	0 ,	1 ,	0 ,	0 ,	0 ,	0 ,	0]
nausea	[0 ,	0 ,	0 ,	1 ,	0 ,	0 ,	0 ,	0]
and	[0 ,	0 ,	0 ,	0 ,	1 ,	0 ,	0 ,	0]
vomiting	[0 ,	0 ,	0 ,	0 ,	0 ,	1 ,	0 ,	0]
complained	[0 ,	0 ,	0 ,	0 ,	0 ,	0 ,	1 ,	0]
headache	[0 ,	0 ,	0 ,	0 ,	0 ,	0 ,	0 ,	1]

Figure 1.3: Matrix of word representation with binary weights

1.3.2 Topic model

Due to several limitations in traditional representation techniques that work on general text such as vector space models seem not work with clinical text. In search of other representation for clinical text in EMRs data, topic models have some advantages to deal with the problem. Basically topic models are assuming a patient record with a mixture of topics, and a topic is a mixture of words. The concept of topic

	The	patient	reports	nausea	and	vomiting	complained	headache
Sentence 1	[1 ,	1 ,	1 ,	1 ,	1 ,	1 ,	0 ,	0]
Sentence 2	[1 ,	1 ,	0 ,	0 ,	1 ,	1 ,	1 ,	1]

Figure 1.4: Matrix of sentence representation with binary weights

model is to represent the words in a hierarchical structure like the way human imagine based on the distribution weights of each word in the document. The graphical representation of topic model for clinical narratives is shown in figure 1.5.

Several works on exploiting clinical text in EMR using topic models with different level of domain incorporation have been studied in the past history. Sarioglu et*al.* have developed a disease-medicine topic model to discover insight knowledge about diseases and medicines [41]. In the study [42], employed a topic model to identify patterns of clinical events in a cohort of brain cancer patients. Though, above works achieved some impressive results, almost little or no medical domain knowledge was embedded in the developed model. However this works considered as the first attempt to bring medical knowledge incorporation to represent the clinical text in the form of topic model.

Even though, it represent the text in hierarchical structures, it misses to discover a temporal information (i.e.time-orientation) which is significant in medical care for disease diagnosis and proposing treatments. Thus, the clinical text demands another powerful representation that depicting time information with hierarchical structure to understand the longitudinal nature of clinical narratives.

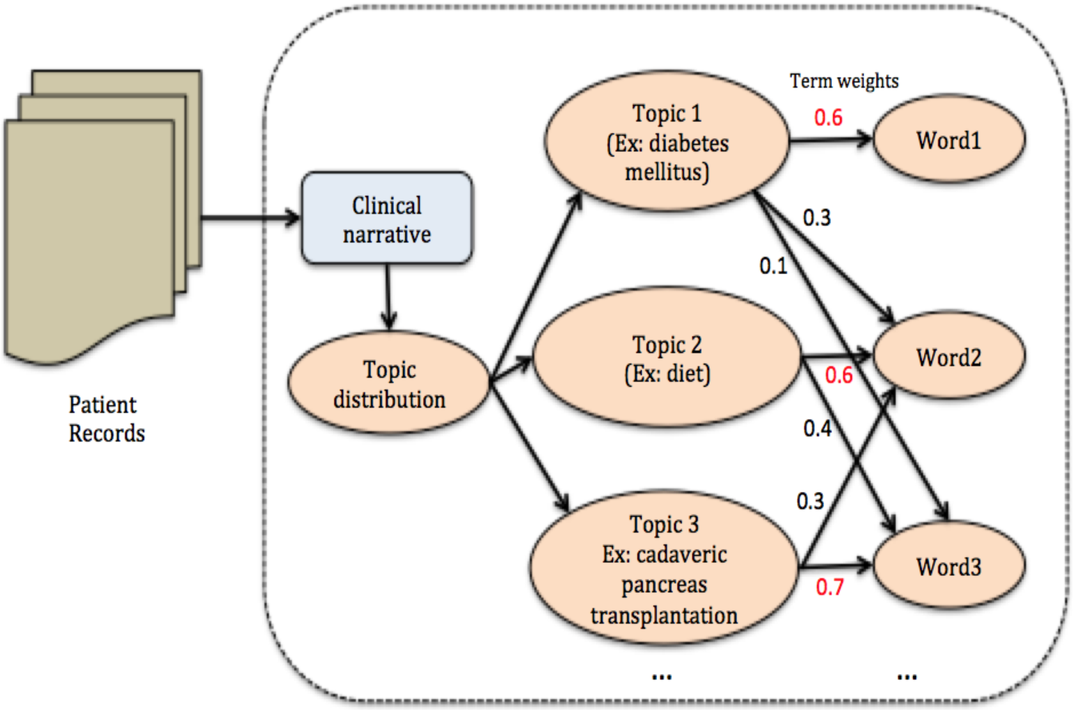

Figure 1.5: Topic model representation for a clinical narrative

1.3.3 Temporal information

The patient clinical information in EMRs have mainly been stored in the discharge summaries, doctor daily notes and nurse narratives, in natural language text. Clinical text contains the patient disease symptoms and disease progress, doctors notes and treatments [43], [13]. In the above sources, most events happened during the patient stay at the hospital are associated with notion of time (dates and time) that linked to the patient disease signs and symptoms and treatments. Therefore, processing of temporal information in clinical text plays a potential role in exploiting clinical text in EMR and medical care to support and develop disease-diagnosing system [16]. For these reasons, there is an increasing interest in temporal information extraction over the clinical text. However the exploitation of clinical text in NLP is still in early stage due to several difficulties in processing the clinical text [44]. Also the utilization of clinical text in EMR is complex task due to the nature of the data.

Temporal information performs the crucial role in interpretation of the patient clinical information such as progress of disease and frequencies of medication information. Though the human can able to understand the temporal information such as events and temporal relations between the events efficiently to interpret the progression of disease status and prognosis, temporal information extraction remains the non-trivial task in NLP.

Basically temporal information in clinical narratives consist of three components. The first and foremost element of temporal information is temporal expressions. Temporal expressions in clinical narratives such as "admitted on Friday", "7 weeks prior to admission", "Twice a day" and a special pre-post expression of "four months post transplantation". Let us also observe this expressions in an history of illness portion from a sample discharge summary have shown in Figure 1.2.

The second element of temporal information is temporal events. In the context of clinical setting, the disguising terms for describing the patient health status with occurring incidents, activities, entities with states are considered as medical events. Moreover events in the clinical narratives are restricted to noun phrase or clinical concept such as laboratory tests, medical problems, proposed treatments, diagnoses, patients complaints among others [45], [46]. In other simple words, any diseases, symptoms, tests, medications and conditions related to the patients health status is considered a medical event.

Let is consider the example discharge summary from figure 1.6 to examine the temporal expressions and events as highlighted.

The third element of temporal information is determining temporal relations with expressions and temporal boundaries of events. This relationship can be classified one among the 13 types of interval algebra proposed by Allen's [6]. The

Admission Date :
2016-08-08
Discharge Date :
2016-08-15

HISTORY OF PRESENT ILLNESS :
The patient is a 37 year old lady with type 1 diabetes mellitus , who is four months post cadaveric kidney transplantation and now has good graft function . She presents for cadaveric pancreas transplantation . Her diabetes mellitus has been complicated by retinopathy and nephropathy as well as peripheral neuropathy . She takes 14 units of NPH insulin twice a day .

Figure 1.6: A clinical narrative: Temporal information representation

event from the example “cadaveric kidney transplantation” has been occured BEFORE entering the hospital for admission, whereas the event “cadaveric pancreas transplantation” is AFTER the event of ADMISSION.

From the above figure, the clinical narratives have the following temporal expressions: “four months”, “twice a day” and “2016-08-08” and “2016-08-15”. Among these, “2016-08-08” and “2016-08-15” can be classified to two types of expressions: (1) it can be considered DATE expressions, (2) on the other hand it is considered SECTIME expressions as ADMISSION and DISCHARGE expressions.

The medical problems, diagnosis, treatment, medication information in clinical narratives considered as temporal events. For example “type 1 diabeted mellitus” considered as diagnosis event, “cadaveric kidney transplantation” considered as treatment event, and “14 units of NPH insulin” considered as medication event. However all these events are covered with temporal boundaries as we explained in temporal expressions.

Therefore, it is obvious to solve the problem of temporal information extraction demands the following sequential tasks which is shown in figure 1.7. Finally these extracted temporal informations are exploited to generate the chronological order of events and expressions, consequently visualization tool such as SMILE tool for visualizing of clinical events and expressions [6]

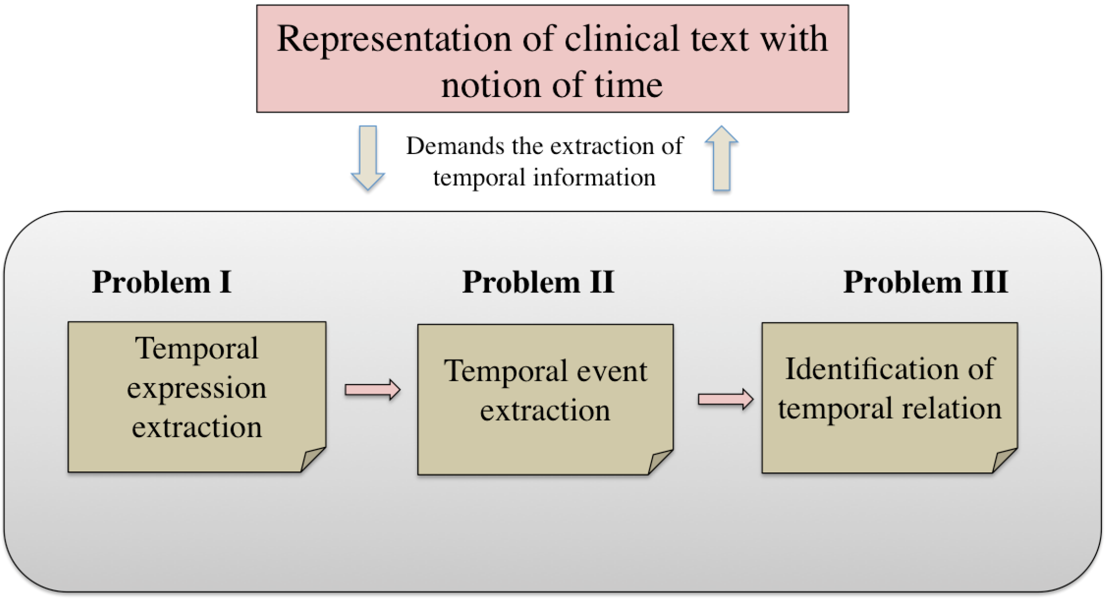

Figure 1.7: Elements of temporal information extraction

1.4 Problem and objective of our study

1.4.1 Motivation and problem formulation

Data representation of clinical text with notion of time, demands to extract temporal expressions, events and temporal relations. Temporal expression is a fundamental element to represent rest of the temporal informations. Moreover, temporal events extraction and temporal relation detection are challenging task and basis for temporal ordering of events in clinical timeline. Especially, discovering the temporal order between the clinical events and expressions from raw clinical text is a crucial part

[6] $http://www.simile-widgets.org/timeline/$

of temporal information extraction. Therefore in this thesis we address the following problems.

1. Extracting absolute and relative temporal expressions from raw clinical text
2. Extraction of temporal events by exploiting unannotated data using semi-supervised methods
3. Discover the temporal relations between temporal expressions and events

1.4.2 Scope of our study

Our work has target to support a study area on automatic reconstruction of clinical text from electronic medical records toward structured representation of unstructured clinical text. In the temporal information extraction, main challenge is to recognize and extraction of clinical events with time expressions, then confirm and classify the crucial relationships among them. Therefore, in this study, we focus to extract temporal expressions, events and classify the relationships among expressions and events.

1.4.3 Major contributions

The main contributions of our thesis have been discussed in this section. We fragment the task of temporal information extraction into multiple sub tasks and solve the problems by developing new methods.

- The first contribution comes from temporal expression extraction. A novel feature set has been proposed to address the problem of temporal expressions extraction. New proposed feature set is obtained from raw clinical text and adopted HeidelTime features that are appropriate for temporal expressions extraction from clinical text.

- The second contribution stemming from temporal event extraction from clinical text. We established a novel semi-supervised framework to exploit abundant unannotated data for extracting temporal events from clinical text. To best of our knowledge and survey from literature, this work is the first to propose semi-supervised method for extracting temporal event from clinical text. This approach innovated the novel idea of gradually extending the training corpus by adding annotated data obtaining from unannotated clinical text. This is the most significant contribution of the thesis to overcome the limitations of the general methods when being applied to the clinical text.

- The third contribution is from temporal relation identification and classification. We formulated a new assumption on generating and identifying the potential candidate pairs from list of temporal events or expressions that can appropriately relate events/expressions in clinical text. Moreover, to address the problem of temporal relation detection, we exploited Naive Bayesian Classifier to detect the temporal relationship among the identified pair's. The effective candidate pair's generation helps to improve the relation classification performance.

1.4.4 Thesis structure

Our thesis consists of 6 chapters. Figure 1.8 shows the graphical representation of our dissertation. The first chapter introduces about the clinical text in EMRs and its challenges, the problem and objective of our research.

In Chapter 2, we give introduction to temporal information extraction, discuss about the importance of TIE and benefits of TI exploitation in medical care and

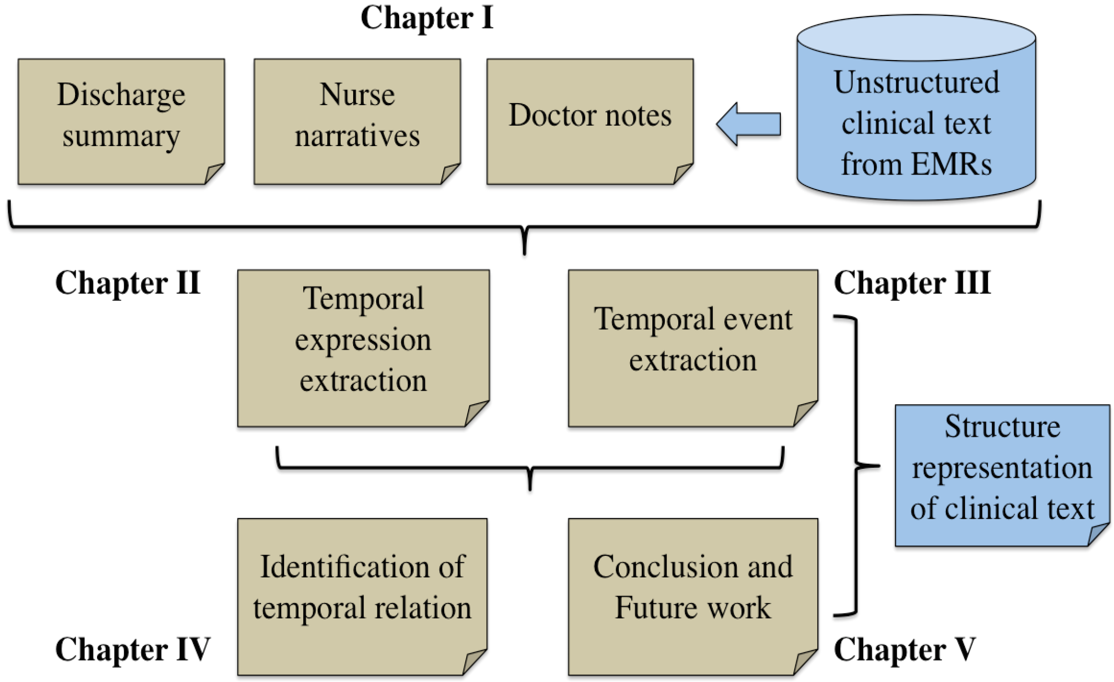

Figure 1.8: Graphical Representation of our Dissertation

research. Then, we will discuss the similarities and differences of temporal information in general and clinical text.

Chapter 3, gives the overview of temporal information and three keys tasks of our research; temporal expression, temporal event extraction and temporal relation classification. Also this chapter focuses to discuss about temporal expression extraction from clinical text by using proposed method with exploiting different combination of feature sets.

Chapter 4, discuss about the extraction of temporal events extraction in clinical text from EMR. Firstly, we have done the deep literature review for the task and moved to the problem formulation and later we discussed about the proposed method to solve the problem. Especially, we described the proposed semi-supervised framework method to extract the temporal events by exploiting abundant amount of unannotated clinical text.

Chapter 5, discusses and establishes the methods for detecting temporal relation between pair of events or expressions. This chapter focuses on methods to generate the candidate pair's from extracted temporal expressions and events. At the end, we will discuss about the temporal relation identification and classification task for temporal relation detection.

In last chapter, we present the detail discussion of our study on temporal information extraction from clinical text and conclusion. Later we discuss future research aspects to our study to develop representation and visualization techniques and applications by exploiting the temporal information.

Chapter 2

A brief history of Temporal Information Extraction (TIE)

This chapter first introduces the temporal information in clinical narratives and its significance. Then, the requirement of annotated copora for temporal information extraction and difficulties in obtaining them. The next part provides a challenges in extracting temporal information from clinical narratives. Last part discusses about the current status of temporal information extraction in clinical narratives.

Temporal reasoning is a very basic yet vital ability of humans to understand the information that we are speaking. [47] temporal reasoning methods to represent the events, state and its positions in their flow of time. Moreover, temporal reasoning is very important for clinical narratives to know and understand the clinical events happened in the patient medical history. General text like newswire articles use the features like tense and aspect to encode temporal information. However, the importance of these features are depends on specific domains. One such area with distinct characteristic of domain-specific text and temporal information is medicine, especially clinical narratives from EMRs.

The central usage of clinical narrative in medicine is to reason the time-oriented data. Continuously observing the clinical events over time with frequently distribute the informations that are required for clinical decision making system, including diagnosis, prognosis, treatment regimen recommendation [48] and planning for patients [49], [16]. The clinical decision support systems aims to help the health-care provider by seeking knowledge from stored clinical data whenever it required [50].

