MEDICAL EVENT TIMELINE GENERATION FROM CLINICAL NARRATIVES

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of The Ohio State University

By

Preethi Raghavan, B.S., M.S.

Graduate Program in Computer Science & Engineering

The Ohio State University

2014

Dissertation Committee:

Prof. Eric Fosler-Lussier, Adviser

Prof. Albert M. Lai, Co-Adviser

Prof. Brian Kulis

ABSTRACT

Extracting information from disparate clinical data sources in electronic health records is crucial to building intelligent systems that can reason with clinical variables and support decision making. This dissertation describes a novel framework for representing and reasoning with medical events and temporal information, in unstructured clinical narratives, by using linguistic insights from clinical text in training machine learning models for timeline extraction. Importantly, we leverage both temporal and semantic representations of medical events in learning structured relationships within and across clinical data sources and creating a longitudinal timeline of events over the patient's history.

To this end, the main problems addressed in this work are medical event coreference resolution and temporal relation learning, both in intra- and cross-document settings, and information fusion across structured and unstructured data. While prior work in clinical informatics has addressed some of these problems in a limited capacity, other problems like cross-narrative temporal ordering and information fusion are being addressed for the first time in this dissertation.

The generated timeline has important implications in various clinical applications with temporal constraints such as patient recruitment for clinical trials, medical document summarization, adverse drug reaction mining, question answering and clinical decision making. We explore the utility of the timeline in resolving temporal eligibility criteria for clinical trial recruitment.

Dedicated to Mummy and Daddy

ACKNOWLEDGMENTS

I begin by expressing immense gratitude to my advisors, Eric Fosler-Lussier and Albert M. Lai, for guiding me through my doctoral studies. It is their scholarly expertise and constant support that has got me to this culmination point.

Eric agreed to advise me in, what was back then, a fairly unfamiliar area of biomedical NLP. I have greatly benefited from his critical insights, excellent writing skills and overall mentorship. I deeply admire his ability to think through complex problems and come up with the most creative solutions. I am extremely fortunate to have had the opportunity to work with and learn from him. If not for Albert, I would not have taken up research in the area of clinical informatics. It is through numerous discussions with him that I got introduced to the world of healthcare and realized how much artificial intelligence can help biomedical informatics. He has always provided sound advice and great insights in informatics, while being very patient and friendly. Through their excellent advise, and prodding and nudging in the right direction, they have been instrumental in my transformation into a researcher; I am really going to miss working with them.

I am grateful to Rajiv Ramnath and Jay Ramanathan for advising me during my Masters and for always supporting and encouraging me even afterwards. I thank Chris Brew for his advice in the beginning of my PhD, and Philip Payne for introducing me to Albert and setting me off on this journey. I would also like to acknowledge my candidacy and dissertation committee members, William Schuler, Brian Kulis, and Harvey J. Miller, for

their feedback, time and support. I would like to thank the NIH and NLM grants that supported my PhD work. I thank Noemie Elhadad and Steve Johnson for providing input during our grant meetings, and everyone at the Clippers weekly discussion group for their feedback. I am grateful to Raghu Machiraju, under whom I did my first TA at OSU, for his support. I thank the CSE and BMI staff members (especially James Gentry) who have patiently helped me navigate through all the administrative work. I am also grateful to the many annotators for marking the clinical narratives with tedious annotations.

I am grateful to all the past and current members of the SLaTe lab for their camaraderie over the years. Through lab lunches, and discussions and laughter about everything under the sun, we maintained our sanity. I would especially like to thank Billy Hartmann, Preethi Jyothi, Rohit Prabhavalkar, Ryan He, Yi Ma, Joo-Kyung Kim and Chaitanya Shivade. Billy made me feel welcome in the lab when I was new, and has always kept us entertained through his unique sense of humor. Preethi, who is also a very dear friend, was the best and most patient sounding board for numerous problems along the way. I'd like to thank Rohit for introducing me to semi-supervised methods that were useful in coreference resolution, Ryan for his input on WFSTs, and everyone else for the many discussions that have helped me along the way.

For making my life in OSU incredible fun, I have many friends to thank. Time spent with Arun, Preethi, Sivaguru, Harsha, Rashmi, Farha, Mukundan, Jeeth, Srikanth and the entire complex gang has always been a lot a fun. I am thankful to them for the unforgettable memories through long tea sessions, random debates, and all-nighters. Endless conversation and laughter with them has cheered me up on many a gloomy day.

My life with Arun has been synonymous with my PhD journey. He has brought a lot of love, happiness, good food and tea in my life. We've seen each other through exams,

research, friendship and marriage in our time at OSU. His support and encouragement has helped me tide through many a difficult day; I thank him for always being there for me. I also thank my sister Usha and her family, and my uncle Srinivas, for their love and support.

My parents, A. Raghavan and Lalita Raghavan, have been my biggest source of strength and encouragement. They have loved and supported me unconditionally through everything I have chosen to undertake in life. I thank my dad for enthusiastically trying to understand my work and providing editorial feedback on my dissertation draft. I thank my mom for always being there for me, and also for being a multi-talented creator of arts and crafts. I love them both a lot, and anything and everything I am today, is because of them.

VITA

September 26, 1983	Born in Mumbai, Maharashtra, India
May, 2005	B.Tech., Computer Science SNDT University, Mumbai, India
December, 2009	MS, Computer Science The Ohio State University, Columbus, Ohio, USA

PUBLICATIONS

Journal Articles

Chaitanya Shivade, **Preethi Raghavan**, Eric Fosler-Lussier, Peter J. Embi, Noemi Elhadad, Stephen B. Johnson, Albert M. Lai, "Systematic review of approaches to identifying patient phenotype cohorts using Electronic Health Records", *Journal of the American Medical Informatics Association (JAMIA)*, 2013.

Conference Papers and Books

Preethi Raghavan, Eric Fosler-Lussier, Noemie Elhadad, Albert M. Lai, "Cross-narrative temporal ordering of medical events", *Association for Computational Linguistics Annual Meeting (ACL)* 2014.

Preethi Raghavan, James L. Chen, Eric Fosler-Lussier, Albert M. Lai, "How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?", *American Medical Informatics Association (AMIA)* 2014.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai, "Learning to Temporally Order Medical Events in Clinical", *Association for Computational Linguistics Annual Meeting (ACL)* 2012, 70-74.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai, "Temporal Classification of Medical Events", *BioNLP*, 2012. pp. 29-37.

Preethi Raghavan, Eric Fosler-Lussier, Albert M. Lai, "Inter-Annotator Reliability of Medical Events in Clinical Narratives by Annotators with Varying Levels of Clinical Expertise", *American Medical Informatics Association (AMIA)* 2012, pp. 1366.

Preethi Raghavan, Eric Fosler-Lussier, and Albert M. Lai, " Exploring Semi-supervised Medical Concept Coreference Resolution using Semantic and Temporal features", *North American Association for Computational Linguistics (NAACL)*, 2012, pp. 731-741.

Preethi Raghavan, Eric Fosler-Lussier, Chris Brew, and Albert M. Lai, "Medical Event Coreference Resolution using the UMLS Metathesaurus and Temporal Reasoning", *ACM International Health Informatics Symposium, (SIGHIT-IHI)*, 2012, pp. 465-472.

Preethi Raghavan, and Albert M. Lai, " Leveraging Natural Language Processing of Clinical Narratives for Phenotype Modeling", Workshop for Ph.D. *Students in Information and Knowledge Management (PIKM), ACM Conference on Information and Knowledge Management CIKM*, 2010, pp. 57-66.

Preethi Raghavan, Rose Catherine K, Shajith Ikbal, Nanda Kambhatla and Debapriyo Majumdar, " Extracting Problem and Resolution Information from Online Discussion Forums", *International Conference on Management of Data (COMAD)*, 2010, pp. 77.