2.1 Introduction to TIE in clinical narratives

Understanding temporal information has become most significant for several language processing applications such as information retrieval, query processing, text processing and various text-mining tasks. To accomplish temporal information processing tasks in NLP, it is important to develop strong annotation standards and corpora for temporal information. Automatic recognition and extraction of essential time-oriented information from clinical text in natural language have become an active area of research in computational linguistics from past decade. Moreover the recent prevalence of clinical text in Electronic Medical Records (EMRs) receives a great attention from researchers due to its significance and rich patient clinical information. The huge collection of EMRs provides a vast but still a underutilized rich source of patient medical information in the real world. However exploiting clinical text from EMR is very challenging due to its heterogeneity nature, lack of structure and writing quality [8], [44], [9].

One major side of exploiting clinical text EMR data is clinical natural language processing and temporal information extraction (TIE). Temporal information extraction indicates the crucial dimension of time-related information extraction from clinical text, which is different from general newswire text [9]. In EMR data, most events happened during the patient stay at the hospital are associated with time stamps (date and time). Basically the clinical text data is unstructured, ungrammatical and medical practitioners represent the temporal information directly or indirectly in clinical text when they explain the symptoms, disease, disease progress, occurrence and treatments.

Most importantly the clinical text is usually written in natural language with unstructured format that contains lot of abbreviations and short forms. As most of the valuable information in EMRs exists in clinical text, encoding such text is intrinsically challenged by issues of text representation [51]. Therefore, it is necessary to represent EMRs into computable forms before using them for further applications in medical care [52]. It is worth noting that clinical text from EMRs is distinguished from non-clinical texts by its domain-specific nature, abbreviations and short forms, which makes the problem of data representation much more difficult [51].

Several computable forms has been developed to reconstruct the clinical patient records [17], [45], [53]. Among these various techniques and data representations, temporal information representation emerges as the most important form since all the clinical events and medical information are noted with time-stamp and indirectly follow a temporal order in the nurse narratives, doctor daily notes and discharge summaries [45]. Thus, the lose of chronological order of clinical events could cause negative effect in disease diagnosing and lead to serious medical errors as well as affect the patient treatment procedure and results [54].

“Is there any improvement in the patient status? ”

This question can only be answered after identifying and extracting the temporal events and relationships from clinical text related to the patients. Therefore, being able to identify and extract temporal information is a very crucial phase and

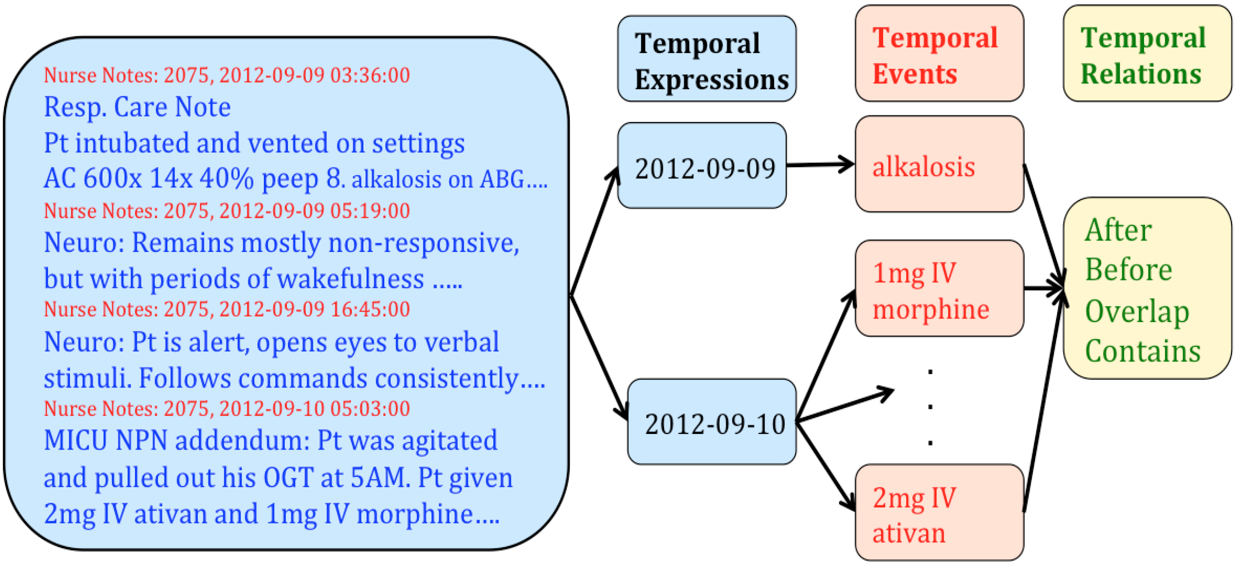

Figure 2.1: Key components of temporal information in clinical text

it becomes the priority task for structured representation of clinical narratives from EMRs.

Temporal information extraction consists of three key components. These are extraction of temporal expressions, temporal events and temporal relations which clearly shown in figure 2.1. Temporal expressions are sequence of words and phrases that expresses a point of time or spans on a timeline such as date, duration, time and frequency. Temporal expression is a fundamental element for the rest of the temporal informations. Temporal events in clinical text is medical incidents that happened in a patient medical history which can be related to clinical timeline. Temporal relations denote the relationship among the pair of temporal events or temporal expressions.

Automatic recognition and extraction of temporal information extraction demand to extract those three elements. Among them temporal expressions are basis to represent other two temporal elements.

Let us consider the following sentence and the list of events and expressions

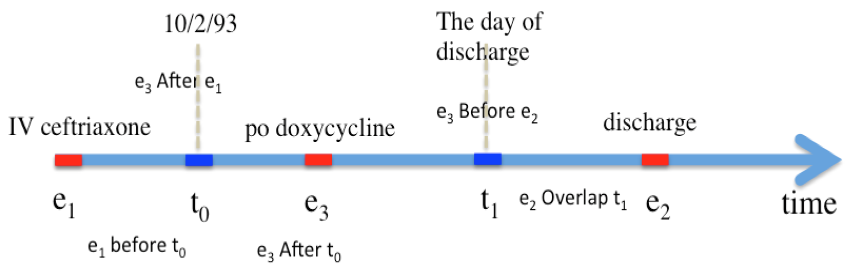

Figure 2.2: Generating chronological order of temporal expressions and events: an example

extracted from the sentence. *The patient was started empirically IV ceftriaxone on 10/2/93, which was changed to po doxycycline on the day of discharge.* The list of temporal expressions and events are the followings, accordingly:

– {10/2/93, the day of discharge}

– {IV ceftriaxone, po doxycycline, discharge}

To chronologically arrange all clinical events as shown in figure 2.2, all the clinical events and expressions should be grounded from the raw clinical text. The expressions are often mentioned in clinical text with the usage of relative time. Also the clinical events and relations often mentioned without referring/specifying of the time and order, respectively. For instance, consider the following sentences:

- "The patient was transferred to the hospital for surgical intervention on *2016-06-11.* The *next day* she administrated a drug *one hour before* the surgery."
- "The patient *consulted* the *toxicologist* and he felt that most likely had *benzo overdose.*"
- "The patient has been admitted to the hospital with frequent severe cold, fever and cough."

The above examples show that the temporal expressions, events and relations have often written with relative time, implicit durations and order. To automatically understand this temporal medical concepts from clinical text by machines, it requires the considerable medical domain knowledge.

Though the human can be able to understand the temporal information such as events and relations between the events, whereas it remains the non-trivial task in machine learning and clinical natural language processing. Despite of these difficulties, the existing works [17], [18], [55] established various methods to extract temporal information with the help of annotated corpora. But all of these works were established with the available small number of annotated temporal corpora, whereas lot of unannotated clinical text are available compared to annotated clinical data.

2.2 Temporal information annotation in clinical narratives

As explained in the previous section, addressing the problem of temporal relation learning with temporal reasoning by machine learning methods demands annotated corpora for training models and evaluating results. However, creating such temporal information annotated corpora from clinical narratives is tedious, expensive and requires much more manual effort with medical domain knowledge. At the same time using doctor's expertise to create annotated corpora is cost expensive.

2.2.1 Motivation

Clinical text annotation is an important task for information extraction [46]. Annotation of clinical text adds more semantic information to the original text document [43],

[56], which can be used later for some kind of temporal information tasks . Temporal annotation plays a significant role in natural language processing and computational linguistics. Successful annotated corpus enables the researchers to perform many tasks such as information extraction, Named Entity Recognition, relation extraction and etc., in many domains such as social media and medical domain. The created corpus addresses the problems such as processing and ordering temporal events, determining temporal relations between the events and temporal expressions.

Rapid development of temporal reasoning and temporal information extraction (TIE) methods initially started with the creation of TimeBank corpus for general newswire text [1], [7]. TimeBank corpus annotated a temporal reasoning elements by using TimeML specifications and guidelines [1]. It is a typical and most popular corpus for temporal reasoning and relation learning in the natural language text. However, temporal reasoning and annotation process can vary newswire domain to clinical domain because of nature of data. Therefore, it is difficult to use the TimeBank guidelines as exactly to annotate the clinical narratives from EMRs. the differences in the nature of data between Timebank and clinical text make it difficult to use Timebank for temporal relation learning in the clinical domain.

The events in linguistics are tuned to be verb tenses and aspects, standard form of some verbs, certain proper names. Sometimes the state of events that are changed by verbs. For example, the events in TimeBank from newswire domain, relies on tense and aspect are to temporally order events, which are used as a key feature to detect them by machine learning methods [57]. In contrast, events in clinical domain need not be verbs and are in fact in most cases noun phrases as all the diagnosis details noted as noun phrases. Therefore, the annotation of temporal events in clinical domain and newswire domain are quite different [45], [56].

[1] $https://catalog.ldc.upenn.edu/docs/LDC2006T08/timeml_annguide_1.2.1.pdf$

In recent days, the interest of temporal reasoning and temporal information extraction from domain-specific clinical narratives has received great attention due to richness of temporality in clinical narratives, significance of temporal information exploitation in medical care and availability of clinical text. Recently, Temporal Histories of Your Medical Events Time Markup Language (THYME-TimeML) project extended the standard text annotation work and created guidelines specifically for clinical domain text. Also they have created the corpus for clinical text. Recently I2B2 created the clinical temporal information corpus with 310 de-identified discharge summaries. There are two methods available to annotate the clinical text data. First is annotate the clinical text manually which is based on the domain knowledge of annotators. Yet, considerable manpower is often required to annotate the corpus manually. This manual annotation process usually involves multiple groups and finally we get the error-pruned corpus. Second is automatic annotation, which is based on the annotation tools developed by various research group. However, accuracy of the annotation tool is very important to annotate temporal information in clinical narratives. Manual intervention and correction is required along with the accuracy of the annotation tool.

Due to this difficulties, only very little amount records annotated records with temporal information such as I2B2 (Informatics for Integrating Biology and the Bedside) dataset (Has 310 patient discharge summaries) and Thyme clinical dataset despite a lot of medical records available online for Natural Language Processing such as MIMIC II clinical database.

We analyze the nature of clinical text in a discharge summary, to understand more about annotation process and creation of annotated corpus from clinical narratives as shown in figure 2.3.

Admission Date : 2016-08-08
Discharge Date : 2016-08-15

HISTORY OF PRESENT ILLNESS :
The patient is a 37 year old lady with type 1 diabetes mellitus , who is four months post cadaveric kidney transplantation and now has good graft function . She presents for cadaveric pancreas transplantation . Her diabetes mellitus has been complicated by retinopathy and nephropathy as well as peripheral neuropathy . She takes 14 units of NPH insulin twice a day .

HOSPITAL COURSE :
She underwent cadaveric pancreas transplantation without complication . She received induction therapy with thymoglobulin intraoperative and postoperatively for five days . She was kept on a similar immunosuppressive regimen as with her kidney transplant . She had excellent pancreas graft function immediately . Her renal function also remained stable in the perioperative period . She was quickly placed on a diet and advanced to regular diet . She was discharged home in stable condition on postoperative day six .

Figure 2.3: Illustration of a patient discharge summary

Clinical narratives: Characteristics As seen in the figure 2.3, discharge summary is composed of various different sections such as history of illness, hospital course, past medical history, allergies, and etc. In comparison to other clinical text in EMR (doctor's daily note and nurse narratives), discharge summaries is composed of little bit structured format with mostly complete sentences. However, it also contains considerable use of medical terminology, and multiple temporal references. However, doctor notes and nurse narratives are less structured but also have the same characteristics of discharge summary. Thus, it is making the task of automatically extract information from clinical narratives (discharge summary, doctor notes and nurse narratives). In order to automatically extract temporal information and ordering them is required annotated data. For this, clinical narrative information should be simplified to create features to facilitate the extraction of events, temporary expressions and temporary relationships. Thus, we define an annotation specification that supports the following tasks:

1. detection of events, expressions in clinical narratives. For example, cadaveric kidney transplantation and retinopathy , 2016-08-08, postoperative day six, four months.

2. Identify temporal relations between events and / or expressions that occur within and across all clinical narratives for a patient. For example, cadaveric kidney transplantation and retinopathy events before admission event, whereas cadaveric pancreas transplantation is after 2016-08-08 expressions and before discharge event.

The tasks of unlocking the clinical elements [58] that are discussed above is to support for generating a chronology of clinical events from across all clinical narratives for each patient [59] . The data used for annotation and training our proposed methods and models is described next.

2.2.2 Annotation of events in clinical narratives

As per the annotation guidelines [2], clinical events include diseases and disorders, symptoms that affects the patient's health status, as well as any treatments, procedures and medication prescribed to the patient. Thus, domain expert annotators were come to conclusion to annotate a word or group of words occurred in a clinical narrative of a patient that has a relevant match in the UMLS medical dictionary as a clinical event. Events in clinical domain includes the following.

1. First and foremost any disease name or disorder. These include symptoms which are typically nouns. For example, heart attack, chest pain, breast cancer.

2. Second is any treatment, test or procedure. These are again generally nouns. For example, blood pressure, temperature and excisional biopsy.

[2] http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

3. All the medication details (i.e.) any medications consumed or used by patients. These are typically denotes the names of drugs such as aspirin and Naproxen.

4. Besides the above, any normal health condition that requires constant health care such as pregnancy.

5. Any habit or observation that affect the patient health status. For instance, drinking and smoking.

2.2.3 Annotation of temporal expressions in clinical narratives

Temporal expressions are sequence of words or phrases that contain time information. The types of temporal expressions that includes date, time, duration, frequency and prepost expressions [43], [45]. TIMEX3 tag is used to annotate a temporal expressions in clinical narratives and it could be any one of the following:

1. DATE: denotes the exact calendar date. For example, Nov 13 Monday, 2016-08-08.

2. TIME: denotes a time of the day. For example, 8:00 AM, seven in the morning. *The patient's CT-scan is scheduled at 10:00 AM*

3. DURATION:Describes a time duration. For example, four months, next two days.

4. FREQUENCY: Describes a set of times. FOr example twice a day.

5. QUANTIFIER: describes like frequency expressions but it does conceive starting point and ending point. For example twice and three times. *The patient vomited "twice" before the treatment.*

6. PREPOSTEXP: The time span of this expressions is implicitly related to event. Therefore, this kind of expressions in clinical narratives are marked with special type as PREPOSTEXP. For example post-surgery.

2.2.4 Annotation of temporal relations in clinical narratives

TLink tag is used to annotate temporal relations between events and/ or expressions. It contains an important information as it relates the previously annotated event and expressions in a meaningful manner to identify relation between them. The temporal relations are adapted from Allens temporal relations as explained in section 1.3.3 and used in Chapter 5. For example of temporal relation annotation on following sentence "The patient was admitted with type 1 diabetes mellitus who is four months post cadaveric kidney transplantation and now good graft function" would be:
<EVENT1=type 1 diabetes mellitus — EVENT2=postcadaveric kidney transplantation— Tlink-type=OVERLAP >
<EVENT2=postcadaveric kidney transplantation — TIMEX3=four months — Tlink-type=AFTER >
<EVENT2=postcadaveric kidney transplantation — EVENT3=good graft function — Tlink-type=OVERLAP >

In this example, EVENT2 is overlapping with EVENT1 and EVENT3, at the same time AFTER the Temporal expressions TIMEX3.

2.3 Challenges of temporal information extraction in clinical narratives

Human being has been developed the ability to reason the natural language text including the complex and implicit informations over the centuries. But, in machine learning and clinical natural language processing, the reasoning of natural language and clinical narrative for temporal information remains an non-trivial task [9], [60], [61]. The several challenges of temporal information extraction form clinical narratives have been shown in figure 2.4 and discussed in detail below.

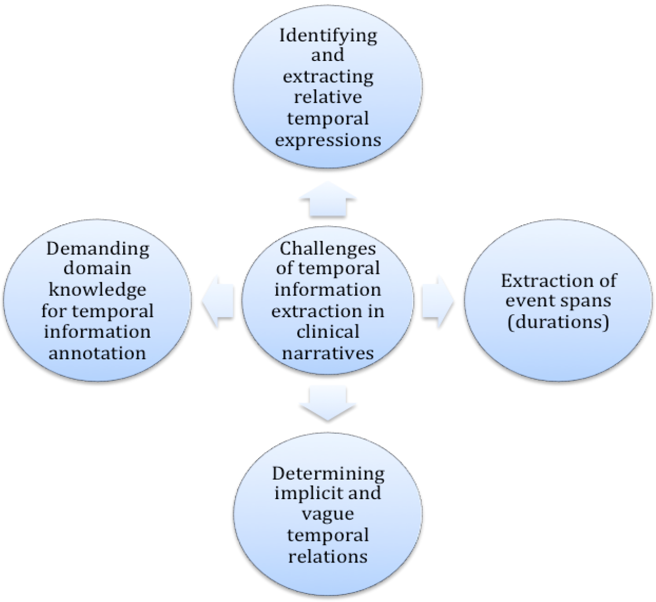

Figure 2.4: Challenges of temporal information detection and extraction

- Temporal relation in clinical text often implicit and vague. The temporal relationship between the temporal events are not noted apparently described by medical professionals in clinical narratives. Therefore determining the order of events is difficult. Let us consider the following example to examine the implicit and vague temporal relations in clinical narratives.

The patient complained cold and fever. He was admitted to the hospital.

From the above example, human can reason the sentence and understand that the patient was sent to the hospital since he complained about his illness. But it is very difficult for a machine to understand and reconstruct the clinical narratives. Moreover, the patient complained two symptoms in the example sentence and it looks like both the symptoms started at the same time. In that case it is difficult to find and say the cold is started before the fever. Consequently, reasoning the end time of symptoms (i.e.) determining whether the symptoms are lasting until now or not.

Vagueness come from the writing style of the clinical narratives. Temporal information can be expressed with time, aspect and tense. The writing quality of clinical narratives is poor since the medical details and facts in clinical narratives are described very compact manner.

- Extraction of relative times

  -Difficult to extract relative time expressions - Ex: The patient administrated the drug one day before the surgery

  Temporal expressions in clinical narratives are frequently written in reference to other time expressions. For example, that night (relative to the date when this phrase was written) and one day before surgery (relative to the date of the surgery). However, temporal reasoning needs to determine absolute date expressions (the exact dates) of relative temporal expressions. Therefore, extraction of temporal expressions fundamental, yet non-trivial task in temporal information extraction

- Extraction of event durations or event spans

-Extracting the completion time of the event is an challenging task Ex: The patient met a doctor

Event duration is important to discover and determining the end points for temporal relations. Sometimes these event durations are mentioned implicitly. From the example, it is difficult to determine *whether the event last for one hour or 10 minutes*

- Annotation corpora for temporal relation

  Unlike creating the annotated corpus for other NLP tasks such as NER, creating qualified corpora for temporal relation is a challenging task. since the relations are expressed implicitly in the clinical text. Also the annotating temporal relations are application specific. Also it demands the domain knowledge to create an annotated corpora for temporal relation. However, to develop stable and reliable automatic temporal information extraction models demands the large annotation corpora like TimeBank. However, the creation of such annotated corpora from clinical narratives requires more time, cost and most importantly huge manual effort. Moreover, the annotation process of clinical narratives demands of medical expertise and domain knowledge for annotator to annotate temporal information.

2.4 Current status of temporal information extraction in clinical narratives

Rapid development of temporal information extraction (TIE) methods initially started with the creation of TimeML corpus for general newswire text [1], [7]. This corpus has three types of temporal information: events, temporal expressions, and temporal relations. Detailed information about these corpora can be found at timeml website as mentioned earlier. For event extraction from text, machine learning methods widely used and have shown good performance, including conditional random fields (CRFs) and supported vector machines (SVMs) [62]. For temporal expression extraction both machine learning and rule-based methods have been investigated [62].

Recent years the interest of temporal information extraction from clincal text has received much attention from researchers with the implementation of Electronic Medical Records (EMR) systems. Many researchers worked on various topics of temporal representation and reasoning with clincal data [10], [11], [12] [13]. A few researches in NLP over clinical text have been opened to address different exploitation strategies of extracted temporal information in medical care and research (develop applications by using temporal informations) [14], [15], [16]. Some of the applications, such as clinical decision support systems are exploiting temporal information [16].

To accelerate the temporal information extraction research works in clinical domain, I2B2[3] (Informatics for Integrating Biology and the Bedside) challenge [17], [18] provided an annotated corpora on temporal relations in the year 2012. Consequently various research groups such as SemEval 2015 [20] and SemEval 2016 [21] contributed an annotated corpora on temporal relations and entities to develop the temporal information extraction methods. Bethard *et al.* adopted the temporal annotation guidelines from ISO-TimeML [22] and developed domain specific guidelines for the clinical domain, which is called THYME-Guidelines to ISO-TimeML (THYME-TimeML)[4], where THYME stands for Temporal Histories of Your Medical Events. Annotation of clinical text adds more semantic information to the original text document [63], [56], which can be used later for some kind of information extraction tasks.

Reeves *et al.* extended temporal awareness and reasoning systems for ques-

[3]https://www.i2b2.org/NLP/TemporalRelations/

[4]http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

tion interpretation (TARSQI) toolkit to detect the temporal expressions in the clinical text [64]. But it performed very poorly as the temporal expressions in clinic notes were very different from those in the newswire domain [65]. On the other hand, following with I2B2, ShARe CLEF 2013/2014 eHealth challenges from Physionet, Clinical TempEval 2015 [20] and Clincial TempEval 2016 [21] also provided annotated data for temporal information from clinical text. In japan, NII Testbeds and Community for Information access Research (NTCIR-MedNLP) provided the text for various topics of medical data exploitation [23], [24]. With available of all these resources, much effort has been done to develop different methods to extract temporal expressions from clinical narratives [17], [18]

To extract the temporal expressions, the well established HeidelTime has been used for generation of rules [18]. HeidelTime is a multilingual, cross-domain temporal expressions tagging system for text data. The system mainly developed based on TimeML corpus and uses the general text data. But it performed very poorly as the temporal expressions in clinic notes, that is very different from newswire text [65]. Such failure in using HeidelTime has suggested us to extend and modify it appropriately for temporal expressions in clinical text, as described later in subsection (3.3).

The work of Tang *et al.* developed a hybrid method combining Conditional Random Fields (CRFs) and Support Vector Machine (SVM) methods and used it to extract the temporal events and event features. In this hybrid method, NER tagger system [66] was extended, which was developed initially to extract the partial event features and temporal events [18]. Various combinations of features (Lexical and UMLS) were used in that NER tagger system for event extraction. Jindal *et al.* developed the pipeline approach to extract the events from clinical narratives. In the first stage, they extracted the time spans and subsequently they extracted the type of events. For the event extraction, Integer Quadratic Program (IQP) was implemented. They have adopted the publicly available time expression system HeidelTime to extract the temporal expressions [67]. There are few limitations in this pipeline approach. Firstly, the accuracy drop in finding the attributes subsequently causes performance decreases in event extraction. Secondly, developing the assumptions and rules are not easy and time consuming tasks. Lastly, in comparison to previous system, the feature extraction performance also diminished [18], [17].

Most of the existing temporal event extraction works were established with the available small number of annotated temporal corpus where as lot of unannotated clinical text are available compared to annotated clinical data. However, learning a stable and reliable model usually requires plenty of annotated data. Motivated by this fact, we proposed and developed a novel two-stage semi-supervised framework to automatically extract the temporal events from clinical text by exploiting unannotated text with gradually increase the number of annotated text in corpus and automatic annotation accuracy.

In case of temporal relation classification, various candidate generation and relation classification approach have been attempted in the existing studies. Chang *et al.* developed the hybrid system which consist of rule-based and Maximum Entropy (ME) based approach to generate the candidate pairs [68]. Finally, the authors have proposed a algorithm to integrate the candidate pairs from both the approach. Even though, they have developed candidate pair's generation process by using Maximum Entropy method, the processing steps are not very transparent and developing rules for each category by analyzing the clinical text is very ambiguous and time taking.

In contrast to the previous work, Tang *et al.* addressed the candidate generation problem with a formulated hypothesis based on the dependency parsing approach [18]. In this study, they have used different strategies for each category of temporal relations (relation between events and section times, intra-sentence and inter-sentence). The hypothesis and approach used to generate the candidate pairs for inter-sentence, intra-sentence temporal relation classifier are not effectively capturing some potential candidate pairs. Though we have various candidate pairs generation and classification methods developed, it is still a totally unresolved problem. This limitation suggested us to formulate a new hypothesis for generating candidate pairs.

2.5 Summary

In this chapter, we introduced the need for an annotated corpus of clinical narratives by comparing the newswire text. We then discussed characteristics of clinical narratives like doctor notes, nurse narratives, discharge summaries and discussed an annotation guidelines for annotating events, expressions, relations for temporal reasoning. Finally, we discussed the challenges and current status of temporal information extraction in clinical narratives. In the forthcoming chapters, we describe and establish methods that leverage the annotated data for extracting elements of temporal information.

Chapter 3

Temporal Expression Extraction from Clinical Narratives

This chapter first introduces the temporal expression in clinical narratives and its significance. Then, the requirement of annotated copora for temporal expression extraction. Also we discusses about similarities and difficulties of temporal expressions in general newswire and clinical narratives. Next part discusses about the current status of temporal event extraction in clinical narratives and our objective. Immediately we proposed a method for temporal expression extraction and established experimental evaluation with available annotated corpora. Finally we provided summary and contributions of this chapter and proposed method respectively.

Developing clinical decision support (CDS) applications in medicine is based on knowledge extracted from the clinical narratives in EMRs to discover and provide the better treatment for the individuals. The core of this CDS has the requirement of identifying and extracting the significant patient clinical details, that denotes the clinical state of events having some relationship with time as time performs the key role in patient course of a disease and treatment regimen. Therefore, extracting temporal extraction is fundamental elements in temporal information extraction and to represent the rest of the information in clinical narratives from patient medical history.

3.1 Temporal expression in clinical text

Temporal expressions are sequence of words or phrases that contain time information. The types of temporal expressions that includes date, time, duration, frequency and prepost expressions [43], [45]. In clinical narratives, temporal expressions are usually written as nouns ("yesterday", "Tuesday", "2016-08-08") , adjective ("two hourse ago", "preoperative") and adverbial ("hourly", "monthy", "recently") pharases. TIMEX3 tag is used to annotate a temporal expressions in clinical narratives like newswire text and it could be any one of the following:

1. DATE: denotes the exact calendar date. For example, Nov 13 Monday, 2016-08-08.
2. TIME: denotes a time of the day. For example, 8:00 AM, seven in the morning. *The patient's CT-scan is scheduled at 10:00 AM*
3. DURATION:Describes a time duration. For example, four months, next two days.
4. FREQUENCY: Describes a set of times. For example twice a day.
5. QUANTIFIER: describes like frequency expressions but it does conceive starting point and ending point. For example twice and three times. *The patient vomited "twice" before the treatment.*

6. PREPOSTEXP: The time span of this expressions is implicitly related to event. Therefore, this kind of expressions in clinical narratives are marked with special type as PREPOSTEXP. For example post-surgery.

The clinical narratives have the special time expression called prepost expressions. The adjectives Pre and Post actually denotes the temporal spans in one perspective that related to event. For example Pre-treatment denotes the until the point of treatment time, post-surgery denotes the any time after the surgery. Thus, these kind of expressions in clinical narratives is considered as special temporal expressions. The following is a example of annotated temporal expressions in clinical narratives obtained from i2b2 to train and evaluate models.

- The patient became agitated with both tent and Bi-PAP machine, although continued to saturate well. He was transferred to Medicine the morning of 2010-08-30 <TIMEX3= “2010-08-30” type= “DATE” value=“2010-08-30” >2010-08-30 </TIMEX3 >, for continued medical management post surgery.
- The patient has had similar pain intermittently for last year <TIMEX3=“last year” type=“DURATION” value=“p1y” >last year </TIMEX3 >.

3.2 Background of temporal expressions extraction in clinical narratives

Temporal expressions extraction has been well-studied in non-clinical settings and exploited for temporal reasoning and representation. For temporal expression extraction from newswire text, the best available systems like HeidelTime [69] and SUTime [70] are rule-based. Among these, HeidelTime is a multilingual cross-domain temporal tagger and was the best system for extraction of temporal expressions from newswire

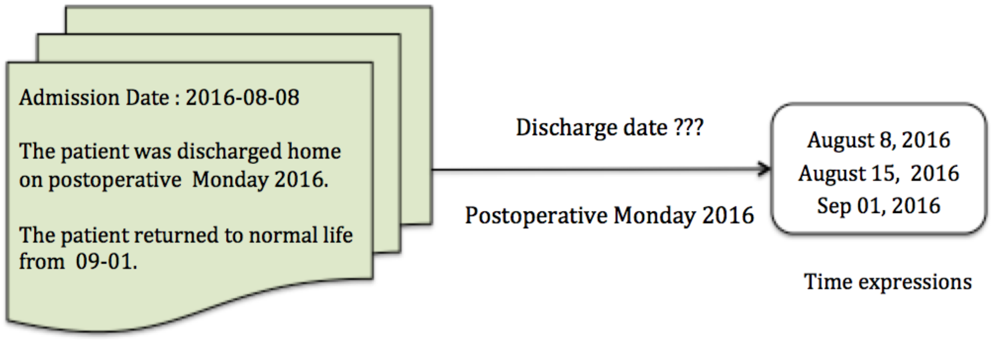

Figure 3.1: Objective of temporal expressions extraction in clinical narratives

text in the TempEval-2 challenge [71]. Reeves *et al.* extended temporal awareness and reasoning systems for question interpretation (TARSQI) toolkit to detect the temporal expressions in the clinical text [64]. But it performed very poorly as the temporal expressions in clinic notes were very different from those in the newswire domain [65]. Thus, *there is a requirement to develop a specific model for temporal expression extraction from clinical narratives.* Sun et al. provide comprehensive overview of systems which perform clinical temporal expressions extraction. 3.1 explained the goal of temporal expression in clinical narratives with example.

Though the aim of this task focuses on the automatic identification and extraction of time expressions from clinical narratives, the existing works requires manual intervention on developing rules. Thus, there is a room for making the task automatically extracting expressions from clinical narratives, which is objective of our work in this section.

3.3 Proposed method for temporal expressions extraction

Temporal expressions are sequence of words and phrases that expresses a point of time or spans on a timeline such as Date, Duration, Time and Frequency. Other temporal expression types in clinical text include Quantifiers and Prepostexp [65]. Table 3.1 shows the various types of temporal expressions in clinical text. According to i2b2 annotation guidelines, annotation of temporal expressions and determining the span for TIMEX3 is straightforward. They have mark the entire temporal expression. The prepositions before or after the temporal expression phrase are usually not included in the TIMEX3 span. However, automatically identifying the explicit and implicit time expressions in clinical narratives without manual intervention is a non-trial task. To accomplish our objective, *we proposed a hybrid method that combines conditional random fields and adopted the state-of-art from newswire domain (HeidelTime system)* for temporal expressions extraction from clinical narratives.

S.No	**Type**	**Example**
1	Time	12:00 PM, After 10 mins
2	Date	Morning of 2010-08-30, Today, Monday
3	Duration	Less than one hour, Next two days
4	Frequency	Twice per week, Monthly
5	Prepostexp	Postoperatively, post-surgery
6	Quantifier	Twice, three times

Table 3.1: Types of temporal expressions in clinical text

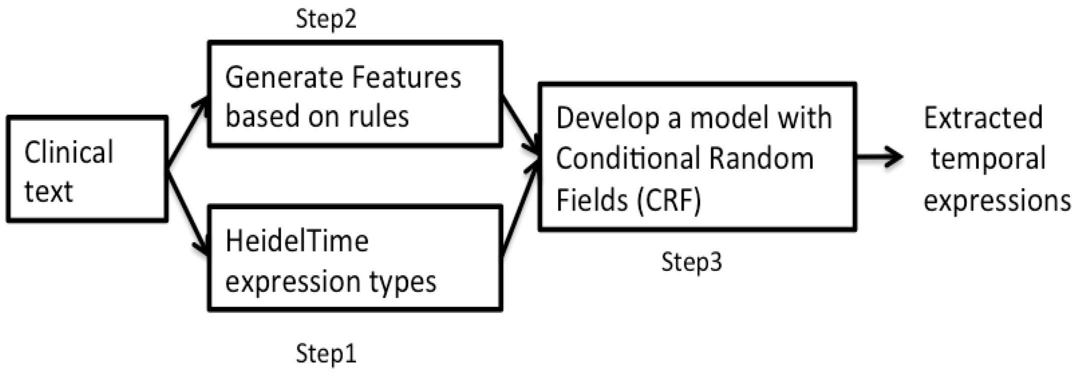

Figure 3.2: Architecture for temporal expression extraction

Heideltime[1] identifies the temporal expressions with three attributes ID, TYPE and VALUE respectively. HeidelTime system tags the frequency expressions as SET, thus we replaced the SET type with FREQ before generating the features using it for our proposed method. As previously mentioned, HeidelTime is not effectively extract all type of temporal expressions from clinical narratives. The following sentence is shows the inefficiency of HeidelTime for certain type temporal expressions in clinical narratives:

Mr. Donald delirium abated with return to baseline mental status *early on 09-01*. He was transferred to Medicine the *morning of 2010-08-30*, for continued medical management *post surgery*.

In the above sentence we have two date expressions such as *morning of 2010-08-30, early on 09-01* and one Prepostexp denotes *post surgery*. HeidelTime is insufficient to extract the prepostexp and date expression such as *early on 09-01*. Thus, we proposed a new framework with novel feature set and adaptation of HeidelTime features to extract the temporal expressions from clinical narratives.

[1]https://github.com/HeidelTime/heideltime/

Figure 3.2 shows our proposed framework for temporal expression extraction from clinical narratives. We used the output time expressions tags of the Java-based HeidelTime system, the best performing system in TempEval2 as a one of the input features in our proposed framework. Along with this features, we generated the lexical features to train the linear chain CRF for extracting the temporal expressions. The process of selecting the feature set is based on our experimental evaluation. We trained and evaluated our proposed model with various combinations of feature sets and chosen the best performing feature set among them. Table 3.2 shows the features used in our proposed framework.

There are three steps to achieve our proposed framework for temporal expressions extractions from raw clinical narratives. The detailed explanation of each steps are given below:

- Step 1: We used HeidelTime on raw clinical narratives to obtain the temporal expressions. HeidelTime system identifies and annotated the temporal expressions with three attributes: TYPE, VAL and MOD attributes. It tags the frequency expressions as SET (which is not present in clinical narratives), thus we replaced the SET type with FREQUENCY. Finally we converted all the tagged expressions as feature called HeidelTime tags.

- Step 2: We generated a lexical features set from raw clinical narratives by using GENIA Tagger for lemmatization, Part-of-Speech(POS), chunking and NLTK package from Python programming for unigram, bigram. Finally we used programming language to generate the "Is numeric" feature.

- Step 3: After the successful generation of features, we trained the linear-CRF model to detect the temporal expressions.

S.No	Features used
1	Word
2	lemma of word
3	POS tag of word
4	Chunking
5	Is numeric
6	next word
7	previous word
8	next two words
9	HeidelTime tags

Table 3.2: Features for temporal expressions extraction

The experimental evaluation of proposed method for event extraction has been well-studied in evaluation 3.4.3 and discussion section 3.4.4 in detail.

3.4 Experimental evaluation

The purpose of this experiment is to evaluate the efficacy of our developed model for temporal expression extraction as stated in the objective. To conduct the experiment we used the i2b2 annotated data set as a key resource.