Preethi Raghavan, Rose Catherine K., Shajith Ikbal and Nanda Kambhatla, "Classification and Retrieval from Mailing Lists and Forums", *Forum for Information Retrieval Evaluation (FIRE)*, 2010.

Preethi Raghavan, Rajiv Ramnath, Jay Ramanathan, Zhe Xu, " Framework for Improving Enterprise Services by Mining Customer Edge Data", *IEEE International Conference on Collaboration, Technologies and Infrastructures (WETICE)*, 2009.

Aman Kumar, **Preethi Raghavan**, Jay Ramanathan, Rajiv Ramnath, " Interaction Ontology for Change Impact Analysis of Complex Systems", *IEEE Asia-Pacific Services Computing Conference (APSCC)*, 2008, IEEE Computer Society, pp. 303-309.

FIELDS OF STUDY

Major Field: Computer Science and Engineering

TABLE OF CONTENTS

Page

LIST OF TABLES

Table **Page**

LIST OF FIGURES

Figure **Page**

CHAPTER 1: INTRODUCTION

Temporal reasoning is a very basic yet vital ability of humans, no matter what language we speak. As noted by Mani [2005], early humans somehow developed a way of reasoning in terms of events and their positions in the stream of time. However, there is considerable cross-linguistic and cross-domain variation in encoding temporal information in natural language. Languages like English use the interplay of tense and aspect to encode temporal information. However, the importance of these features may vary across various domains and tasks. One such natural language domain with distinct sub-language and temporal characteristics is medicine. In this dissertation, we study clinical data in electronic health records and propose methods for extracting a timeline of medical events over the patient's history.

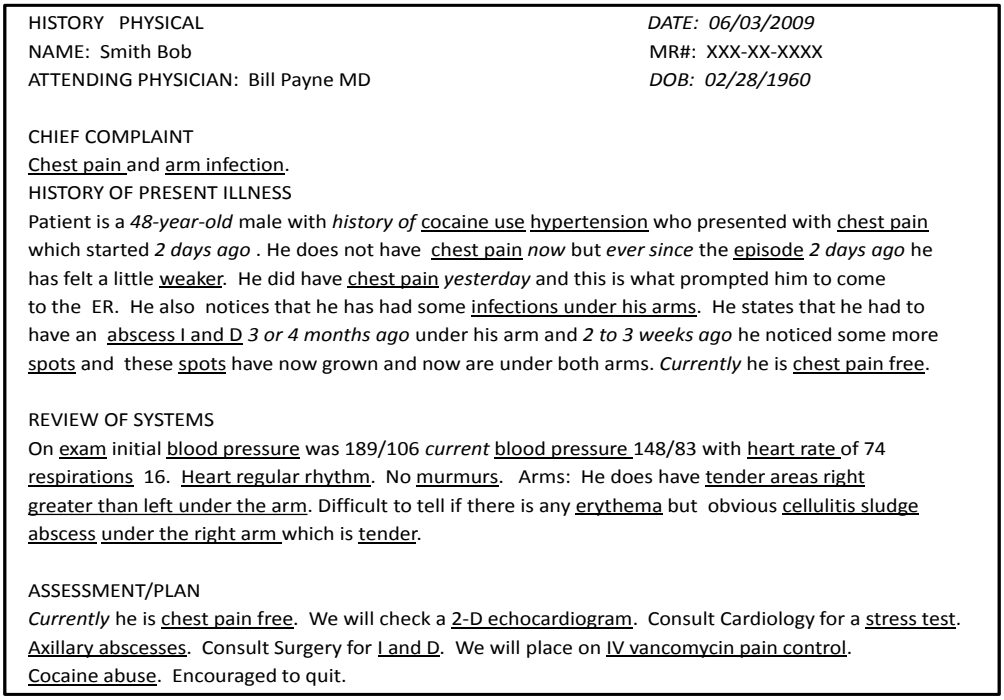
1.1 Electronic Health Records

Whenever a patient visits a health care delivery setting such as a clinic or a hospital, one or more unstructured electronic notes describing his present medical condition, diagnoses and treatments, along with his past medical history gets generated. These notes, usually written by a nurse or a physician, include discharge summaries, radiology and pathology reports, history and physical reports, and are collectively termed as clinical narratives. Unstructured clinical narratives, along with structured patient data, forms part of the electronic health record (EHR) of the patient. The adoption of EHRs in hospitals in the United States

is over 80% as of 2013.[1] A patient could have a large number of different types of clinical narratives in the EHR.

Unstructured clinical narratives in the EHR describe various medical events related to the patient's condition and health care and also contain some information on when these events occurred, as seen in the sanitized "history and physical" report is shown in Figure 1.1. The narrative begins with a semi-structured portion which includes details like the patient and physicians names, the note creation or admission date, the medical record number (MR# in the figure) and the patient's date of birth (DOB). This is followed by an unstructured portion that captures the patient's condition and the health care being provided to the patient. The text in this unstructured portion has mentions of various symptoms, diseases, tests, medications and other medical events concerning the patients healthcare. Some examples of these medical events (underlined phrases in Figure 1.1) include "cocaine use," "hypertension," "chest pain," "blood pressure," "cocaine abuse." Clinical text is also very temporal in nature with frequent mentions of temporal expressions indicating when a medical event occurred relative to other events in the patient's history. However, the text is often temporally incoherent. The narrative goes back and forth in time describing events that happened at different points of time in the past, in the context of current events. For instance, the "history of present illness" section mentions the lack of "chest pain" now, an "episode" that happened 2 days ago, followed by "chest pain" that happened yesterday. Some temporal expressions co-occuring with medical events (highlighted in italics in Figure 1.1) include *2 days ago, now, yesterday, 2 to 3 weeks ago, currently*. The nature of this temporal information is varied and complicated; temporal expressions are often sparse

[1] http://www.healthit.gov



HISTORY PHYSICAL — *DATE: 06/03/2009*
NAME: Smith Bob — MR#: XXX-XX-XXXX
ATTENDING PHYSICIAN: Bill Payne MD — *DOB: 02/28/1960*

CHIEF COMPLAINT
Chest pain and arm infection.
HISTORY OF PRESENT ILLNESS
Patient is a *48-year-old* male with *history of* cocaine use hypertension who presented with chest pain which started *2 days ago* . He does not have chest pain *now* but *ever since* the episode *2 days ago* he has felt a little weaker. He did have chest pain *yesterday* and this is what prompted him to come to the ER. He also notices that he has had some infections under his arms. He states that he had to have an abscess I and D *3 or 4 months ago* under his arm and *2 to 3 weeks ago* he noticed some more spots and these spots have now grown and now are under both arms. *Currently* he is chest pain free.

REVIEW OF SYSTEMS
On exam initial blood pressure was 189/106 *current* blood pressure 148/83 with heart rate of 74 respirations 16. Heart regular rhythm. No murmurs. Arms: He does have tender areas right greater than left under the arm. Difficult to tell if there is any erythema but obvious cellulitis sludge abscess under the right arm which is tender.

ASSESSMENT/PLAN
Currently he is chest pain free. We will check a 2-D echocardiogram. Consult Cardiology for a stress test. Axillary abscesses. Consult Surgery for I and D. We will place on IV vancomycin pain control. Cocaine abuse. Encouraged to quit.

Figure 1.1: Sanitized history and physical report. Underlined phrases are medical events and italicized phrases are temporal expressions.

and vague [Zhou and Hripcsak, 2007]. Moreover, clinical text also exhibits a distinct sub-language abundant with domain-specific terminology, abbreviations and ambiguous terms. For example, *abscess I and D, consult surgery for I and D*. Besides unstructured clinical narratives, the EHR also contains structured and semi-structured data sources like lab reports, problem lists, discharge lists, and encounter lists, medication lists and patient demographics. There tends to be a lot of redendant information embedded across these disparate data sources at different levels of semantic and temporal granularity.