3.4.1 Data sets

We obtained the preprocessed and annotated training and testing data from I2B2 temporal relation challenge organizers. This dataset was prepared based on the I2B2 guidelines (adopted from TimeML and THYME guidelines) for clinical text [45]. We used this annotated dataset for our experiments and evaluations. This annotated dataset consists of training data (180 annotated records from the I2B2 temporal relations shared task, which contains nearly 10,000 sentences) and testing data (120 annotated records from the I2B2 with ground-truth labels).

Our experiments consist 180 annotated records of training set and 120 annotated records of testing set. The training set contains 16468, 2368 and 33660 temporal events, expressions and relations, respectively. The testing set includes 13594, 1820 and 27530 temporal events, expressions and relations, respectively.

3.4.2 Baselines

In the first stage of the proposed framework, we developed the supervised CRF model with lexical features consisting of base form, part-of-speech tags, chunking and BIO-events (beginning of event, inside of event, outside of event) as a target label for temporal event extraction. To improve the accuracy, we incorporated the domain knowledge by exploiting UMLS through metathesaurus tool. Metamap CUI and semantic group features from UMLS helps to improve the performance for event extraction task as the result shown in Table 4.2. In the second stage of semi-supervised framework, we trained our semi-CRF model with the above mentioned features on the I2B2 training set and unannotated dataset. Finally, we evaluated our model with the I2B2 testing dataset for temporal event extraction.

3.4.3 Evaluation metrics and results

We evaluated our models with three measures, which are Precision, Recall and F-measure as given by:

$$Precision = \frac{|System.Output \cap Annotated.Corpus|}{|System.Output|} \tag{3.1}$$

$$Recall = \frac{|System.Output \cap Annotated.Corpus|}{|Annotated.Corpus|} \tag{3.2}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{3.3}$$

The results of our proposed approach for temporal expressions, events extraction and relation classification on I2B2 testing set is presented in Table 3.3, Table 4.2 and Table 5.1, respectively.

Expressions	**Precision**	**Recall**	**F-measure**
HeidelTime	77.62	79.80	78.69
HeidelTime + features	81.53	79.11	*79.95*

Table 3.3: Results from temporal expression extraction

3.4.4 Error analysis

The I2B2 dataset is annotated with 6 types of temporal expressions. We manually analyzed the temporal expressions extraction performance of our method very carefully. It revealed that the source of errors is from expression recognition. We listed our findings below.

- Unclear expressions: Many of the date expressions (like “the day prior to discharge” “this time”, “that time” and “the night prior”) were fairly significant and ambiguous. Identifying and recognizing of this type of expressions is quite difficult. In this situation our system is insufficient on such expressions.
- Difficulty in deciding the frequency expressions: As most of frequency expressions were written in abbreviation form such as t.i.d, b.i.d, qd and etc. We planned to use the bag-of-words model to detect this kind of expressions in our future work. Also, frequency expressions and duration expressions are nearly similar with the same kind

of words with slight variation. For example “two years” belongs to duration where as “every two hours” belongs to frequency.

3.5 Summary

In this chapter, we have presented a hybrid approach for temporal expressions extraction from clinical narratives. The novel framework was developed with the combination of two key components: adoption of state-of-art HeidelTime system from newswire domain and a novel feature set. We have converted the extracted time expressions as features for training model. Finally we have chosen the best feature set for temporal expression extraction in the final step of our proposed method. Based on the findings and experiments, we found that the proposed method is surprisingly simple and easily extract temporal expressions.

Beside the advantages, our method still has some drawbacks, which is discussed in error analysis section 3.4.4 in detail. In our future work, we plan to investigate the rules for latin expressions of time such as “tid” or “t.i.d”, etc.

Chapter 4

Temporal Event Extraction

This chapter first introduces the temporal events in clinical narrative and its significance. Then, the requirement of annotated copora and difficulty to obtain for temporal event extraction. Next part discusses about the current status of temporal event extraction in clinical narratives, advantages of semi-supervised learning and our objective. Also we discusses about similarities and difficulties of temporal events in general newswire and clinical narratives. Immediately we proposed a semi-supervised method for temporal event extraction and established experimental evaluation with available annotated corpora. Finally we provided summary and contributions of this chapter and proposed method respectively.

An EVENT is anything relevant to the clinical timeline (disease diagnosis and treatment details related to such timeline), (i.e.) anything that would appear on a detailed timeline of the patients care or life. We considered all the events in clinical text as temporal events relating to document creation time except temporally span-less events such as people and organization cannot be a temporal event. according to the i2b2 event annotation guideline, candidates for EVENTs include verb phrases, adjective phrases, noun phrases, and in some cases, even adverbs. Naturally, verb phrases that describe clinically relevant actions are considered EVENTs. For example, in the patient reports a headache, the verb reports refers to a clinically relevant action of the patients complaint, and hence is counted as an EVENT. Basically temporal events have annotated with 6 types of events in I2B2 clinical data. Standard types of temporal events are: Problem (disease and symptoms), Tests, Treatments, Evidential, Clinical departments and Occurrence.

Let us consider the following example sentences to understand more about temporal events in clinical narratives.

1. The patient had a CT scan, which showed fatty infiltration of her liver diffusely with a 1 cm cyst in the right lobe of the liver.

2. The Patient has no relief from antacids or H2 blockers.

- The Patient had a CT scan <EVENT= "a CT scan" type="TEST" modality="FACTUAL" polarity="POS" >a CT scan </EVENT >, which showed fatty infiltration of her liver diffusely <EVENT="fatty infiltration of her liver diffusely" type="PROBLEM" modality="FACTUAL" polarity="POS" >fatty infiltration of her liver diffusely </EVENT>with a 1 cm cyst in the right lobe of the liver <EVENT="a 1 cm cyst in the right lobe of the liver" type="PROBLEM" modality="FACTUAL" polarity="POS" >a 1 cm cyst in the right lobe of the liver </EVENT>.

- The patient has no relief <EVENT="relief" type="OCCURRENCE" modality="FACTUAL" polarity="NEG">relief </EVENT>from antacids <EVENT= "antacids" type="TREATMENT" modality="FACTUAL" polarity="POS" >antacids </EVENT>or H2 blockers <EVENT= "antacids" type="TREATMENT" modality="FACTUAL" polarity="POS" >H2 blockers </EVENT>.

[72]

4.1 Structured sequential labeling

4.1.1 Conditional random fields

Conditional Random Fields(CRF) [73] is a undirected probabilistic model which assign the label sequence to a given set of observed data sequence. It has all advantage of Maximum Entropy Markov Models (MEMMs) and other probabilistic models such as Hidden Markov Models (HMMs). The basic idea of CRF is that of calculating a conditional probability distribution over entire label sequences given a particular observation sequence, rather than a joint distribution over both label and observation sequences [73]. **X** is a random variable over data sequence to be labeled, and
Yis a random variable over corresponding label sequences.

Conditional probability **P(label sequence y | observation sequence x)** rather than joint probability P(y, x)

- Specify the probability of possible label sequences given an observation sequence

Conditional probability of a label sequence y given an observation sequence x to be written as

$$\mathrm{p(y|x,}\lambda) = \frac{1}{Z(x)} \exp(\Sigma\ \lambda_j F_j\mathrm{(y,x)})$$

Where Z(x) is a normalization term and

$$\lambda_j F_j(y,x) = \begin{cases} 1 & if condition = true \\ 0 & otherwise \end{cases}$$

Generally we say that model of conditional distribution is:

$$\mathrm{p(y|x)}$$

To predict the label for unknown text is

$$y^* = \underset{x}{\operatorname{argmax}}\ \text{p(y|x)}$$

Here we provided the example of sequential label prediction for Part-of-Speech Tagging. Consider the following text for Part-of-Speech(POS) Tagging:

There have been many earthquakes in Tokyo.

Word	There	have	been	many	earthquakes	in	Tokyo	.
X $\rightarrow$	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8

Word	There	have	been	many	earthquakes	in	Tokyo	.
Y (POS tag) $\rightarrow$	EX	VBP	VBN	JJ	NNS	IN	NNP	.

In the above example X might range over natural language sentences and Y range over part-of-speech tagging of those sentence ,with y the set of possible part-of-speech(POS) tags. Sutton and McCallum provided the overview of linear-chain CRFs, general CRFs, varies of CRF types and applications of CRFs[74]. Figure 4.1 shows the simple graphical representation of linear-chain CRF and general CRF models.

4.1.2 Conditional random fields on text processing

Text processing exploited the probabilistic model for various applications. Generally text processing refers to automatically understanding of electric text which denotes Natural Language Processing(NLP). Most of NLP tasks exploited the probabilistic model, especially Hidden Markov Models(HMM), Maximum Entropy Markov Models(MEMM), Markov Random fields such as linear chain CRFs and Generalized CRF's for text processing applications. NLP subfields includes, Information Retrieval(IR), Information Extraction(IE), Discourse analysis, dependency parsing and etc.

The basic task of Information Extraction problem has been treated as a sequence labeling problem [73]. Specifically Conditional Random Fields(CRFs) established for many

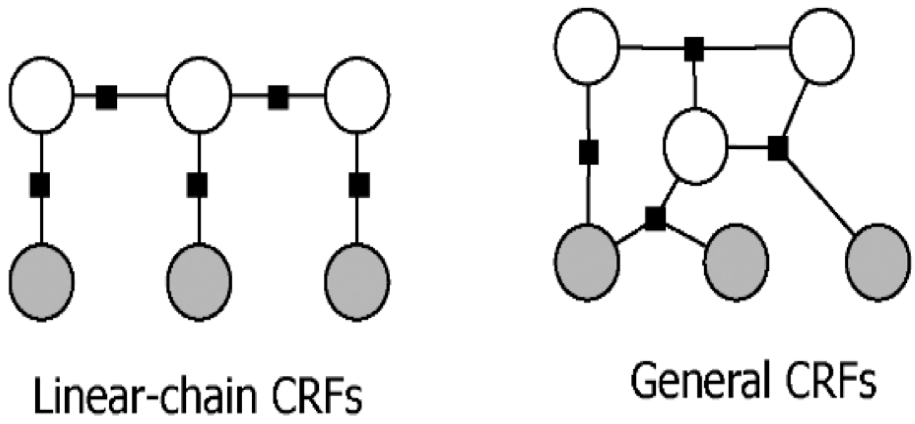

Figure 4.1: Diagram of Conditional Random Fields: Linear-chain CRF and General CRF

subfields of Information Extraction [75], such as Part-of-Speech Tagging [76] [77], Named Entity Recognition(NER) [78],[79], Co-reference extraction and relationship extraction [80].

4.1.3 Conditional random fields for temporal information extraction

Temporal information extraction plays very significant role in Natural Language Processing. In the literature of general text, temporal expression and event extraction [81], [82], [83] considered sequence labeling problem and temporal relation extraction has been treated as a classification problem [84], [3]. In TempEval 2007, Verhagen *et al.,* proposed many machine learning approaches for temporal relation identification and classification. Later [85] temproal relation identification has been treated as pair-wise classification problem, and exploited CRF to identify the relationship.

In clinical text, conditional random fields has been used to extract temporal event and expression [86], [87], [55], [88] and temporal relation extraction [89] . Clinical text is a special kind of text as mentioned earlier. From the literature review of temporal information extraction from general and clinical text, we found that, conditional random fields (CRF) is outperforming than other machine learning and rule-based methods.

4.2 Temporal event extraction in clinical narratives

The most techniques for temporal information extraction on clinical text are mainly based on available techniques used in the general newswire text domain. The rule-based methods deeply analyzed the nature of the data and developed the hand-coded rule to extract the event [87], which demands domain knowledge besides much of manual effort and time. In the machine learning-based approach, the feature set is determined through NLP data structures, part-of-speech, n-gram, or external resources [55], [87] before applying probabilistic models such as Hidden Markov Models(HMM), Conditional Random Fields(CRF) [18]. The best performing system Xu *et al.* from I2B2 shared task Sun *et al.*, using the hybird approach with hand-coded rules, CRFs and SVMs by exploiting the various feature sets [90].

On the other hand, Jindal *et al.* developed the pipeline approach to extract the events from clinical text. In this approach, they recognized the attributes of the events in first stage, then they implemented Integer Quadratic Program (IQP) for the event extraction. For extracting the temporal expressions, they have adopted the publicly available time expression system HeidelTime [67]. However, this approach have the limitation with accuracy. The event accuracy of this method is decreased in comparison to existing approaches [18].

From the above literature review, we discovered that all the existing temporal event extraction work were established with the available small number of annotated temporal corpus. However, a lot of rich unannotated clinical text are available compared to annotated clinical data. Moreover, the temporal annotation task is time-consuming, often requires the domain knowledge and considerable manpower to annotate the corpus man-

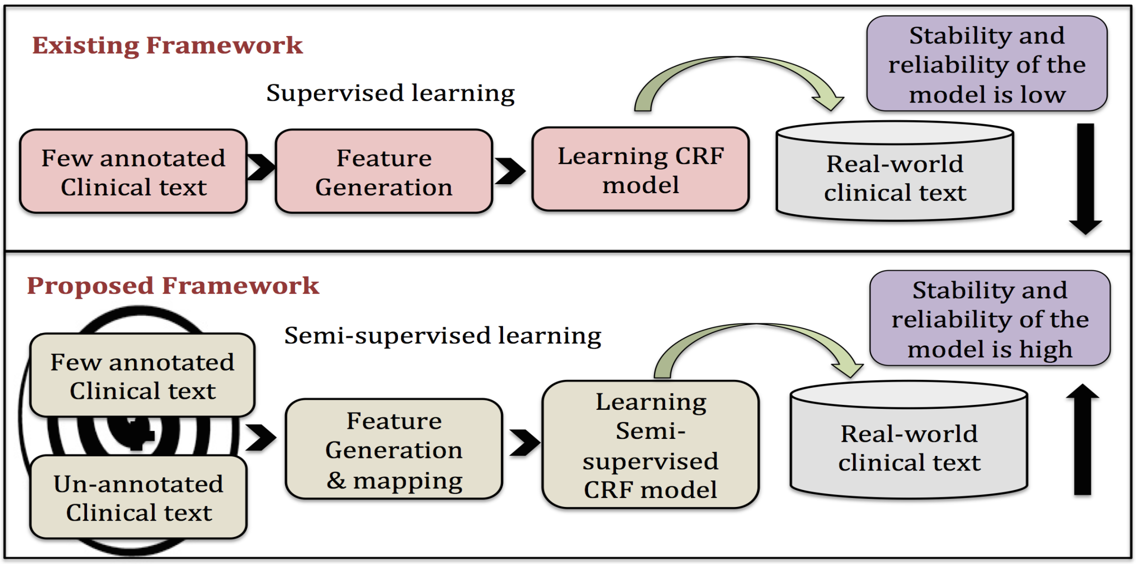

Figure 4.2: Problem space and overview

ually. Also, unannotated testing corpus is not likely to be same as with their training or testing dataset. Besides that, from the literature review, we discovered unannotated data have significantly contributed to enhance the model when we combine with annotated data [91]. Therefore, we proposed and developed a semi-supervised framework to automatically to detect the temporal events from clinical text by exploiting unannotated text with gradually increasing of the number of annotated text in the corpus, which increases the automatic annotation accuracy as shows in figure 4.2.

4.3 Semi-supervised learning

Semi-supervised learning lies somewhere between supervised and unsupervised learning. Supervised learning methods have been successfully implemented for many text processing tasks such as sequence labeling, information extraction, but they demands annotated dataset which is expensive, time consuming and requires lot of manual effort to obtain, where as unannotated data available for less cost and easy to obtain. Making use of unlabeled data helps to improve the performance of NLP tasks. Hence we investigate semi-supervised approach to combine both labeled and unlabeled data. Semi-supervised learning will be most useful when there are huge amount of unlabeled data than labeled data. Semi-supervised learning technique have already been successfully exploiting many NLP tasks including information extraction [92] [93], word sense disambiguation [94]. The data set of standard semi-supervised learning can be divided into two parts: the points $X_l = (x_1, x_2, \ldots, x_l)$, for which labels $Y_l = (y_1, \ldots, y_l)$ are provided, and the points $X_u := (x_l + 1, \ldots, x_l + u)$, the labels of which are not known. The earliest idea about using unlabeled data is self-learning, which is also known as self-training, self-labeling, or decision-directed learning. The interest in semi-supervised learning has been increased from 1990s, mostly due to applications in natural language problems and text classification [94]. Since then, there are many works has been done in Natural language processing and text mining using semi-supervised approach. Some popular representative models of semi-supervised learning is:

- Self-learning (Self-training or bootstrapping) and Co-training
- Generative probabilistic models
- Graph-based semi supervised method

Generally semi-supervised methods forms strong model assumptions (Smoothness , cluster and low density assumption) to exploit unlabeled data. Choose the best assumption that fit to the nature of the problem.

4.3.1 Self-learning (self-training or bootstrapping)

This is the simplest and most common semi-supervised learning method which is also called self-learning or bootstrapping method. Mostly supervised model trained with labeled data, subsequently the model applied n unlabeled data to accomplish the semi-supervised objective. Then typically most confident scores of the labeled sequence will be added to the labeled set and the model is re-trained until its converge [91]. This method has been successfully applied NLP tasks such as word sense disambiguation and temporal expression extraction [95] as mentioned earlier.

4.3.2 Semi-supervised CRF for information extraction

Jiao *et al*, successfully developed the semi-supervised CRF through *extending the minimum entropy regularization.* Semi-supervised approach derives benefit from the usage of entropy regularization on unlabeled data, which *combines the likelihood of CRF from labeled training data and conditional entropy on unlabeled data* [93]. Exploited the developed semi-supervised CRF for information extraction (To detect protein-gene mentions) from biomedical text and achieved better results compared to supervised CRF models. This work compared to self-training algorithm developed by Yarowsky 1995 [94], proposed semi-supervised CRF provides better performance. Motivation of our work observed from the work of Granvalet *et al.,* semi supervised learning by entropy minimization [96]. It gives *the assumption that when the classes has small overlap, unlabeled data provides promising benefit.* Overlap between the classes dealt with kullback-Leibler divergence.

Despite of having some works focused on NER and event detection from biomedical text [97], [93] using semi-supervised CRF where as there is no work has been done for temporal information extraction clinical text.

Although many researchers implemented semi-supervised learning for information extraction such as Named Entity recognition [98],[97] and word sense disambiguation [94] [99] only limited number of works focused on temporal information processing such as temporal expression extraction [95] and temporal relation extraction[85], [100], [101] from newswire text.

4.4 Semi-supervised framework for temporal event extraction

Probabilistic models have been used for information extraction for past several years. Conditional random fields is a undirected probabilistic model which assign the label sequence to a given set of observed data sequences [73]. It is often used for labeling or parsing of sequential data, such as natural language text or biological sequences. From considering the above advantages of CRF and literature review from general and clinical text, we found that, CRFs is outperforming in temporal information extraction than other machine learning and rule-based methods. Therefore we focus on semi-supervised CRF to extract events from both annotated and unannotated clinical text. Semi-supervised CRF has been investigated for biomedical text in [93].

Events in clinical text refers to medical concepts such as symptoms occurred, disease/disorder names diagnosed, medications prescribed by doctors and treatment procedures followed by patients. Therefore event extraction task in clinical text demands the domain knowledge as previously mentioned. Due to the high cost of annotation, time, manual effort and requirement of domain knowledge, annotation corpus for temporal information in clinical texts are limited to a few hundred patient records. All the existing methods used supervised CRF method to extract the temporal events from clinical text with such limited size of annotated data [18], [17]. We proposed a novel semi-supervised framework for event extraction from clinical text. Our proposed semi-supervised framework of temporal event extraction has been divided into two stages shown in figure 4.3. In the first stage, we developed a supervised Conditional Random Fields[1] model to recognize the temporal events with various features sets, which is described in section 4.4.1 and 4.4.2 in detail. Table 4.1 shows the features used in our proposed framework.

In stage II we utilize the rich quantity of unannotated data to develop the semi-supervised CRF from stage I and increase number of records in annotated corpora step by step (spi-

[1] https://sourceforge.net/projects/crfpp/

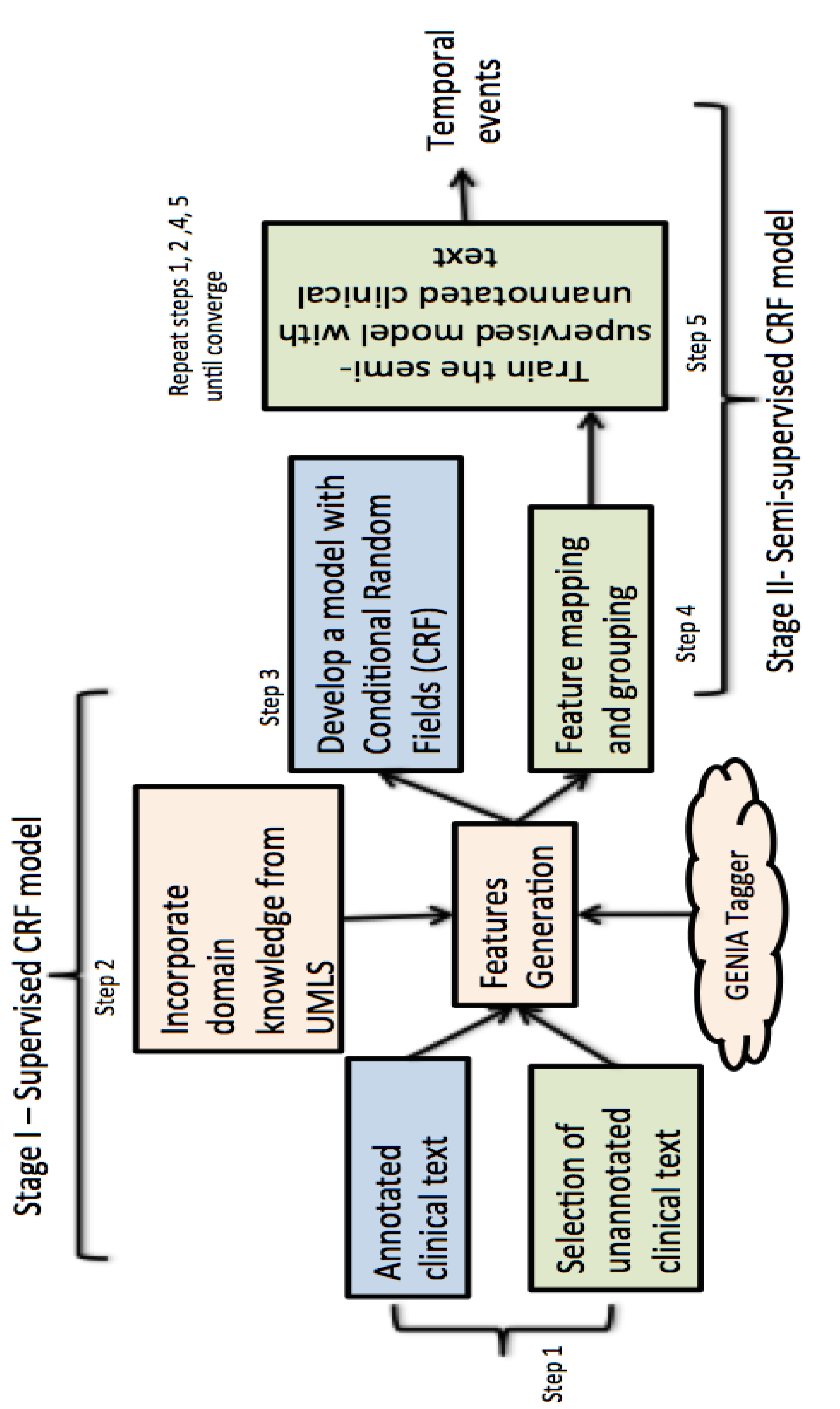

Figure 4.3: Overview of the proposed architecture for temporal event extraction.

S.No	Features used
1	Word
2	lemma of word
3	POS tag of word
4	Chunking
5	previous word
6	next word
7	POS of next word
8	POS of Previous word
9	bigram
10	trigram
11	Concept unique identifier from Metamap
12	Semantic groups from Metamap

Table 4.1: Features for temporal event extraction

rally) to subsequently improve the event extraction accuracy. To the best of our knowledge, this proposed semi-supervised framework to temporal event processing from clinical text is a novel trial.

The second stage of proposed semi-supervised framework consists of two steps: *unannotated data selection* for selecting the high quality data from plentiful unannotated data, and *feature generation and mapping* for extending the labels to unannotated training data. After the feature generation and mapping, we trained the semi-supervised CRF (step 5) with Mallet tool[2].

[2]http://mallet.cs.umass.edu/semi-sup-fst.php

4.4.1 Selection of unannotated training data

The step 1 of the proposed semi-supervised framework requires a training data from an unannotated dataset. As previously mentioned, we have an abundant amount of unannotated data. However, we cannot assume that all the patient records in unannotated data are equally distributed among all the disease groups (according to international classification of diseases from WHO) and contains rich and significant temporal information. In this situation, to maintain the fairness of data selection in the empirical experiment, we leveraged the advantage of using a clustering algorithm for the unannotated training data selection instead of selecting the records randomly. This method helps us to isolate rich and sparse training data from massive unannotated dataset. To select the unannotated training records, first we should group all the unannotated patient records according to disease similarities by incorporating the domain knowledge.

There are various document clustering methods has been proposed in the existing researches [102]. The documents grouping or clustering using Latent Dirichlet Allocation (LDA) is based on latent topics. These topics are constructed over the probability distribution of words. However, in this method, we should initialize the number of latent topics in advance and the topic constructs over the distribution of words [103]. Based on the topics learned, the document clustering will be carried out with similarity measures. Here in our work, deciding the number of topic in advance by incorporating domain-knowledge (International Classification of Diseases (ICD) codes) is more challenging and difficult in clinical narratives. Therefore documents clustering using LDA leads to complex and time-consuming task to select unannotated training data for semi-supervised approach.

On the other hand, K-means clustering is used widely for text document grouping and suitable for larger datasets with less number of iterations [102]. Therefore we chose the basic and prevailing K-means clustering algorithm to group the unannotated training records since we are grouping the patient records based on disease groups from ICD codes.

To accomplish the clustering, we exploited the K-means clustering algorithm [104] with TF-IDF. After clustering, we obtained K clusters with similar group of patient records and we selected N training records from each cluster K based on our heuristic analysis. The selected training records from each cluster contain sparse and rich temporal information. The selected unannotated data is included with the annotated data for next step in feature generation and mapping. Subsequently, the influence of unannotated data towards the trained model would be high and the model achieves high accuracy and stability.

4.4.2 Feature generation and mapping

In step 2 and 4 of the semi-supervised framework, we have created a list of feature groups to train and evaluate the performance of the model.

Feature generation To generate the features, various external tools have been exploited. These feature sets are selected based on the significance of each feature. The first group of features consist of lexical features, such as base form, surrounding words, POS tags, chunking and language model. The lemmatization is the only one feature belongs to second group. In the third group, feature set is generated from UMLS metathesaurus [3], consist of semantic group and Concept Unique Identifier (CUI) of words.

- *Lexical and syntactic features:* We have used GENIA Tagger[4] to generate the lexical features such as lemmatization, Part-of-Speech (POS) tags, chunking the phrases. To define the bigram, trigram, near-by words and their POS tags we have exploited Python programming language with NLTK package[5]. The n-gram language model plays the key role as a feature in the label detection.

- *Metamap features:* As we mentioned earlier, the clinical text contains the rich quantity of medical terms and it requires the external medical knowledge system to detect the medical concepts. Thus, we have used UMLS system through metamap

[3]https://metamap.nlm.nih.gov/
[4]http://www.nactem.ac.uk/GENIA/tagger/
[5]http://www.nltk.org/api/nltk.tag.html

tool to get of crucial features such as semantic group and Concept Unique Identifier (CUI) of words from clinical texts.

Feature mapping and grouping After generating the features from the training dataset, to develop an effective training procedure, we first need to extend the labels for the unannotated dataset. Unlike the general text, the clinical text has the advantage to group the words based on semantic group from UMLS. Therefore we grouped the features and words based on the semantic groups from UMLS. We selected 35 predominantly contributing groups from 133 semantic groups in UMLS. The experimental evaluation of proposed method for event extraction is well studied in evaluation section 3.4.3 and discussion section 4.5.4 in detail. The following figure 4.4 shows the example of feature mapping based on semantic group from UMLS.

4.5 Experimental evaluation

The purpose of this experiment is to evaluate the efficacy of our developed semi-supervised framework for temporal event extraction from clinical narratives as stated in the objective. To conduct the experiment we used the i2b2 annotated data set and supervised model (for selecting feature set) as a key resources.

4.5.1 Data sets

We obtained the preprocessed and annotated training and testing data from I2B2 temporal relation challenge organizers. This dataset was prepared based on the I2B2 guidelines (adopted from TimeML and THYME guidelines) for clinical text [45]. We used this annotated dataset for our experiments and evaluations. This annotated dataset consists of training data (180 annotated records from the I2B2 temporal relations shared task, which contains nearly 10,000 sentences) and testing data (120 annotated records from the I2B2 with ground-truth labels). To develop a semi-supervised framework for temporal event ex-

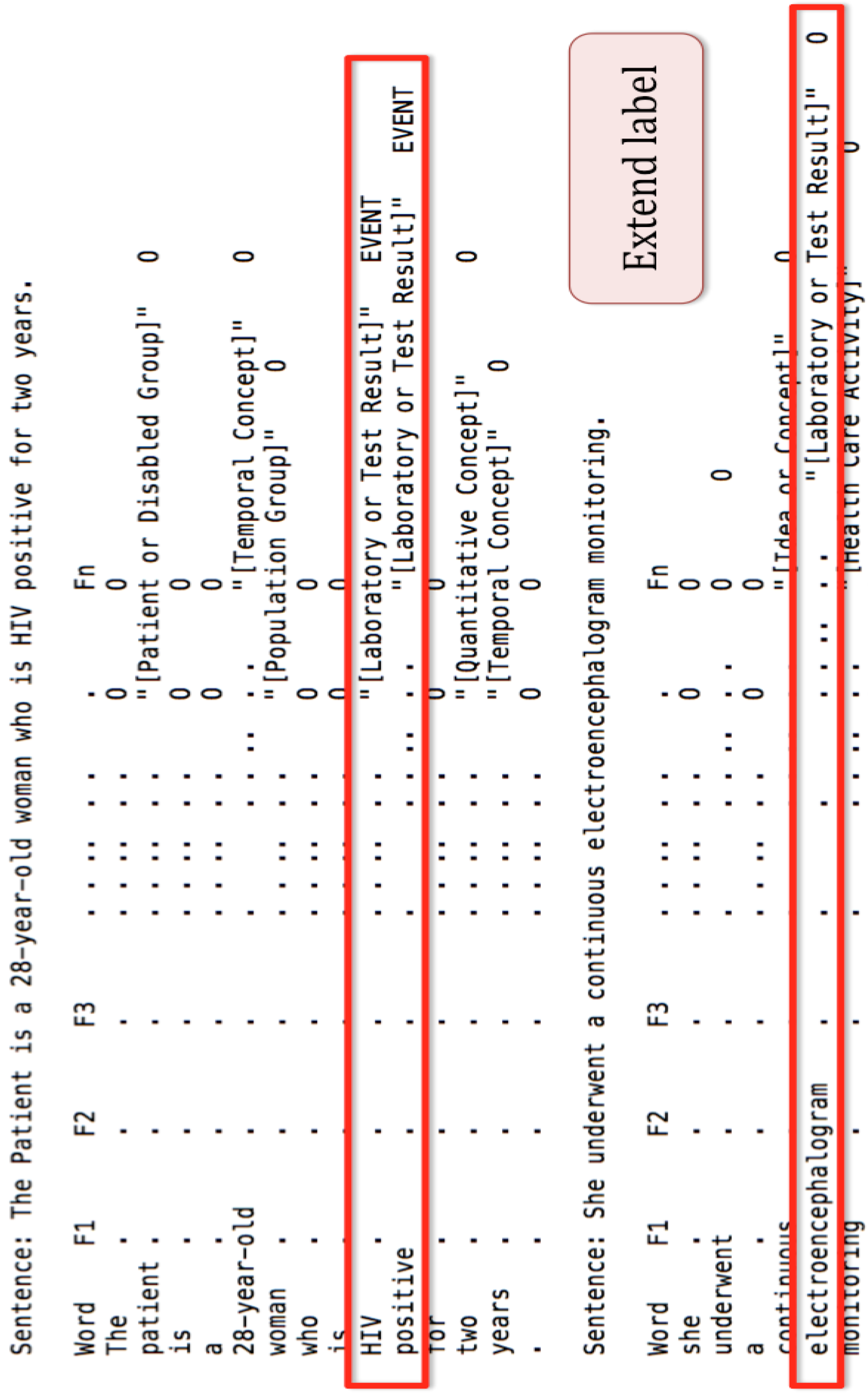

Figure 4.4: Example of feature mapping based on semantic group from UMLS.

traction, we have added unannotated training data (900 unannotated records from I2B2 medication and relations shared task, which contains more than 85,000 sentences) along with the above annotated dataset.

Our experiments consist 180 annotated records of training set and 120 annotated records of testing set. The training set contains 16468, 2368 and 33660 temporal events, expressions and relations, respectively. The testing set includes 13594, 1820 and 27530 temporal events, expressions and relations, respectively. For unannotated dataset, we used 2010-I2B2 relation challenge and 2009-medication extraction data along with annotated data for temporal event extraction.

4.5.2 Experimental design

In the first stage of the proposed framework, we developed the supervised CRF model with lexical features consisting of base form, part-of-speech tags, chunking and BIO-events (beginning of event, inside of event, outside of event) as a target label for temporal event extraction which is considered baseline model. To improve the accuracy, we incorporated the domain knowledge by exploiting UMLS through metathesaurus tool. Metamap CUI and semantic group features from UMLS helps to improve the performance for event extraction task as the result shown in Table 4.2. In the second stage of semi-supervised framework, we trained our semi-CRF model with the above mentioned features on the I2B2 training set and unannotated dataset. Finally, we evaluated our model with the I2B2 testing dataset for temporal event extraction.

Evaluation metrics

We evaluated our developed models with three measures, which are Precision, Recall and F-measure as given by:

$$Precision = \frac{|System.Output \cap Annotated.Corpus|}{|System.Output|} \tag{4.1}$$

$$Recall = \frac{|System.Output \cap Annotated.Corpus|}{|Annotated.Corpus|} \tag{4.2}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{4.3}$$

4.5.3 Experiment results

The results of our proposed approach for temporal events extraction on I2B2 testing set is presented below in Table 4.2.

Events	**Precision**	**Recall**	**F-measure**
Baseline	69.7	78.1	73.66
Stage I CRF (run1)	85.27	72.53	78.39
Stage I CRF (run2)	85.52	77.56	81.34
Stage II Semi-CRF (run3)	86.41	82.25	***84.21***
Stage II Semi-CRF (run4)	91.56	88.14	***89.76***
Stage II Semi-CRF (run5)	91.68	88.36	***89.98***
Top performing system from I2B2 [18]	93.74	86.79	90.13

Table 4.2: Evaluation results of temporal event extraction.

The proposed method of temporal expression extraction with added new feature set achieved F-measure of 79.95, which is relatively better than HeidelTime system. The precision increased considerably, whereas recall did not have much difference. Therefore, the overall performance has slight improvement compared to HeidelTime system as shown in Table 3.3.

To develop the semi-supervised method, we applied the K-means clustering algorithm on the unannotated dataset to select the rich data. In this process, we obtained 30 clusters and selected 30 records per each clusters. Finally we obtained 900 sparse unannotated training records with rich temporal information. From this 900 records, we used 300 records initially for training for the event extraction model. Consequently, the size of annotated data is increased from unannotated data, whereas unannotated data is added spirally from the selected unannotated records. Finally we checked the stability of our developed event extraction model with the testing data.Table 4.3 summarizes the different combination of feature sets in experimental settings and evaluated systems for temporal event extraction with number of training records.

	Lexical	**Syntactic**	**UMLS**	**Data selection**	**No of records**
Baseline	•				180
CRF (run1)	•	•			180
CRF (run2)	•	•	•		180
SemiCRF(run3)	•	•	•	Random selection	300
SemiCRF(run4)	•	•	•	K-means clustering	500
SemiCRF(run5)	•	•	•	K-means clustering	900

Table 4.3: Experimental settings for temporal event extraction and evaluation.