All of these factors, combined with the increasing size of historical patient data in EHRs, makes it difficult for clinicians to reliably search, identify and review specific information related to the patient. This presents opportunities for natural language processing (NLP) to enable unstructured clinical data-analysis and information extraction.

1.2 Clinical Motivation

Clinicians typically formulate hypotheses based initially on the patient's chief complaint and the biomedical context and then modify those conjectures as more information emerges. They evaluate the likelihood, severity, and urgency for treatment of the diseases being considered. Much of this information depends on, among many other factors, when certain medical events took place in the patient's history, their temporal relationship with other medical events, and their recurrence pattern.

Clinical reasoning at its core involves modeling medical knowledge about the patient and making probabilistic inferences. Thus, machine learning provides the appropriate framework to model noisy and uncertain clinical information for temporal reasoning, allowing us to generate a probable timeline of medical events over the patient's history. Clinicians often rely on subjective probabilities or beliefs, also known as heuristics. Heuristics improve cognitive efficiency as clinicians wade through piles of findings. Designing features based on clinical domain specific heuristics helps in training effective in machine learning models for information extraction and temporal reasoning from unstructured clinical narratives. This is important as the clinical sub-language has distinct domain-specific characteristics that may require novel NLP methods for information extraction. Some challenges include dealing with information redundancy within and across unstructured narratives, as well as understanding implicit, explicit and relative temporal cues about when

certain events occurred in the patient's history. Consolidating all of the medical event and temporal information found in clinical narratives, and also attempting to merge this with structured data in the EHR, helps generate a complete and concise picture of the events that occurred in a patients medical history.

Taking all of this into account, we extract semantic and temporal representations of medical events in clinical narratives to help address problems like coreference resolution and temporal reasoning to enable **medical event timeline generation**. Such a timeline could be used to influence physician behavior and improve the quality of health care in general through applications like patient recruitment for clinical trials, information retrieval from bio-repositories, medical document summarization, adverse drug reaction mining among others. We discuss two such applications next.

1.3 Why Extract a Medical Event Timeline from EHRs?

In this section, we describe some applications of extracting and structuring information from clinical data sources in the EHR.

Patient Accrual for Clinical Trials. Clinical trials are research studies that try to answer scientific questions and to find better ways to prevent, diagnose, or treat a disease.[2] Consider a scenario where a clinician needs to answer the following question. "Which patients have a history of another primary malignancy $\leq$ 3 years, with the exception of non-melanoma skin cancer and carcinoma in situ of uterine cervix?" The answer to this question will help decide if the patient is eligible for a particular clinical trial.

The Ohio State University Wexner Medical Center (OSUWMC) has various clinical trials in place including trials for Chronic Lymphocytic Leukemia (CLL). CLL is a type of

[2] http://www.clinicaltrials.gov

cancer in which the bone marrow makes too many lymphocytes. Around 800 CLL patients are treated annually at OSUWMC for CLL.[3] The selection of patients for a clinical trial is general done on a prospective basis . When a new trial for CLL is initiated, any incoming patients suffering from CLL are examined for eligibility. Occasionally, patient selection is also done retrospectively by examining clinical narratives of CLL patients and determining if they match multiple eligibility criteria. However, performing this task manually is extremely challenging as one has to read to multiple of clinical narratives to try and infer the answer. Moreover, the nature of the medical sub-language and implicit temporal cues make the task even more complex and tedious. The process of prospective patient selection is very slow, and also depends on how many patients come to the hospital for treatments or follow-up once the trial has been initiated [Weng et al., 2010]. There are multiple instances of delays in commencing the trial because of shortage of eligible patients. There could also be a section of CLL patients who visit the hospital less frequently, say yearly, and hence never get considered for these trials prospectively [Raghavan and Lai, 2010].

While there have been significant efforts to move to structured data collection, clinical narratives are pieces of clinical documentation used for capturing nuances of a patient's progress that are difficult to capture in a structured manner. Therefore, clinical narratives remain a critical data source for tasks such as the scenario described above. The ability to identify patients for clinical trials automatically, or even reduce the search space, would be of immense value to the clinical research community [Thadani et al., 2009].

Patient cohort selection for clinical trial recruitment needs to examine structured relationships between diseases and diagnoses in the patient's history. This may involve performing temporally conjunctive queries where knowledge of the timeline of a patient's

[3] http://cll.osu.edu

medical history is critical to understanding the progression of diseases and the efficacy of treatments [Jung et al., 2011]. In addition, studies show that temporal constraints are present in 38% of clinical trial eligibility criteria and such constraints can be coarse (*intake of ATT during the past 5 years*) or fine-grained (*antacids for 4 hours before and 4 hours after itraconazole*) [Luo et al., 2011]. Finally, a longitudinal timeline of medical events can often be important to clinical decision making and the practice of evidence based medicine. Physicians often evaluate the likelihood, severity, and urgency for treatment of diseases based on when certain medical events took place in the patient's history, their relationship with other medical events, and their recurrence pattern, among other factors.

Retrieval of Specimens from a Biospecimen Repository. Biospecimen repositories provide long term storage of tissues that can be used for future research. Given that biospoecimen repositories to date are frequently not characterized apart from some very superficial information about the patient from whom the sample was retrieved from, the retrieval of specimens with a particular phenotype is challenging. Also, the labels on the specimen tend to have limited information which makes enabling an automatic search difficult. Thus, augmenting these specimen labels with patient characteristics extracted from the EHR facilitates more specific tissue retrieval from the repository.

Moreover, any clinical application that requires generating a chronological summary of the patient's medical conditions to help clinical decision making would benefit from such a timeline. The timeline can also be mined for temporal patterns that may benefit adverse drug reaction mining. Thus, to enable temporal constraint resolution to help such clinical applications, and keeping in mind the characteristics of clinical narratives, we address a number of problems that help achieve the goal of timeline generation from a patient's

EHR. We briefly describe these problems and the contributions made in addressing these problems in the following sections.

1.4 Dissertation Outline

The natural language processing problems addressed in the generation of such a timeline are captured in Figure 1.2. The contributions made in addressing each of these problems are discussed in Section 1.5. The organization of the rest of the dissertation is described next.

In Chapter 2, we review the background and preliminaries relevant to the topics covered in this dissertation. In doing so, we trace the development of the notion of "events" and "time" in linguistics and philosophy and describe medical events in the context of these notions. We examine prior work in the representation of events, and relationships between events, while noting their relevance to our work.

In Chapter 3, we describe the clinical dataset used throughout the dissertation for various experiments, including the types of clinical narratives and their characteristics, and describe our annotation schema and evaluation metrics.

We then investigate the problem of temporal relation learning by learning to assign medical events to coarsely defined time-bins within each narrative in Chapter 4. The time-bins along with the admission and discharge dates on each note allow us to derive a coarse partial temporal ordering of events across all narratives of a patient. Moreover, these time-bins serve as an important feature in both coreference resolution and fine-grained temporal ordering.