The proposed semi-supervised framework for temporal event extraction with K-means data selection method achieved F-measure of 89.76. This result is almost closer to the top performing system based on hybrid approach [18], which is shown in Table 4.2. Due to the unannotated dataset addition to the annotated data in the semi-supervised CRF model, it shows a significant improvement in Precision, Recall and F-measure. Apart from the accuracy and performance from the testing dataset, we evaluated the stability and reliability of the developed model by applying on the testing dataset as shown in figure 4.5. Our approach is flexible regarding addition of unannotated training data, and the corpus size can be easily increased. Subsequently, reducing time and effort increases the accuracy and stability of the system. Finally, we accomplished our objective in automatically annotating the unannotated clinical narratives with temporal events, which reduces time, cost and manual effort.

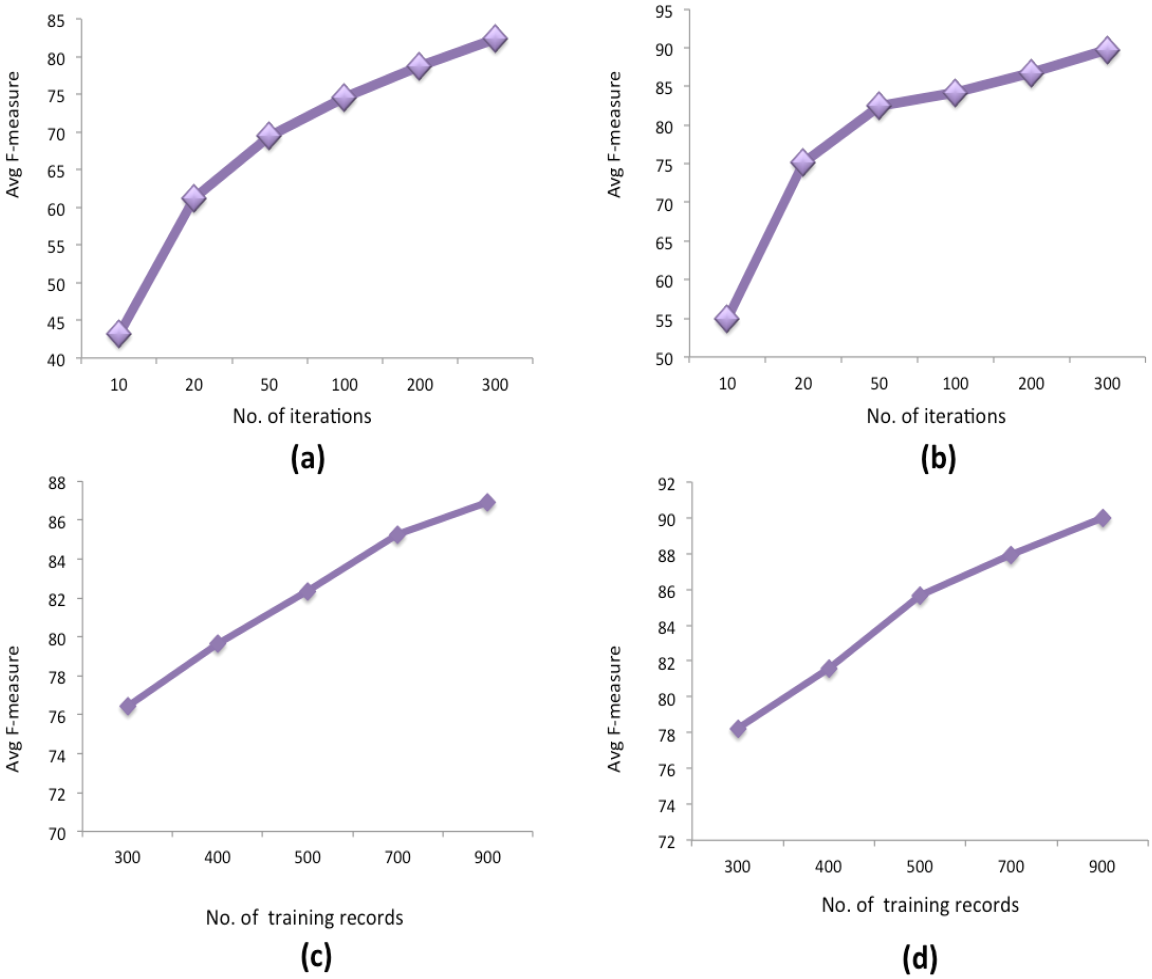

Figure 4.5: Performance of temporal event extraction from proposed semi-supervised CRF model (a) Random selection and (b) K-means clustering.

4.5.4 Error analysis

- Lack of annotated dataset (unseen events not always coinciding with annotated text, unknown abbreviations (SBE, URI, TPN, CK, etc.)) is a leading reason for less performance in event detection. We focus on to develop stage II of semi-supervised framework with utilization of unlabeled dataset to address this problem.

- Our system identified some of the events partially as some symptoms and disease name are very long. For example, "connected up with social services", "infarct of the cerebellar hemispheres bilaterally , the right occipital lobe , the right thalamus and bilateral pons". Our method could identify this type events partially, which subsequently affected the performance.

The size of the annotated dataset is small and it could not cover all possible events compared to vast amount of unannotated clinical narratives available. Also, some of the events are very long and it is difficult to identify them effectively by using CRF.

The above limitations we was overcome by processing unannotated data in Stage II of our method. This proposed approach helps to significantly improve the performance. In stage II, we utilized the K-means clustering algorithm for data selection and introduced feature mapping to extend the labels for unannotated training data.

We have carried out our experiment with an increasing number of training records and iterations. The performances of existing supervised methods is normally affected by new testing data as the number of annotated data is limited, while our proposed method allows gradual addition of unannotated data which helps to identify new events effectively. Therefore the performance of event extraction is steadily increasing in each step. Moreover, the precision is increased slightly and the recall is increased significantly due to the contribution of unannotated clinical text as shown in Table 4.2. From this we concluded that unannotated data contributed to increase the accuracy of temporal event detection as expected. We planned to use CRFs with word embeddings as in [105] to extract very long events in future.

4.6 Summary

In this chapter, we have presented semi-supervised framework for the temporal event extraction in clinical narratives. The novel framework was developed for the semi-supervised approach to extract temporal events from the rich unannotated text by extending annotated corpora. The proposed method has two stages to achieve the semi-supervised frame work. In stage I, we developed CRFs model on I2B2 annotated data with various features set. From this we have chosen the best feature set for temporal event extraction for next stage. In the second stage, besides using annotated data, we utilized the simple K-means clustering algorithm to select the rich and sparse data from abundant unannotated data.

We reported the achieved results for event extraction from stage I and stage II. We also witnessed the strong and steady improvement in event extraction accuracy in semi-supervised framework. Compared to the baseline system, performance and stability of the semi-supervised method for event extraction is significantly increased due to utilization of unannotated data, our novel assumption, incorporation of concept unique identifier and semantic group features from UMLS. In stage II, we extended the unannotated data by feature mapping and trained the semi-supervised CRF. In this work we proved that the influence of unannotated clinical text helps to improve the accuracy of temporal event extraction significantly. In our future work, we plan to investigate the expansion of lexical features set by using medical dictionary and semantic ontologies for event extraction.

Chapter 5

Temporal Relation Detection

This chapter first introduces the notion of time in clinical natural language text and how it plays the major role in ordering the events. Next part discusses about the current status of temporal relation classification in clinical narratives and our objective. Our objective consists of two steps: candidate pair's generation and temporal relation classification. Immediately we proposed a method for candidate pair's generation and temporal relation classification. Also we established the experimental evaluation with available annotated corpora and conducted the detailed analysis on obtained results. Finally, we provided summary and contributions of this chapter and future research works respectively.

With increasing importance and exploitation of unstructured clinical narratives from electronic medical records(EMRs), one of the key thing is to order the clinical events with clinical time-line. Basically, the clinical observation of a patient details are used to be noted chronologically in nurse narrative, doctor notes and discharge summary. Learning temporal relations from clinical narratives key for temporal reasoning and representation. However understanding and identifying this implicit temporal order of the events and/or expressions is very ambiguous and crucial task in clinical domain. Let us consider the following sentence for understanding the importance of learning temproal relations for temporal reasoning:

"The patient was *discharged* home on *postoperative day six* with *stable condition* and *excellent pancreas graft function.*"

This sentence apparently discusses a change in a patient's health status and the response that was taken by the hospital. In order to represent this response chronologically as it happened, several steps are necessary. First and foremost extraction of all the temporal expressions and it should grounded on a timeline. Second, all the available medical events from the sentence must be noted. Finally, the relationship between the extracted events and/or temporal expressions should identified in order to generate the timeline on the above sentence.

5.1 Temporal relations in clinical narratives

The definition of a temporal relationship is the timing between a factor and an outcome which can be used to assign causality to a relationship. In the medical context, a temporal relationship between a triggering event and an illness is medically acceptable if the time frame between the event and illness would reasonably support the conclusion that the event caused the illness. These relations exist between events and/or expressions for the reason to allow us to arrange and order the all the available events and/or expressions accurately on a chronological order as it happened in patient medical history. In the newswire domain, tense and aspect are often used as a key features to temporally order events and/or expressions. But, most the sentences in the clinical narratives are written in the past tense and many clinical events are noun phrases. Therefore, tense becomes a less informative and important feature in clinical narratives.

To annotate the temporal relations in clinical narratives, various interval logic has been proposed in past temporal reasoning history. Allen's proposed the interval-based temporal logic for defining the relationship between two events [106]. In contrast to the interval temporal reasoning, a point of time reasoning has been proposed by Marc Vilain and Henry Kautz [107]. Among these representations, Allens interval logic is more suitable for representing the events and expressions from clinical text with clinical timeline [11],[43]. In Allens interval logic, there are 13 kinds of relations is shown in figure 5.1. From this relation types, at least one typical type of relationship should exist between a pair of event or expressions.

According to the I2B2 annotation guidelines and annotated corpora, In this work, we have used three relationships (AFTER, BEFORE and OVERLAP) between pairs of event or expressions among 13 types of relations.

TLink tag is used to annotate temporal relations between events and/ or expressions. The temporal relations are annotated based on Allens interval algebra as explained in figure 5.1 and used for our experiments. For example: the annotation of temporal relation on following sentence "The patient was admitted with type 1 diabetes mellitus who is four months post cadaveric kidney transplantation and now good graft function" would be:

<EVENT1=type 1 diabetes mellitus | EVENT2=postcadaveric kidney transplantation | Tlink-type=OVERLAP >

<EVENT2=postcadaveric kidney transplantation | TIMEX3=four months | Tlink-type=AFTER >

<EVENT2=postcadaveric kidney transplantation | EVENT3=good graft function | Tlink-type=OVERLAP >

In this example, EVENT2 is overlapping with EVENT1 and EVENT3, at the same time AFTER the Temporal expressions TIMEX3.

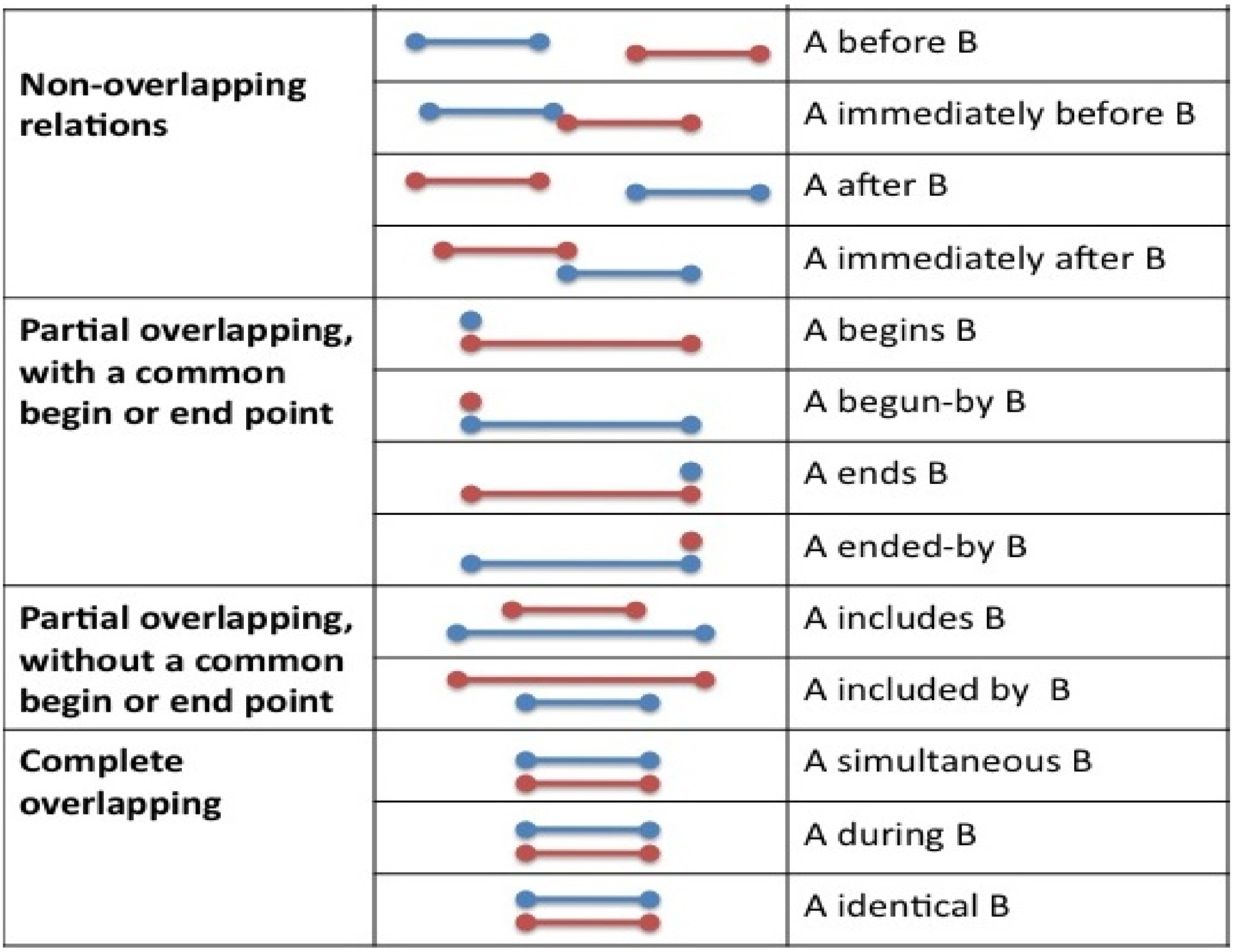

Figure 5.1: Allen's temporal relations

5.2 Background of temporal relation identification and classification

Though the temporal relation identification and classification is well-studied in non-clinical settings, classify the temporal relations in clinical text is a key significant, yet nontrivial task, in order to arrange the clinical events according to time stamp [9]. However temporal relation classification is not possible without generating candidate pairs from extracted events from raw clinical text. Therefore in this section, we would like to discuss about the candidate pairs generation from raw clinical narratives.

Various candidate generation and temporal relation classification approach have been attempted the existing studies. The best performing system for temporal relation classification is achieved 69.43 in I2B2 temporal relations challenge [18]. They implemented CRF++, LIBLINEAR of SVMs for temporal relation classification. The hybrid method classified the relations successfully and performed top of the submitted results in I2B2 temporal relations challenge. On the other hand [7] used Naive Bayesian classifier to classify temporal relations in TimeBank Corpus. Though we have various candidate pairs generation Chang *et.*al developed the hybrid system which consist of rule-based and Maximum Entropy (ME) based approach to generate the candidate pairs [68]. Finally proposed the algorithm to integrate the candidate pairs from both the approach. Even though they have developed candidate generation algorithm by using using machine learning algorithm, the processing steps are not very transparent and developing rules for each category by analyzing the clinical narratives is very ambiguous and time taking.

In contrast to the previous work, Tang *et.*al addressed the candidate generation problem with formulated hypothesis based on dependency parsing approach [18]. In this study, they have used different strategies for each category of temporal relations (relation between events and section times, intra-sentence and inter-sentence). The hypothesis and approach used to generate the candidate pairs for inter-sentence temporal relation classifier is not effectively capturing some potential candidate pairs. Therefore we proposed new hypothesis for this section to generate the candidate pairs. We implemented both LIBLINEAR SVM classification and Naive Bayesian classifier to classify temporal relations from clinical narratives using gold annotated I2B2 dataset and obtained results. Finally we conferred the best results from these two algorithms, (i.e.) Naive Bayesian provides the better results than LIBLINEAR classification.

5.3 Proposed method for temporal relations detection and classification in clinical narratives

Detection of temporal relation between a pair of events/expressions or events/events is very significant in temporal ordering of the clinical events in timeline and the medical applications [14]. Moreover, relation detection is considered as a classification problem in temporal relation annotated corpus. Therefore, many relation classification approaches have been proposed with the available temporal relation annotated corpora in the existing studies [68].

In this work, we have classified three relationships (AFTER, BEFORE and OVERLAP) between pairs of event or expressions among 13 types of relations.

The proposed approach for temporal relation extraction consists of two tasks. They are:

- Generating candidate pairs from extracted temporal events and expressions,
- Classifying the temporal relationship between the pair of events/expressions.

Figure 5.2 shows these two tasks of candidate pairs generation and temporal relation classification in a graphical representation.