Chapter 5 addresses the problem of medical event coreference resolution. Given multiple mentions of a medical event within and across clinical narratives of a patient, we

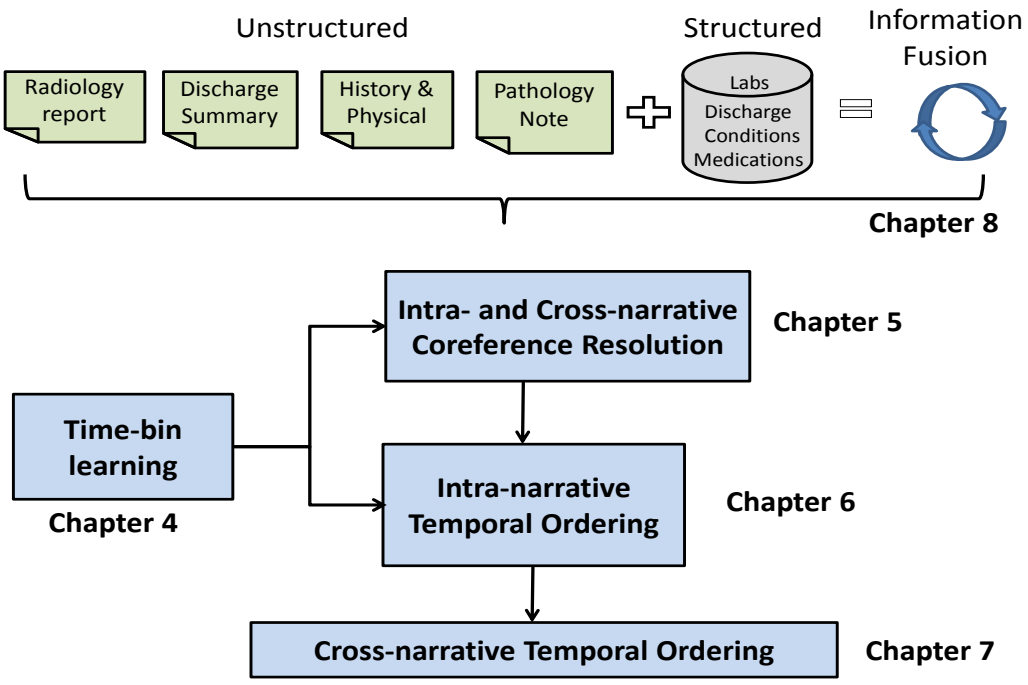


Figure 1.2: Problems addressed in this dissertation. Solving all of these problems helps extract medical events and temporal information from unstructured clinical narratives and generate a timeline over the patient's history. We also describe efforts towards integrating medical events from structured data into the generated timeline.

want to resolve these mentions to the same instance of the medical event. Some example coreference chains include {*heart attack, myocardial infarction, myocardial infarction*}, {*chest pain, episode, chest pain*}, {*abscess, wound*}. Coreference resolution is an important step with respect to our final goal of probabilistic timeline generation as over 20% of the medical events within and 40% of medical events across clinical narratives corefer. We propose training semi-supervised models for the task of medical event coreference resolution by determining semantic and temporal relatedness between medical events. Resolving coreferences not only helps identify unique instances of medical events, but also helps in

identifying additional temporal expressions associated with each mention that may help determine when the medical event occurred.

Next, we address the problem of learning a fine grained temporal ordering of medical events in the same clinical narrative in Chapter 6. In previous work, temporal relation learning has been explored in limited capacity. These include rule-based methods [Zhou et al., 2006] used on discharge summaries, and some classification-based approaches in the CLEF project [Roberts et al., 2008] and the i2b2 tasks [Aramaki et al., 2006]. In this work, we describe a novel ranking method to learn intra-narrative temporal relations. We leverage characteristics of the clinical discourse and narrative structure, including the fact that every narrative always has a temporal grounding in the form of an admission or discharge date (for inpatient notes) or encounter date or date of service (for outpatient notes), for inducing a temporal ordering of events within each narrative. The learned time-bins and coreference information serve as useful features in this process. We demonstrate that this model works better for temporal ordering of medical events within a clinical narrative than the traditionally used pairwise-classification approach by [Mani et al., 2006; Roberts et al., 2008].

Once this task is complete, we have multiple sequences of medical events corresponding to clinical narratives of a patient. Next, we address the problem of cross-narrative temporal ordering in Chapter 7. We learn to combine temporally ordered medical event sequences into a single sequence of temporally ordered medical events across all clinical narratives of a patient. In general, there is little prior work on the problem of cross document temporal relation learning and no prior work in the biomedical domain. Our main contribution here is modeling the problem as a multiple sequence alignment task using a

weighted finite state transducer-based framework to efficiently find the most likely temporally ordered sequence of medical events. We align sequences based on cross-narrative coreference and temporal relation information learned from a corpus of patient narratives. We show that this method performs better than dynamic programming and integer linear programming-based solutions to multiple sequence alignment.

Finally, we investigate the problem of information fusion where we extract medical events along with timestamps from structured notes, such as medication lists and lab tests, in the EHR, as described in Chapter 8. We then learn to map medical events from the structured data to corresponding events in the unstructured data. This mapping across data sources provides additional temporal information for all the learning tasks including intra- and cross-narrative coreference resolution and temporal ordering, thus helping us generate a comprehensive longitudinal account of medical events in the patient's history.

1.5 Dissertation Contributions

The main contributions of our work are described in this section. We break the problem of timeline generation from clinical text into multiple sub-problems, and develop methods to address each of them.

- **Coreference Resolution (Chapter 5).** We approach with the problem of information redundancy within and across longitudinal clinical narratives by addressing the problem of medical event coreference resolution. Although coreference resolution is a well studied problem in computational linguistics [Ng, 2010; Soon et al., 2001], there has been very little research on medical event coreference resolution in clinical text. An important barrier to training supervised models for this task is obtaining expert annotations. Zheng et al. [2012] and the i2b2 challenges [Savova et al.,

2010b] have recently explored the application of supervised learning algorithms to this problem based on various features in clinical text. However, none of the existing research has successfully demonstrated the use of semi-supervised methods for this task. Moreover, in prior work, coreference resolution of medical events has not been modeled on the basis of temporal and semantic similarity.

In this work, we train semi-supervised models for medical event coreference resolution with limited annotated data. We demonstrate that these models perform almost as well as a supervised model on a dataset of clinical narratives. The ability to perform coreference resolution with limited annotations is of immense value to the clinical community and can be used to enable not only applications like temporal reasoning as in our work, but also, search, medical document summarization and cross-document relation learning.

- **Temporal Relation Learning (Chapters 4, 6, 7).** Chronologically ordering medical events in unstructured and temporally incoherent clinical text and generating a comprehensive timeline of medical events across the patient's history has enormous utility in clinical applications with temporal constraints. Our main contributions in addressing the task of timeline generation are in developing novel NLP methods for temporal ordering in clinical text, at both an intra-narrative and cross-narrative level.

  In the case of intra-narrative temporal ordering, we leverage an interval based representation of medical events to enable temporal ordering by ranking events in relative

order of occurrence. We show that the ranking method works better than traditionally used classification-based approaches. Thus, demonstrate the need to rethink resources and methods used for the temporal relation learning task on real-world data like clinical text.

To the best of our knowledge, this is the first work that addresses the problem of cross-narrative reasoning between medical events in clinical text. We enable cross-narrative temporal ordering across all the clinical narratives of a patient with the help of a novel WFST-based approach for multiple sequence alignment. We empirically demonstrate that this method outperforms iterative pairwise dynamic programming, and another state-of-the-art ILP-based method [Do et al., 2012], for the task of multiple sequence alignment. Moreover, the proposed WFST-based framework may be useful in modeling multi-alignments across a variety of domains such as spoken dialog systems and speech.

- **Information Fusion (Chapter 8).** Information is captured in both structured and unstructured formats in the EHR. While most of our work focuses on the NLP of unstructured data for the problem of temporal reasoning, we also explore how merging concepts across structured and unstructured data sources can help better address this problem. Our main contribution here is the development of a temporal model from timestamped structured data to predict the probability of medical event occurrences. This probability can be used as an informative feature in training models for temporal reasoning from unstructured data.

In summary, we propose a novel framework for timeline generation from a patient's EHR, where the framework addresses a number of problems essential to the task of generating a medical event timeline. In the chapters that follow, we describe methodologies to address each of these problems and explore the generated timeline's utility in solving temporal clinical trial eligibility criteria. However, to better motivate and understand these problems, we need understand the background and prior work in representing and reasoning with events and time, and familiarize ourselves with the dataset and annotations used for our experiments. In Chapter 2, we explore the evolution of the definition of "events" in natural language, and the representation of "time," while noting relevant prior work in learning to represent and reason with events in general as well as medical events. We then describe the nature of our dataset including the type of clinical narratives and the annotation schema and attributes in Chapter 3.

CHAPTER 2: A BRIEF HISTORY OF TIME AND EVENTS

Events are woven together in time, and hence time is an intrinsic parameter in the interpretation of event occurrences. An appropriate representation of events, their attributes, and their temporal and semantic relationships is required to enable the automatic learning of relations between events in discourse. To this end, this chapter brings you a historical perspective on events in linguistics. This includes event representation and interpretation, temporal and coreference relations between events, both in the clinical domain and in general. We also review the necessary background for understanding the topics in this dissertation and past work related to these topics. We begin with a discussion about event definition and representation by philosophers and linguists over the years and how this differs from what we consider as a medical event.

2.1 What are *Events*?

The question of event representation raises a deeper question about defining an event. Philosophers and linguists have long debated the meaning of an event and propounded multiple theories of the definition of an event.

In philosophy, events are objects in time or instantiations of properties in objects. In the clinical domain, these objects in time may correspond to medical conditions affecting the patient or treatments given to the patient. The structure of events, including incidents and states, could consist of a host of participants, props, times and locations [Martin and Jurafsky, 2000]. However, in clinical text, the main protagonist of an event is usually the

patient. Further, in defining event representation, we need to ensure all desired inferences can be directly derived from the representation. We first briefly describe our representation of medical events, and then proceed to compare it with event representations used by linguists over the years.

2.1.1 Medical events

We consider medical events as cover terms for mentions that are incidents, activities, states, or entities associated with the patient's medical condition and health care. These include diseases like *mitral valve prolapse, myocardial infarction*, procedures like *surgery* and lab tests like *blood test*, *echocardiogram*, as well as normal health situations like *pregnancy* and *smoking*. Physicians often refer to adverse medical conditions as events, for instance, cardiac events such as *myocardial infarction.* Medical events can be instantaneous, for example, "The patient stated that the *cough* had become bothersome." They could last a period of time, for example, "The patient gives a history of *fever* associated with *chills* for the last 1 month." We also consider as events predicates describing states or circumstances in which something holds true, for example, "The patient had *fever* yesterday." Syntactically, in many studies using clinical text, medical events are restricted to concepts or noun phrases, as seen in the i2b2 [Guo et al., 2006] data annotations. However, we consider noun phrases along with verb phrases (*coughing*), adjectives (***polymicrobial*** *infection*) and event nominals (*intubation*) as medical events. We also cover certain entities that participate in events (patient has *high blood pressure*). In general, any diseases, symptoms, tests, medications and conditions related to the patient's health or healthcare is considered a medical event.

2.1.2 The Evolution of Event Definition in Linguistics

We now study how events are defined and interpreted in literature and their relevance to the notion of a medical event used in various medical reasoning systems including the work described in this dissertation.

The common interpretation of an event in linguistics corresponds to a verbal predicate. meaWe begin with looking at first-order logic (FOL) that uses quantified variables, a domain of discourse over which the quantified variables range, finitely many predicates defined on that domain, and a recursive set of axioms that are believed to hold for those things [Bechhofer et al., 2002].

First-order logic-based representation. FOL is often used for creating semantic representations with events, objects, and properties; for example, a predicate *suffer* with subject patient and object palpitations *suffer(patient, palpitations)*. The advantages of FOL representations for medical events is that they have well-defined syntax and clean mathematical semantics.

Rule-based medical reasoning models like MYCIN [Buchanan and Shortliffe, 1984] have their own logical structure and FOL allows for precise, and compact, representation of these models. However, logic is not the suitable for problems such as medical decision making under uncertainty, and in handling unknown data.

Consider the example rule: $\forall x$ *symptom(x, fever)* $\rightarrow$ *disease(x, pneumonia)*. However, this rule by itself is insufficient as the patient can have several other diseases. We can augment the rules as follows: $\forall x$ *symptom(x, fever)* $\rightarrow$ *disease(x, pneumonia)* $\rightarrow$ *disease(x, influenza)* $\rightarrow$ *disease(x, ear infection)*.

However, due to the overwhelming number of facts in medicine, it is not possible to explicitly represent all those facts as rules. Moreover, first-order representations cannot

handle discourse anaphora and coreference which is an important part of our work in timeline construction.

Another important limitation of FOL is fixed number of arguments or arity. Recognizing that *suffer(patient, palpitations)* and *suffer(patient, palpitations, admission, hospital)* are the same event becomes difficult. In order to overcome this, Davidson [2001] introduced the idea that sentences are indefinite descriptions of eventualities. Thus, he introduced the notion of an "event argument" along with subject and object for each verb. For instance, $\exists e$ *suffer(e, patient, palpitations, admission, hospital)*. This event variable gives us a handle on the event in question, however still leaves issues such as capturing ancillary facts with additional predications (time). For instance, $\exists e$ *suffer(e, patient, palpitations)* $\wedge$ *time(e, admission)* $\wedge$ *location(e, hospital)*.

Neo-Davidsonian representation. Parsons [1990] noted that this can be overcome by distilling the event representation to a single argument that stands for the event itself (known as neo-Davidsonian representation) . For instance, $\exists e$ *suffer(e)* $\wedge$ *sufferer(e, patient)* $\wedge$ *symptom(e,palptitations)* $\wedge$ *time(e,admission)* $\wedge$ *location(e, hospital)*. This eliminates the need for a fixed number of arguments and allows for as many roles and fillers as appear in the input. Now, we can add temporal variables and temporal predicates relating an end point to the current time as indicated by the tense of the verb. However, the relation between verb tenses and points in time is complicated — a present tense verb may be used to refer to a past or future event. Thus, Reichenbach and Reichenbach [1956] introduced the notion of reference time. In representing medical events, we use a similar notion of reference time, usually corresponding to the admission time or date of creation of the narrative to which the event belongs.

Vendler's classification and Bach's eventualities. Based on these related notions, event expressions are traditionally divided into certain classes beginning with Vendler's influential classification of verbs (Figure 2.1) based on temporal properties [Vendler, 1967]. Bach coined the term "eventualities" to include all aspectual types, both stative and eventive [Bach, 1986; Tenny and Pustejovsky, 2000]. Recent work has adopted the use of "event" as the cover term for Bach's eventuality [Tenny and Pustejovsky, 2000]. Approaches within event semantics take eventualities as basic entities in the domain of discourse, along with individuals and times [Dowty, 1986]. Our medical event representation covers the traditional eventualities under the constraint that it has to be a medically relevant term. Since medical events are domain specific concepts that can also be noun phrases, in many instances, they are an argument (object) of the verbal predicate (eventuality). An important point to note is that event expressions can easily be shifted from one class to another. Consider the following examples. (1) "The patient *coughed*." This has no natural endpoint, and hence may be an activity. (2) "The patient *coughed* yesterday." This has a temporal endpoint and may be an accomplishment. Thus, the classification of an event is not just governed by the verb but by the semantics of the entire expression in context.