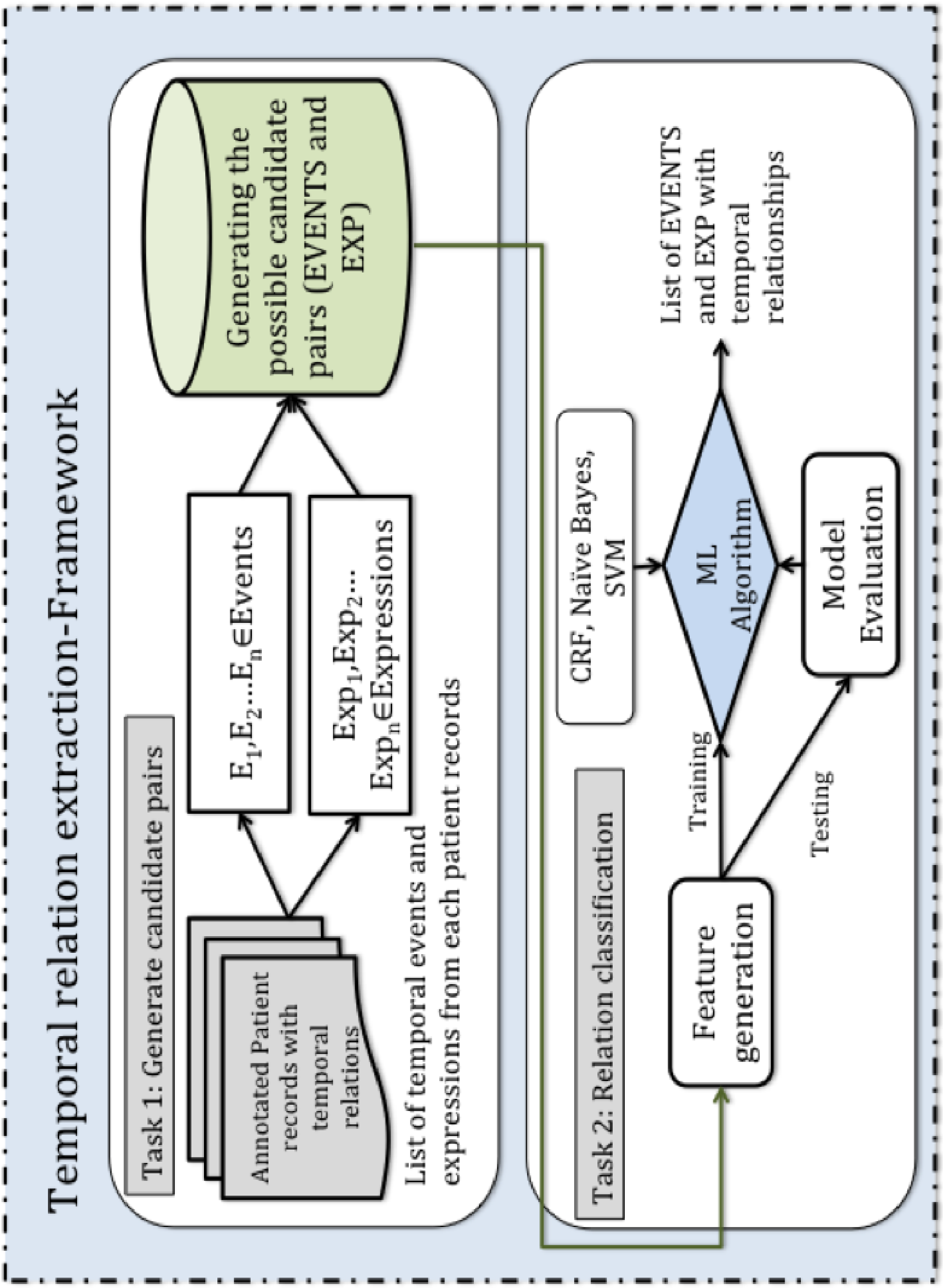

Figure 5.2: Framework for temporal relation classification

5.3.1 Candidate pairs generation

Candidate pair generation from the list of extracted events and expressions is a fundamental task for temporal relation classification. In our proposed approach, we have devised a new hypothesis for candidate pair generation based on the attributes of events and expressions along with the dependency parsing approach. In the proposed assumption, the temporal relationship exist between events or expressions that has an event types (PROBLEM, TEST, TREATMENT, EVIDENTIAL, DATE) in a sentence as shown below. The event type is identified and classified based on domain knowledge system in UMLS and HeidelTime. If the mixture of event types is in a sentence, it is considered the candidate pairs. The following hypothesis formulated based on our heuristic analysis:

a) PROBLEM,PROBLEM

b) PROBLEM and PROBLEM

c) PROBLEM and DATE

d) TREATMENT of PROBLEM or TREATMENT for PROBLEM

e) TEST for PROBLEM

f) EVIDENTIAL from TEST

g) TREATMENT and DATE

h) EVIDENTIAL and DATE

The following example sentences explain the mixture of event/ expression types as explained in the above hypothesis:

1. The patient diabetes mellitus has been complicated by retinopathy, nephropathy and peripheral neuropathy (assumption a, b).

2. The report shows that the patient takes 14 units of NPH insulin to control blood sugar twice a day (assumption c, d, g, h).

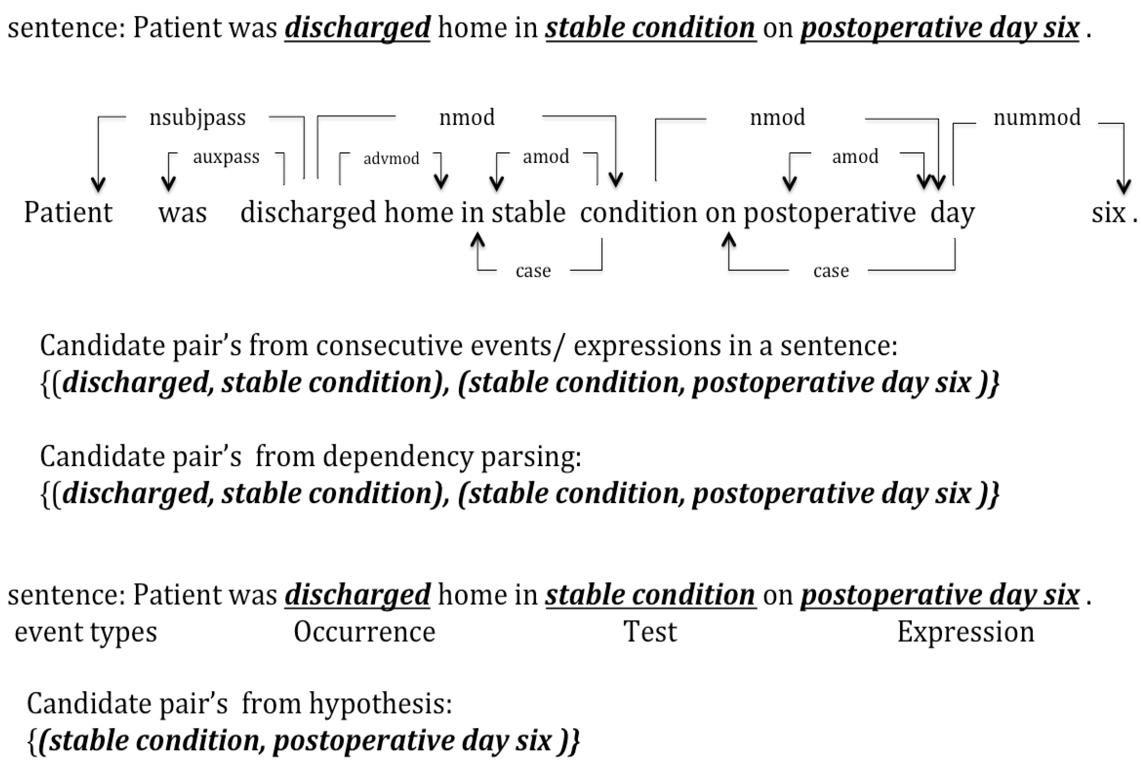

Figure 5.3: An example of candidate pair generation for intra-sentence temporal relation classification.

Patient underwent ***cadaveric pancreas transplantation*** without ***complication*** .

.....

Patient ***renal function*** also ***remained stable*** in the perioperative period .

.....

Patient was ***discharged*** home in ***stable condition*** on ***postoperative day six*** .

Candidate pair's from dependency parsing :
{(*cadaveric pancreas transplantation, complication*), (*renal function, remained stable*)(*discharged, stable condition), (stable condition, postoperative day six)}*

Candidate pair's from hypothesis: {*(cadaveric pancreas transplantation, complication),* ***(cadaveric pancreas transplantation, stable condition), (cadaveric pancreas transplantation, postoperative day six)*** *}*

Figure 5.4: An example of candidate pair generation for inter-sentence and section-cross temporal relation classification (Highlighted the pairs generated by proposed hypothesis).

3. Oral glucose tolerance tests show that the patient has type 1 diabetes (assumption e, f).

We adopted the dependency parsing strategy as explained in [18] to generate the candidate pairs from the list of events and expressions. However, this approach is unable to discover all possible candidate pairs. Therefore, along with existing approaches, we come up with the above newly formulated assumption to discover the potential candidate pairs as shown in Figure 5.3 and Figure 5.4. Finally, we integrated the candidate pairs from both approaches and eliminate duplicates. This method helps to discover almost all possible candidate pairs from clinical narratives. After successfully generating the candidate pairs with target labels from annotated dataset, we move to the temporal relation classification task. This classifier model subsequently improves the performances for inter-sentences, intra sentences and section-cross classifiers.

5.3.2 Temporal relation classification

We have exploited Naive Bayes Classifier from Weka toolkit[1] to train the classifier for temporal relation classification. In this approach, we are estimating the probability distribution of label for the unknown candidate pairs $f : X \rightarrow Y$ or equally P(Y|X). In other words, X = $\left\{ X_1, X_2, X_3 ... X_n \right\}$ denotes the features of raw clinical text and Y denotes the label of temporal relationship between the candidate pairs. This is the best way of developing the classifier model P(Y|X) to exploit the training data to estimate the P(X | Y) and P(Y). The proposed framework for temporal relation classification is shown in figure 5.2. The below steps are followed to train the classifiers and estimate probability distribution for predicting unknown label in intra-sentence, inter-sentence and section cross relationships of our proposed method.

1. Step 1: Preprocessing the training and testing data
2. Step 2: Generating feature set consisting of dependency related features, position information, time-related features, event related features and distance between two events/expressions for three different classifiers (inter-sentence, intra-sentence and section-cross)
3. Step 3: Training the classifier models using Naive Bayesian Classifier with the prepared training data
4. Step 4: Estimating the probability distribution for training data using developed model
5. Step 5: Evaluating the established classifier models with the test run data for inter-sentence, intra-sentence and section-cross relationships.

Let us consider the following sentence in clinical narrative from Figure 1 processed through our proposed classification method : "The patient was admitted with type 1

[1] http://www.cs.waikato.ac.nz/ml/weka/

diabetes mellitus who is four months post cadaveric kidney transplantation and now good graft function".

As a result, we obtain the following output: <EVENT1=type 1 diabetes mellitus | EVENT2=postcadaveric kidney transplantation | Temporal relation=OVERLAP >

<EVENT2=postcadaveric kidney transplantation | TIMEX3=four months | Temporal relation=AFTER >

<EVENT2=postcadaveric kidney transplantation | EVENT3=good graft function | Temporal relation=OVERLAP >

After the achieving of generated features, Temporal relation classification experiment was carried out with the intra-sentence, inter- sentence and section-cross relations to obtain the baseline result and the best results on the generated candidate pairs. We trained our proposed Naive Bayes classifiers and LIBLINEAR SVM classifier on candidate pairs generated by dependency parsing approach and candidate pairs generated by proposed hypothesis. The obtained results clearly show the improvement in accuracy of classification in relation classification from clinical narratives. The experimental results of our proposed method for temporal relation classification is discussed in the evaluation section 3.4.3 and discussion section 5.4.3 in detail.

5.4 Experimental evaluation

The purpose of this experiment is to evaluate the efficacy of our proposed method for temporal relation identification and classification from clinical narratives with Naive Bayesian classifier as stated in the objective. To conduct the experiment we used the i2b2 annotated data set as a key resource.

5.4.1 Evaluation metrics

As explained in previous chapters, we evaluated our developed models with three measures, which are Precision, Recall and F-measure as given by:

$$Precision = \frac{|System.Output \cap Annotated.Corpus|}{|System.Output|} \tag{5.1}$$

$$Recall = \frac{|System.Output \cap Annotated.Corpus|}{|Annotated.Corpus|} \tag{5.2}$$

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall}. \tag{5.3}$$

5.4.2 Experimental results

Temporal relations	**Precision**	**Recall**	**F-measure**
Baseline (SVM)	71.42	39.54	50.90
Dependency - SVM (Overall)	77.17	50.32	60.91
Dependency + Hypothesis - SVM (Overall)	81.43	53.75	64.76
Baseline (Naive Bayes)	83.39	32.54	46.81
Naive Bayes (Intra-sentence)	74.4	75.6	74.2
Naive Bayes (Inter-sentence)	66.3	66.4	66.34
Naive Bayes (Section-cross)	69.2	72.4	70.7
Dependency - Naive Bayes (Overall)	67.12	63.76	***65.39***
Dependency + Hypothesis-Naive Bayes (Overall)	66.8	68.4	***67.1***

Table 5.1: Evaluation results of temporal relation classification

We initially developed and evaluated the baseline model using gold standard annotated data of temporal relations. From Table 5.1, We can notice the improved performance of temporal relation classification with 3 different classifiers results (inter-sentence, intra-sentence and section-cross). Moreover, overall F-measure of temporal relation classification is increased from 46.81 to 67.1. From these results, it can be understood that fluctuating recall and biased precision are affecting the classification performance. However, the recall of proposed method has increased drastically from baseline model and biased precision has stabilized. These effects are due to incorporation of effective candidate pairs generation hypothesis from our proposed method.

5.4.3 Error analysis

Though the proposed method could improve the performance of temporal relation classification (precision and recall), still it has a room for improvement on generating candidate pairs. Since some of the pairs generated by our proposed approach does not have any relationships. Hence, in the future, we plan to focus on finding new strategies for candidate pairs generation and subsequently improve the precision, recall and f-measure of temporal relation classification.

5.5 Summary

In this chapter, we have presented hybrid approach with a novel assumption on candidate pairs generation and Naive Bayesian classification for temporal relation detection and classification. Understanding the temporal relationship between events consist of temporal boundaries covered by the explicit and implicit expressions. We formulated a new assumption on generating and identifying the potential candidate pairs from list of temporal events or expressions that can appropriately relate events/expressions in clinical narratives based on their attributes. Moreover, to address the problem of temporal relation detection, we exploited Naive Bayesian Classifier to detect the temporal relationship among the identified pair's. The effective candidate pair's generation helps to improve the relation classification performance.

Chapter 6

Conclusion and future research

This chapter provides the overall summery, contribution and conclusion of our thesis. Also we discussed about our future research works to exploit the extracted temporal information on developing applications.

Due to the large availability and ambiguity of temporal information in clinical narratives, computers have difficulties in temporal reasoning and extracting temporal information from EMRs automatically. We aim to support on reconstructing the clinical narratives from patient records with structured representation of temporal information for further processing by machines such as temporal reasoning and visualizations. Temporal Information extraction over the clinical text is an active area for the past several years. Current research trend implies a promising future applications of temporal information extraction in the clinical text and its exploitation in medical care.

In this dissertation, we presented three tasks for the temporal information extraction in clinical narratives as such for temporal expressions, temporal events and temporal relations as stated in the objective.

In chapter 3, temporal expressions recognition in clinical narratives, A novel feature set has been proposed to address the problem of temporal expressions extraction. Our proposed framework has feature set from two key components. First set of features are generated and adopted from HeidelTime system, whereas the second feature set is obtained from raw clinical narratives that are appropriate for temporal expressions extraction. Existing methods either having the advantage of HeidelTime or developing rules/ machine learning models, but not the integrated components of both. Those properties helps to extract temporal expressions effectively.

In chapter 4, we have presented semi-supervised framework for the temporal event extraction in clinical narratives. The novel framework was developed for the semi-supervised approach to extract temporal events from the rich unannotated text by extending annotated corpora. The proposed method has two stages to achieve the semi-supervised frame work. Key contribution in this chapter is, besides using annotated data, we utilized the simple K-means clustering algorithm to select the rich and sparse data from abundant unannotated data. We reported the achieved results for event extraction from our proposed method. We also witnessed the strong and steady improvement in event extraction accuracy in semi-supervised framework. Compared to the baseline system, performance and stability of the semi-supervised method for event extraction is significantly increased due to utilization of unannotated data, our novel assumption, incorporation of concept unique identifier and semantic group features from UMLS. In semi-supervised framework, we extended the unannotated data by feature mapping and trained our model. In this work we proved that the influence of unannotated clinical text helps to improve the accuracy of temporal event extraction significantly.

In chapter 5, temporal relation identification and classification, we formulated a novel hypothesis for generating candidate pairs and exploited the Naive Bayes Classifier. Effectiveness of generated candidate pairs helps to increase the performance of relation classification between events or expressions. We conferred the achieved results for relations classification from our proposed approach in a long-winded discussion in section 5.4.2. In future, we are planning to expand candidate generation heuristics along with the proposed hypothesis to improve the performance of temporal relation classification.

Publications during doctor course

[1] MOHARASAN Ganddhimathi and Tu Bao Ho. Extraction of temporal information from clinical narratives, *Special Issue on Healthcare Knowledge Discovery and Management (Springer)*, Submitted on Journal of Healthcare Informatics Research, 2019 (Published).

[2] MOHARASAN Gandhimathi and Tu Bao Ho. Extraction of Temporal Events from Clinical Text Using Semi-supervised Conditional Random Fields, *Accepted on International Conference on Data Mining and Big Data (DMBD)*, pages=409-421, July 27th -August 1st 2017, Fukuoka, Japan.

[3] MOHARASAN Gandhimathi and Tu Bao Ho. A semi-supervised approach for temporal information extraction from clinical text, *Accepted on The 12th IEEE-RIVF International Conference on Computing and Communication Technologies (IEEE-RIVF)*, pages=7-12, 7-9 November 2016, Hanoi, Vietnam.

[4] MOHARASAN Gandhimathi and Tu Bao Ho. Challenges and Opportunities of Temporal Information Extraction from Electronic Medical Records, *Accepted on Asia Conference on Information Systems (ACIS)*, pages=386-393, 1-3 December 2014, Nha Trang, Vietnam.

[5] MOHARASAN Gandhimathi and Tu Bao Ho. Multilayered Approach to Extract Temporal Events and Expressions from Clinical Narratives, *Poster accepted on The 5th Annual Translational Bioinformatics Conference*, pages=262-266, 7-9 November 2015, Tokyo, Japan.

[6] MOHARASAN Gandhimathi and Tu Bao Ho. NTCIR-12: Temporal Intent Disambiguation (TID) subtask: Naive Bayesian Classifier to predict temporal classes, *NTCIR Conference*, pages=262-266, 7-10 June 2016, NII, Tokyo, Japan.

Bibliography

[1] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. In *Corpus linguistics*, volume 2003, page 40, 2003.