Dowty's representation and the development of Vendlers representation. Dowty explicitly rejects approaches that try to reduce the differences between different verb classes to purely temporal properties, because they do not capture important lexical semantics of verbs, and hence provide no adequate motivation for the different behavior of the verbal classes [Tenny and Pustejovsky, 2000]. Classifications of verbal predicates into classes that build on the work of Vendler and Dowty have been used for the analyses of a number of grammatical phenomena. Thus, we consider properties of the verbal predicate along with its arguments, as well as semantic, temporal and discourse context, in representing

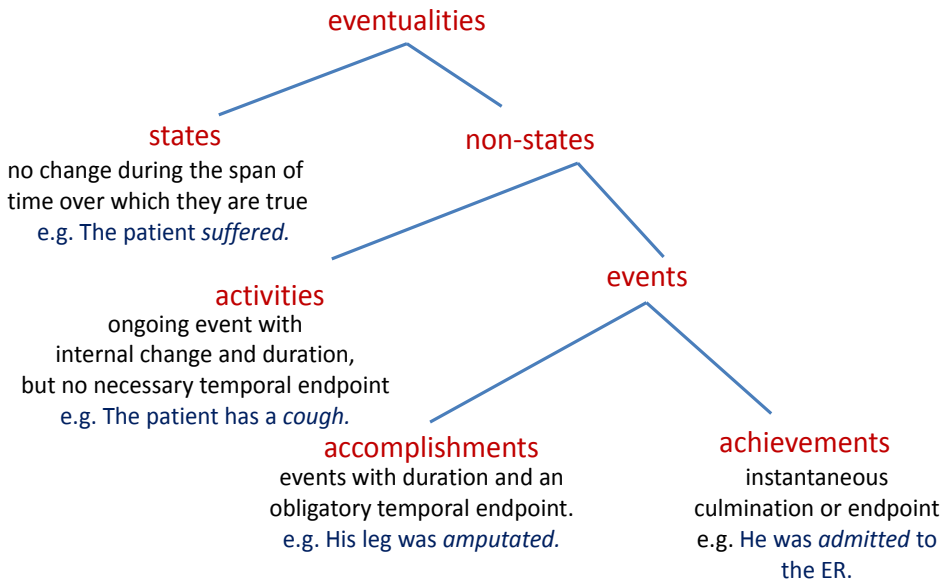


Figure 2.1: Vendler's four categories of verbal predicates [Vendler, 1967].

medical event structure. However, in order to seek explanations for why events fall into these classes requires representing the meaning of arguments with semantic roles and selectional restrictions. We do not include this level of deep representation in our medical event structure.

Following the work of Dowty, Vendlerian classification was further developed within tense logic and event semantics. Tense locates an eventuality in time (past, present, future), whereas aspect distinguishes between events which are ongoing or completed. For example, "The patient *coughed*" and "The patient was *coughing*" are both in past tense, but are aspectually different. Kamp and Reyle [1993] make a distinction between states and non-states, as only non-statives can be used as answers to the question "What happened?." They argue that states and non-states have different temporal consequences. The patient

was *ill* yesterday (state). The patient had a *blood test* yesterday (non-state). Further, lexical meaning can be best captured using different levels of representation including event structure, argument structure, qualia structure and inheritance structure. Thus, event structure is one level of semantic specification.

In the context of our work, we want to be able to answer the following types of questions using the medical event representation: Does the patient have a history of *hypertension*? Was an *echocardiogram* done last week? Did the patient suffer from a *heart attack* 2 days ago? These questions refer to the temporal aspects of the properties of medical events in question. Eventualities give rise to temporal relations, sharing of participants, and coreferential relations.

Coreferential relations among eventualities plays an important role for facilitating access to content and extracting relevant information. Along with eventualities, medical events include domain-specific concepts that are usually noun phrases describing medical conditions; they also could be verb phrases, adjectives or event nominals. Similar to neo-Davidsonian approach described earlier, we can think of medical events as first order individuals, existentially quantified, where participants to the event are conjoined relations between individuals and the event [Parsons, 1990]. We include additional arguments in the event structure to capture temporally related discourse relations, as well as cross-narrative relations. Although the way we interpret the logical medical event representation is similar to neo-Davidsonian theories of event structure, the syntactic details of the representation will vary due to our definition of medical events.

On the one hand, in representing medical events, considering the entire verbal predicate and its arguments may us to include richer grammatical categories such as aspect, telicity,

semantic roles, transitiveness as part of the event structure. Aspect helps in understanding inter-clausal temporal relations, and facilitates reasoning about temporal progression in clinical discourse. It also allows us to reason about the persistence of events. Further, discourse relations like narration, causality, explanation, parallel, are related to the movement of time in discourse. Discourse is coherent if it exhibits these structural relationships between its various segments [Hobbs, 1977]. This can be enabled with a rich event representation relating event and argument structure.

On the other hand, even if we consider verbs co-occurring with medical events, they are not always accurately reflective of the medical event's temporal nature. Moreover, in discharge summaries, almost all medical events or co-occurring verbs are in the past tense (before the discharge date). This is complicated by the clinical sub-language, and the fact that the reference time and medical event with respect to which the tense of the verb is expressed is not always clear. Based on the type of clinical narrative, when it was generated, the reference date for the tense of the verb could be in the patient's history, admission, discharge, or an intermediate date between admission and discharge.

The biggest challenge in considering a rich event representation includes getting annotators with the requisite expertise in linguistics and medicine. This may not be practically feasible as it will further extend the time taken to annotate longitudinal clinical text, leading to increased costs.

Ultimately, in order to provide events with the ability to relate to one another, it would be necessary to enrich them with a structure enabling identification of their temporal characteristics, their context, and their meaning. Such a rich computable structural representation would be immensely beneficial to solving many problems in the clinical research community. Based on this motivation, we develop a medical event representation that is

computable and helps us reason about relationships between medical events in clinical discourse. In developing a representation for medical events, we consider two main problems suggested by Bunt [2007], that are as follows.

1. **Ambiguity**: Interpreting the meaning of natural language expressions requires a lot of context information, such as clinical domain knowledge, how the discourse was generated, and what occurred earlier in the discourse. Thus, handling ambiguity with the help of well defined attributes from the clinical discourse is important in identifying distinct occurrences of events and in learning relationships between those events in clinical text. In the absence of such information, natural language expressions could be very ambiguous.

2. **Robustness**: Linguistic semantic theories are often developed as components of grammatical theories and informed by the analysis of carefully constructed, grammatically perfect sentences. However the clinical sub-language is distinct with domain-specific abbreviations and grammatically irregular sentences. Our medical event structure tries to capture the elements of such sentences, without delving deep into possibly inconsistent grammatical details of medical events across clinical discourse.

We describe in detail the annotation template for medical event representation and relationships between medical events in Chapter 4. As discussed earlier, events give rise to temporal relations, sharing of participants, and coreferential relations. In clinical discourse, often the primary participant in a medical event is the patient. This is of course barring the cases where family history or social history of the patient mentions a relative. We first describe the history of time-based analysis and temporal reasoning in medical systems and review prior work in temporal relation learning in both the clinical domain and

in general. We then explore work in coreference resolution; specifically, we examine the different types of coreferences including coreference between event pairs, and coreference between entity pairs, and describe their relevance to medical event coreference resolution.

2.2 Time and Time Again: The Representation of Time

Effective reasoning with temporal knowledge in natural language requires an appropriate representation of time. Temporal reasoning systems have used either a point-based representation or an interval-based representation of time. Early problem-solving systems represented time as a sequence of time slices, where a set of facts hold true in a particular time-slice [McDermott, 1982]. Before the interval-based representation by Allen [1981], there were was not much prior work on computer representations of time that would enable temporal reasoning in natural language. In the 1970s, Kahn and Gorry [1977] describe a system that maintains temporal relations and provides the rest of the system with the tools to test, retrieve, add, and delete temporal information. Allen's representation varies from that of Kahn and Gorry [1977] as it allows relative temporal relationships to be maintained in a highly structured fashion. It also includes a notion of the present time (i.e., "now"), which is maintained in a manner that does not require knowledge of the exact present time [Allen, 1981].

In general, a point-based representation is desirable if every event is assigned a date. Unfortunately, in real applications, many events cannot be assigned a precise date. In natural language, time references are rather relative and vague. In such cases, time intervals are convenient. References to temporal relations in natural language are often relative, implicit and fuzzy in nature and maybe implicitly introduced by tense and by the description of how events are related to other events. This is true of temporal references in clinical

text as well. An example of a sentence that exhibits an implicit temporal relation, "He is on Dilantin while receiving busulfan." The temporal connective "while" indicates that the time when the medical event regarding administration of *Dilantin* occurred is during the time when he was being receiving *busulfan*. The tense indicates that the medications are being administered at the present time with respect to when the clinical note was written.

As illustrated in Figure 2.2, Allen [1981] models intervals by using endpoints where he assumes a model consisting of a fully ordered set of points of time. An interval is an ordered pair of points with the first point less than the second. The relations that can be defined between the endpoints is shown in Figure 2.2. Such a representation is convenient in modeling medical events as well as it gives us the flexibility of learning relationships between these endpoints in a relative manner. This works well as much of the temporal cues in clinical text are implicit and relative to an anchor date such as admission.

2.3 Temporal Reasoning in Clinical Systems

The ability to reason with time-oriented data is central to the practice of medicine. Monitoring clinical variables over time often provides information that drives medical decision making, including diagnosis and treatment planning.

Back in the 1970s, time-oriented databanks [Fries, 1972] stored and dealt with explicitly timestamped clinical variables. Early expert medical systems such as MYCIN [Buchanan and Shortliffe, 1984] and ONCOCIN [Shortliffe et al., 1981] recognized the need to associate time information with clinical data. However, such systems were based on a fixed set of rules, which were in many cases disease specific, and hence had limited ability to exploit the temporal nature of clinical data for reasoning. In most cases these systems dealt with

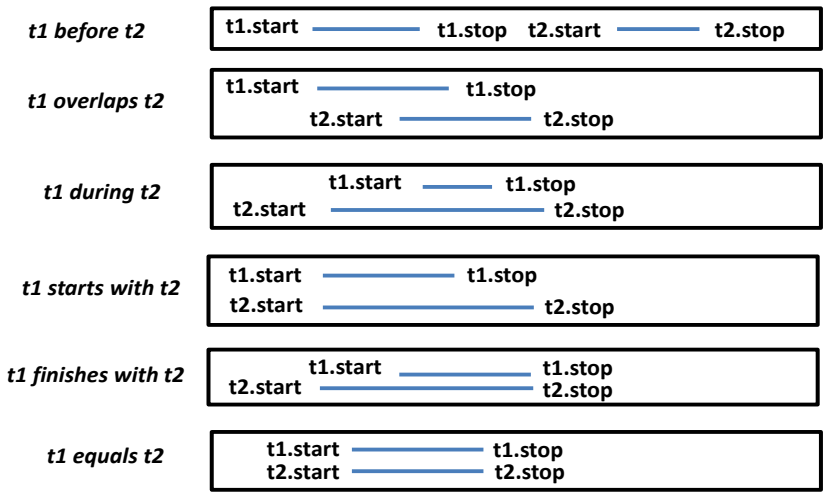


Figure 2.2: Allen's temporal representation using endpoints. Ordering endpoints of time allows learning of temporal relationships {before, after, overlaps, equals, during, finishes with, starts with} [Allen, 1981]

limited categorical clinical variables such as WBC count, bilirubin levels, blood pressure readings etc.

Further, in the 1970s and 1980s, there were some important studies like the Linguistic String Project (LSP) at New York University [Grishman et al., 1973]. LSP was one of the first NLP systems for English and was eventually adapted to medical text. This was developed as part of the time program that takes the output of the NLP system, recognizes and analyzes time information, and formalizes the variant time expressions in a representation which consists of the following fields: relation, reference point, direction, quantity and time-unit. It also describes a representation for medical events which can be used to form directed graph. The vertices of such a graph correspond to points in time, and the directed

edges to the intervals of time between the points. The output of the time program can be used to answer time-related queries.

Story and Hirschman [1982] identified words and structures that carry time information in medical data and emphasized that the identification should embrace explicit temporal information like verb tense, adverbial expressions of time, as well as implicit information like multiple references to the same event in a text, and narrative time progression. Obermeier [1986] developed a system called "grammatical representation of objective knowledge" for analyzing temporal information in medical text. The authors defined "key events" as domain-specific concepts which is used to order and group events. The 1990s saw a lot of research in temporal reasoning and temporal data management, but most work focused on structured medical records stored in databases [Böhlen, 1995; Pedersen and Jensen, 1998; Shahar, 1999].

More recently, the 21st century has seen considerable research in processing time in natural language in many domains. Modern statistical approaches and advances in natural language processing now also enable representations and learning over large-scale patient data in the biomedical domain. However, privacy concerns do not allow easy creation of publicly accessible clinical corpora. The difficult nature of the medical sub-language also makes it difficult to annotate clinical text with temporal information. In spite of these hurdles, the past decade has seen a lot of research in processing time in clinical text using rule-based methods as well as statistical machine learning.

Zhou and Hripcsak [2007] survey work done by researchers in medical decision support systems, where some systems model time implicitly, whereas others do it explicitly. An example of an implicit temporal statement includes "significant weight loss during last year." On the other hand, explicit time modeling has a model of time in which various

factors are associated to the model of time and uses the association of entities to the time model to draw inferences.

Although there has been a push towards capturing clinical information as structured data, physicians have a tendency to record patient information in the "free text" input boxes of the EHR. Often, narratives are also transcribed from notes dictated by physicians and only available in unstructured format. Problems in processing unstructured clinical data include extracting precise and contextually relevant medical concepts, for e.g. diseases, symptoms, tests, findings and medications. Automatically encoding the extracted medical concepts using medical terminologies is important for various clinical applications including billing systems. Further, extraction of temporal and semantic relationships between medical concepts from the unstructured data is challenging. Various medical data understanding systems like "The Special Purpose Radiology Understanding System" and "Symbolic Text Processor" have been used to processes specific clinical datasets such as radiology reports, but they do not capture explicit temporal information. However, they do identify a change of state. Possible values for states include {unchanged, improved, recurrence, worsened}. One limitation of most medical reasoning systems is that they tend to ignore referential relations between discourse units [Zhou and Hripcsak, 2007]. Recently developed medical NLP systems use a conceptual representation structure to address anaphoric reference relations spanning sentences [Savova et al., 2010a].

More recently, Zhou et al. [2006] have modeled the temporal information contained in clinical discharge summaries as a Simple Temporal Problem (STP). Based upon this work, they further proposed an architecture for representing, extracting and reasoning about temporal information in clinical narrative reports, and incorporate this as part of the MedLEE NLP system [Chiang et al., 2010].

In the last few years, a shared task has been defined called the i2b2 challenge.[4] Their challenges include datasets for coreference resolution and temporal relation learning. Researchers have used these datasets to extract a limited set of temporal relations using rule-based and machine learning methods [Sun et al., 2013].

2.3.1 Learning Temporal Relations using TimeML and Timebank

Until around the year 2003, there were no shared resources or prior work on automatically learning temporal relations from unstructured text. Most of the prior work in this area focused on temporal data management using temporal databases [Jensen and Snodgrass, 1999]. Since then, there have been efforts to develop a shared corpus of newswire text annotated with event and time information. An annotation scheme that was developed by Pustejovsky et al. [2003a], called TimeML, was used to annotate events, times, and their temporal relations in newswire text and develop the Timebank corpus [Pustejovsky et al., 2003b]. Tensed verbs, adjectives, or nominals are considered as "events " by the TimeML scheme. Event attributes include tense, grammatical aspect, polarity (negative or positive), modal operators which govern the event being tagged, and cardinality of the event if its mentioned more than once. Temporal expressions are annotated based on TIMEX scheme, defined as part of TimeML, and temporal relations tagging events to other events and/or times are annotated using the TLINK tag. An example simplified annotated sentence from Timebank is shown below.