[2] Roser Saurí, Jessica Littman, Robert Gaizauskas, Andrea Setzer, and James Pustejovsky. Timeml annotation guidelines, version 1.2. 1. 2006.

[3] Nathanael Chambers, Shan Wang, and Dan Jurafsky. Classifying temporal relations between events. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 173–176, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[4] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, and James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 75–80. Association for Computational Linguistics, 2007.

[5] Naushad UzZaman and James F Allen. Trips and trios system for tempeval-2: Extracting temporal information from text. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 276–283. Association for Computational Linguistics, 2010.

[6] James F. Allen. An interval-based representation of temporal knowledge. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'81, pages 221–226, 1981.

[7] Andrea Setzer and Robert J Gaizauskas. Annotating events and temporal information in newswire texts. In *Proceedings of LREC-2000*, volume 2000, pages 1287–1294, 2000.

[8] DM Roden, H Xu, JC Denny, and RA Wilke1. Electronic medical records as a tool in clinical pharmacology: Opportunities and challenges. *Clin Pharmacol Ther.*, 91(6):322–329, 2012.

[9] Ozlem Uzuner Weiyi Sun, Anna Rumshisky. Temporal reasoning over clinical text: the state of the art. *Journal of the American Medical Informatics Association*, 20(5):814–819, 2013.

[10] Simon Parsons George Hripcsak Li Zhou, Carol Friedman. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. *American Medical Informatics Association*, 2005:869, 2005.

[11] George Hripcsak Li Zhou. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, 40(2):183–202, 2007.

[12] Randall A. Brown Yu-Kai Lin, Hsinchun Chen. Medtime: A temporal information extraction system for clinical narratives. *Journal of Biomedical Informatics*, 46:s20–s28, 2013.

[13] Xiaodan Zhu, Colin Cherry, Svetlana Kiritchenko, Joel Martin, and Berry De Bruijn. Detecting concept relations in clinical text: Insights from a state-of-the-art model. *J. of Biomedical Informatics*, 46(2):275–285, April 2013.

[14] Yi Liu, Paea LePendu, Srinivasan Iyer, and Nigam H Shah. Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Summits on Translational Science proceedings*, 2012:47, 2012.

[15] Chien-Chih Wang, Ming-Nan Chien, Chua-Huang Huang, and Li Liu. A rule-based disease diagnostic system using a temporal relationship model. In *Fuzzy Systems and Knowledge Discovery, 2007. FSKD 2007. Fourth International Conference on*, volume 4, pages 109–115. IEEE, 2007.

[16] Juan Carlos Augusto. Temporal reasoning for decision support in medicine. *Artificial Intelligence in*, 33(1):1–24, 2005.

[17] Uzuner Ozlem Sun Weiyi, Rumshisky Anna. Evaluating temporal relations in clinical text: 2012 i2b2 challenge,. *Journal of the American Medical Informatics Association*, pages amiajnl–2013, 2013.

[18] Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C Denny, and Hua Xu. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835, 2013.

[19] Ozlem Uzuner, Peter Szolovits, and Isaac Kohane. i2b2 workshop on natural language processing challenges for clinical records. In *Proceedings of the Fall Symposium of the American Medical Informatics Association*. Citeseer, 2006.

[20] Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814, 2015.

[21] Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. Semeval-2016 task 12: Clinical tempeval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval 2016), San Diego, California, June. Association for Computational Linguistics*, pages 962–972, 2016.

[22] Harry Bunt Laurent Romary James Pustejovsky, Kiyong Lee. Iso-timeml: An international standard for semantic annotation. *LREC 2010*, 2010.

[23] K. Uehara S. Higashiyama, K. Seki. Clinical entity recognition using cost-sensitive structured perceptron for ntcir-10 mednlp. 2013.

[24] E. Aramaki H. Imachi, M. Morita. Ntcir-10 mednlp task baseline system. *Proceedings of NTCIR*, 10(710-712), 2013.

[25] Tran-Thai Dang, Phetnidda Ouankhamchan, and Tu-Bao Ho. Detection of new drug indications from electronic medical records. In *Computing & Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2016 IEEE RIVF International Conference on*, pages 223–228. IEEE, 2016.

[26] Kerstin Denecke and Yihan Deng. Sentiment analysis in medical settings: New opportunities and challenges. *Artificial intelligence in medicine*, 64(1):17–27, 2015.

[27] Sujan Perera, Amit Sheth, Krishnaprasad Thirunarayan, Suhas Nair, and Neil Shah. Challenges in understanding clinical notes: Why nlp engines fall short and where background knowledge can help. In *Proceedings of the 2013 international workshop on Data management & analytics for healthcare*, pages 21–26. ACM, 2013.

[28] Rebecca S Crowley, Melissa Castine, Kevin Mitchell, Girish Chavan, Tara McSherry, and Michael Feldman. caties: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research. *Journal of the American Medical Informatics Association*, 17(3):253–264, 2010.

[29] Guergana Savova Dingcheng Li, Karin Kipper-Schuler. Conditional random fields and support vector machines for disorder named entity recognition in clinical texts. *Association for Computational Linguistics*, pages 94–95, 2008.

[30] Lamont Antieau Yonggang Cao Hong Yu Zuofeng Li, Feifan Liu. Lancet: a high precision medication event extraction system for clinical text. *Journal of the American Medical Informatics Association*, 17(5):563–567, 2010.

[31] Bryan Rink, Sanda Harabagiu, and Kirk Roberts. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association*, 18(5):594–600, 2011.

[32] Brian L Cairns, Rodney D Nielsen, James J Masanz, James H Martin, Martha S Palmer, Wayne H Ward, and Guergana K Savova. The mipacq clinical question answering system. In *AMIA Annual Symposium Proceedings*, volume 2011, page 171. American Medical Informatics Association, 2011.

[33] Yonggang Cao, Feifan Liu, Pippa Simpson, Lamont Antieau, Andrew Bennett, James J Cimino, John Ely, and Hong Yu. Askhermes: An online question answering system for complex clinical questions. *Journal of biomedical informatics*, 44(2):277–288, 2011.

[34] Gaurav Trivedi, Phuong Pham, Wendy Chapman, Rebecca Hwa, Janyce Wiebe, and Harry Hochheiser. An interactive tool for natural language processing on clinical text. *arXiv preprint arXiv:1707.01890*, 2017.

[35] Omar Alonso, Michael Gertz, and Ricardo Baeza-Yates. On the value of temporal information in information retrieval. In *ACM SIGIR Forum*, volume 41, pages 35–41. ACM, 2007.

[36] Hector Ambit and Consuelo Gonzalo. Clinical narrative analytics challenges. In *Rough Sets: International Joint Conference, IJCRS 2016, Santiago de Chile, Chile, October 7–11, 2016, Proceedings*, volume 9920, page 23. Springer, 2016.

[37] Yonghui Wu, Joshua C Denny, S Trent Rosenbloom, Randolph A Miller, Dario A Giuse, and Hua Xu. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. In *AMIA annual symposium proceedings*, volume 2012, page 997. American Medical Informatics Association, 2012.

[38] Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, pages 299–307, 2014.

[39] Sebastien Dubois, David C Kale, Nigam Shah, and Kenneth Jung. Learning effective representations from clinical notes. *arXiv preprint arXiv:1705.07025*, 2017.

[40] Yangqiu Song and Dan Roth. Unsupervised sparse vector densification for short text similarity. In *HLT-NAACL*, pages 1275–1280, 2015.

[41] Efsun Sarioglu, Kabir Yadav, and Hyeong-Ah Choi. Topic modeling based classification of clinical reports. In *ACL (student research workshop)*, pages 67–73, 2013.

[42] Corey Arnold and William Speier. A topic model of clinical reports. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, pages 1031–1032. ACM, 2012.

[43] Blaylock Nate Lucian Galescu. A corpus of clinical narratives annotated with temporal information. *ACM*, pages 715–720, 2012.

[44] G.Niklas Norn, Johan Hopstadius, Andrew Bate, Kristina Star, and I.Ralph Edwards. Temporal pattern discovery in longitudinal electronic patient records. *Data Mining and Knowledge Discovery*, 20(3):361–387, 2010.

[45] Ozlem Uzuner Weiyi Sun, Anna Rumshisky. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46:s5–s12, 2013.

[46] Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. Temporal annotation in the clinical domain.

[47] Alexander Nakhimovsky. Temporal reasoning in natural language understanding: the temporal structure of the narrative. In *Proceedings of the third conference on European chapter of the Association for Computational Linguistics*, pages 262–269. Association for Computational Linguistics, 1987.

[48] Leilei Sun, Chuanren Liu, Chonghui Guo, Hui Xiong, and Yanming Xie. Data-driven automatic treatment regimen development and recommendation. In *KDD*, pages 1865–1874, 2016.

[49] Michael Stacey and Carolyn McGregor. Temporal abstraction in intelligent clinical data analysis: A survey. *Artificial intelligence in medicine*, 39(1):1–24, 2007.

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

[51] Ricardo Martinho. Text mining applied to electronic medical records. 2015.

[52] Peter B Jensen, Lars J Jensen, and Søren Brunak. Mining electronic health records: towards better research applications and clinical care. *Nature reviews. Genetics*, 13(6):395, 2012.

[53] Lance De Vine, Mahnoosh Kholghi, Guido Zuccon, Laurianne Sitbon, and Anthony Nguyen. Analysis of word embeddings and sequence features for clinical information extraction. 2015.

[54] Abha Agrawal. Medication errors: prevention using information technology systems. *British journal of clinical pharmacology*, 67(6):681–686, 2009.

[55] Dingcheng Li Siddhartha Jonnalagadda Cui Tao Ravikumar Komandur Elayavilli Hongfang Liu Sunghwan Sohn, Kavishwar B. Wagholikar. Comprehensive temporal information detection from clinical text: medical events, time, and tlink identification. *JAMIA*, 20(5):836–842, 2013.

[56] M. Palmer S. Pradhan P. C. de Groen B. Erickson T. Miller C. Lin G. Savova J. Pustejovsky S. Bethard, S. Finan. Temporal annotation in the clinical domain. 2013.

[57] Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. A hybrid approach for event extraction. *Polibits*, (46):55–59, 2012.

[58] Sumithra Velupillai, D Mowery, Brett R South, Maria Kvist, and Hercules Dalianis. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics*, 10(1):183, 2015.

[59] Özlem Uzuner, Amber Stubbs, and Weiyi Sun. Chronology of your health events: Approaches to extracting temporal relations from medical narratives. *Journal of biomedical informatics*, 46:S1, 2013.

[60] Moharasan Gandhimathi and Ho Tu Bao. Challenges and opportunities in temporal information extraction from electronic medical records.

[61] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57:28–37, 2015.

[62] James Pustejovsky and Marc Verhagen. Semeval-2010 task 13: evaluating events, time expressions, and temporal relations (tempeval-2). In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 112–116. Association for Computational Linguistics, 2009.

[63] M. Hepple G. Demetriou Y. Guo A. Setzer I. Roberts A. Roberts, R. Gaizauskas. Semantic annotation of clinical text: The clef corpus. *Proceedings of the LREC 2008 workshop on building and evaluating resources for biomedical text mining*, pages 19–26, 2008.

[64] Ruth M Reeves, Ferdo R Ong, Michael E Matheny, Joshua C Denny, Dominik Aronsky, Glenn T Gobbel, Diane Montella, Theodore Speroff, and Steven H Brown. Detecting temporal expressions in medical narratives. *International journal of medical informatics*, 82(2):118–127, 2013.

[65] Timothy A Miller, Steven Bethard, Dmitriy Dligach, Chen Lin, and Guergana K Savova. Extracting time expressions from clinical text. 2010.

[66] Mei Liu S Trent Rosenbloom Subramani Mani Joshua C Denny Hua Xu Min Jiang, Yukun Chen. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *Journal of the American Medical Informatics Association*, 18(5):601–606, 2011.

[67] Prateek Jindal and Dan Roth. Extraction of events and temporal expressions from clinical narratives. *Journal of Biomedical Informatics*, 46, Supplement(0):S13 – S19, 2013. 2012 i2b2 {NLP} Challenge on Temporal Relations in Clinical Data.

[68] Johnny Chi-Yang W Jian-Ming Chen Richard Tzong-Han Tsai Wen-Lian Hsu Yung-Chun Chang, Hong-Jie Dai. Tempting system: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of Biomedical Informatics*, 46:s54–s62, 2013.

[69] Jannik Strötgen and Michael Gertz. Multilingual and cross-domain temporal tagging. *Language Resources and Evaluation*, 47(2):269–298, 2013.

[70] Angel X Chang and Christopher D Manning. Sutime: A library for recognizing and normalizing time expressions. In *LREC*, volume 2012, pages 3735–3740, 2012.

[71] Marc Verhagen, Roser Sauri, Tommaso Caselli, and James Pustejovsky. Semeval-2010 task 13: Tempeval-2. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 57–62. Association for Computational Linguistics, 2010.

[72] Buzhou Tang, Yonghui Wu, Min Jiang, Joshua C Denny, and Hua Xu. Recognizing and encoding discorder concepts in clinical text using machine learning and vector space model. In *CLEF (Working Notes)*, volume 1179. 2013.

[73] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289, 2001.

[74] Charles Sutton and Andrew McCallum. An introduction to conditional random fields. *arXiv preprint arXiv:1011.4088*, 2010.

[75] Fuchun Peng and Andrew McCallum. Information extraction from research papers using conditional random fields. *Information processing & management*, 42(4):963–979, 2006.

[76] Pranjal Awasthi, Delip Rao, and Balaraman Ravindran. Part of speech tagging and chunking with hmm and crf. *Proceedings of NLP Association of India (NLPAI) Machine Learning Contest 2006*, 2006.

[77] Miikka Silfverberg, Teemu Ruokolainen, Krister Linden, Mikko Kurimo, et al. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *ACL (2)*, pages 259–264, 2014.

[78] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics, 2004.

[79] Yassine Benajiba and Paolo Rosso. Arabic named entity recognition using conditional random fields. In *Proc. of Workshop on HLT & NLP within the Arabic World, LREC*, volume 8, pages 143–153. Citeseer, 2008.

[80] Markus Bundschus, Mathaeus Dejori, Martin Stetter, Volker Tresp, and Hans-Peter Kriegel. Extraction of semantic biomedical relations from text using conditional random fields. *BMC bioinformatics*, 9(1):1, 2008.

[81] Hector Llorens, Estela Saquete, and Borja Navarro-Colorado. Timeml events recognition and classification: learning crf models with semantic roles. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 725–733. Association for Computational Linguistics, 2010.

[82] Hector Llorens, Estela Saquete, and Borja Navarro. Tipsem (english and spanish): Evaluating crfs and semantic roles in tempeval-2. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 284–291. Association for Computational Linguistics, 2010.

[83] Anup Kumar Kolya, Asif Ekbal, and Sivaji Bandyopadhyay. A hybrid approach for event extraction. *Polibits*, (46):55–59, 2012.

[84] Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. Machine learning of temporal relations. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association

for Computational Linguistics, ACL-44, pages 753–760, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[85] Anup Kumar Kolya and Asif Ekbal. A supervised machine learning approach for temporal information extraction. 2011.

[86] Min Jiang Yukun Chen-Joshua C Denny-Hua Xu Buzhou Tang, Yonghui Wu. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association*, 20(5):828–835, 2013.

[87] M. Filannino J. A. Keane G. Nenadic A. Kovačević, A. Dehghan. Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *Journal of the American Medical Informatics Association*, pages amiajnl–2013, 2013.

[88] Kirk Roberts, Bryan Rink, and Sanda M Harabagiu. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association*, 20(5):867–875, 2013.

[89] Yueh-Lin Yang, Po-Ting Lai, and Richard Tzong-Han Tsai. A hybrid system for temporal relation extraction from discharge summaries. In *Technologies and Applications of Artificial Intelligence*, pages 379–386. Springer, 2014.

[90] Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, I Eric, and Chao Chang. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):849–858, 2013.

[91] Xiaojin Zhu. Semi-supervised learning literature survey. 2005.

[92] Wenhui Liao and Sriharsha Veeramachaneni. A simple semi-supervised algorithm for named entity recognition. In *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pages 58–65. Association for Computational Linguistics, 2009.

[93] Feng Jiao, Shaojun Wang, Chi-Hoon Lee, Russell Greiner, and Dale Schuurmans. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, pages 209–216. Association for Computational Linguistics, 2006.

[94] David Yarowsky. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd annual meeting on Association for Computational Linguistics*, pages 189–196. Association for Computational Linguistics, 1995.

[95] Jordi Poveda, Mihai Surdeanu, and Jordi Turmo. An analysis of bootstrapping for the recognition of temporal expressions. In *Proceedings of the NAACL HLT 2009 workshop on semi-supervised learning for natural language processing*, pages 49–57. Association for Computational Linguistics, 2009.

[96] Yves Grandvalet and Yoshua Bengio. Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems*, pages 529–536, 2004.

[97] Andreas Vlachos. *Semi-supervised learning for biomedical information extraction.* PhD thesis, University of Cambridge, 2010.

[98] Andrew Carlson, Justin Betteridge, Richard C Wang, Estevam R Hruschka Jr, and Tom M Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 101–110. ACM, 2010.

[99] Liu Yang, Rong Jin, and Rahul Sukthankar. Semi-supervised learning with weakly-related unlabeled data: Towards better text categorization. In *Advances in Neural Information Processing Systems*, pages 1857–1864, 2009.

[100] Seyed Abolghasem Mirroshandel and Gholamreza Ghassem-Sani. Towards unsupervised learning of temporal relations between events. *Journal of Artificial Intelligence Research*, 2012.

[101] Paramita Mirza. Extracting temporal and causal relations between events. *arXiv preprint arXiv:1604.08120*, 2016.

[102] Charu C Aggarwal and ChengXiang Zhai. A survey of text clustering algorithms. In *Mining text data*, pages 77–128. Springer, 2012.

[103] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[104] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE transactions on pattern analysis and machine intelligence*, 24(7):881–892, 2002.

[105] Carson Tao, Michele Filannino, and Özlem Uzuner. Prescription extraction using crfs and word embeddings. *Journal of biomedical informatics*, 72:60–66, 2017.

[106] James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[107] Marc B Vilain and Henry A Kautz. Constraint propagation algorithms for temporal reasoning. In *Aaai*, volume 86, pages 377–382, 1986.