- *On the other hand, it's*

  <EVENT eid="e1" class="OCCURRENCE"> ***turning*** </EVENT>

  out to be another very

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

<EVENT eid="e369" class="STATE"> ***bad*** </EVENT> *financial*

<TIMEX3 tid="t83" type="DURATION" functionInDoc="NONE"> ***week***

</TIMEX3>

In the above sentence, the annotation identifies an event *turning*, with ID e1, and assigns it a class of type "occurrence" indicating something that happens or occurs in the world. Similarly, it also identifies *bad* as an event with ID e369, and *week* as a temporal expression with ID t83 and of type "duration." The temporal expression annotation has certain attributes like functionInDoc that indicates whether the temporal expression provides a temporal anchor for other temporal expressions in the document. The event *turning* is linked to temporal expression *week* with the temporal relation "simultaneous" as indicated in the example TLINK annotation below.

- <TLINK lid="l52" relType="SIMULTANEOUS" eventID="e369"
  relatedToTime="t83"/>

Similarly, events may also be related to other events via a temporal relation. TimeML uses 14 temporal relations between event-event pairs and event-time pairs, which reduce to a disjunctive classification of 6 temporal relations {SIMULTANEOUS, IBEFORE, BEFORE, BEGINS, ENDS, INCLUDES}. These temporal relations correspond to Allen's interval-based representation for temporal relations [Allen, 1981].

The Timebank corpus has evolved as a community resource for temporal relation learning in the NLP community [Boguraev et al., 2007]. The TempEval framework was created for evaluating systems that automatically annotate texts with temporal relations using the TimeML format Verhagen et al. [2009]. The TempEval challenges included creating systems for temporal expression extraction, identifying temporal relations between a set of

events and all time expressions appearing in the same sentence, identifying temporal relations between events and the document creation time and identifying the temporal relations between contiguous pairs of matrix verbs. HeidelTime [Strötgen and Gertz, 2010], a rule based system for extraction and normalization of temporal expressions achieved an F-score of 86%, for extraction in TempEval-2. Systems developed for the TempEval challenge also include the rule-based system by Kolya et al. [2010], TRIPS and TRIOS [UzZaman and Allen, 2010], among others [Lee and Katz, 2009]. While the TempEval tasks are mostly restricted to a set of three temporal relations {*before, after, overlaps*}, there are various other researchers who have tried to learn all of Allen's temporal relations between event pairs.

Mani et al. [2006] train a Maximum Entropy classifier to classify event pairs into one of the 6 Allen's temporal relations. They utilize the hand-tagged features in Timebank including tense, aspect, modality, polarity and event class for learning. They also expand the training dataset by doing a transitive closure on temporal relations in the corpus and report improved results. Chambers and Jurafsky [2008] report that they were unable to replicate this performance on Timebank. Instead they propose a two stage architecture that first learns the hand-tagged attributes of events in Timebank. These attributes include aspect, tense, aspectual class, and then use the learned features along with other linguistic features to classify temporal relations between event pairs. Using this two-stage architecture, they report a 3% improvement over results reported by Mani et al. [2006]. Lapata and Lascarides [2006] trained a classifier based on syntax and clausal ordering features to learn inter-sentential events.

However, since the nature of language and events occurring in the new domain is very different from clinical language used in the EHR, the Timebank corpus cannot be directly

used for learning temporal relations between medical events in clinical text. In our work, we leverage the advances in machine learning for natural language processing to train models of information extraction from clinical narratives based on clinical domain heuristics. More specifically, we propose an end-to-end system which will help generate a probabilistic timeline of medical events from within and across unstructured longitudinal clinical narratives. In this system, we use an event representation (described in Chapter 3) and enable temporal ordering of medical events within and across clinical narratives. Another important component in this system is performing medical event coreference resolution. We explore some relevant prior work in coreference resolution next.

2.4 Coreference Resolution

Coreference is defined as when two or more expressions in a text have the same referent, i.e., they refer to the same person or thing. Coreference resolution is a well-studied problem in discourse analysis and is considered a difficult natural language processing task, typically involving the use of sophisticated knowledge sources and inference procedures [Charniak, 1972].

Different types of coreference occur in natural language including anaphora, noun phrase or entity coreference and event coreference. Anaphora resolution is the problem of trying to identify an antecedent for an anaphoric, where an anaphoric is a noun phrase that depends on the antecedent. On the other hand, noun phrase coreference resolution, the task of determining which noun phrases in a text refer to the same real-world entity. The ACE terminology [Doddington et al., 2004] defines an entity as an object or a set of objects in the world, for instance, person, place, or organization. While anaphora and entity coreference resolution have been widely studied on standard NLP corpora, there is limited

prior work on event coreference resolution. In general accordance with event definitions and interpretations described in Section 2.1, ACE defines an event in natural language as a specific occurrence involving participants, where the event trigger or mention is the word that most clearly expresses an event's occurrence. An event could have attributes such as type, polarity, modality, genericity, and tense. Events are typically verbs such as *tore* and *exploded*, although event nominals may also be considered as events. Event coreference resolution addresses the task of grouping all the mentions of events in a document into equivalence classes so that all the mentions in a given class refer to a unified event [Chen and Ji, 2010].

2.4.1 Entity coreference resolution

Entity or noun phrase coreference resolution is a well studied problem with many successful techniques proposed over the years [Strube and Ponzetto, 2006; Haghighi and Klein, 2010; Raghunathan et al., 2010a]. Clustering, as well as pairwise coreference resolution models, have been explored by different researchers to match entity pairs using various syntactic and semantic features. The WordNet[5] semantic class feature is widely used for coreference resolution, although it is known to have limited coverage. Bean and Riloff [2004] propose a semi-supervised method to extract case frames from large corpora. They then use case frames to represent the contextual role of NPs, where a case frame is a frequent pattern surrounding an NP. The intuition is that NPs frequently appearing in the same case frame are likely to be semantically related or equivalent.

Machine learning approaches to noun phrase coreference resolution are described in detail in the survey paper by Ng [2010]. The problem is usually cast as a classification task where a pair of noun phrases is classified as coreferring or not based on constraints

[5] http://wordnet.princeton.edu

that are learned from an annotated corpus. A separate clustering mechanism identifies possibly contradictory pairwise classifications and constructs a partition on the set of NPs. Soon et al. [2001] applied an NP coreference system based on decision tree induction to two standard coreference resolution data sets [Van Deemter and Kibble, 2000], achieving performance comparable to the best-performing knowledge-based coreference engines. A popular off-the-shelf coreference resolution system is the Stanford NLP group's coreference resolution system.[6] This was the top ranked system at the CoNLL-2011 shared task. The system is based on work by Raghunathan et al. [2010b] who propose a coreference architecture called *seives*, where tiers of coreference models at different levels of precision are applied one after the other. This clustering-based model propagates global information by sharing attributes across mentions in the same cluster. This method outperforms many state-of-the-art unsupervised and supervised coreference resolution methods on standard corpora.

More recently, Do et al. [2013] have proposed methods for event detection and coreference resolution from newswire text. Durrett and Klein [2013] use hand-crafted, shallow, heterogeneous semantic, syntactic and discourse features in training a coreference model that outperforms both the Stanford coreference resolution system and the IMS system [Björkelund et al., 2013] which was the best performing English coreference resolution system.

2.4.2 Event coreference resolution

Compared to the extensive work on entity coreference, the related problem of event coreference remains relatively under-explored, with minimal work on how entity and event

[6] http://nlp.stanford.edu/software/dcoref.shtml

coreference can be considered jointly on an open domain [Van Deemter and Kibble, 2000; Bagga and Baldwin, 1998; Humphreys et al., 1997; Haghighi and Klein, 2010].

The problem of determining if two events are identical was originally studied in philosophy. As discussed earlier, Davidson [2001] argued that two events are identical if they have the same cause and effect. However, he abandoned this model and adopted Quinean theory on event identity [Malpas, 1992], where each event refers to a physical object that is well defined in space and time, and two events are identical if they have the same spatio-temporal location. Based on a similar notion, we consider two medical event as identical if the events correspond to the same occurrence with the same spatio-temporal location. However, syntactically the notion of a medical event is quite different from what is considered an event in linguistics. Noun phrases, verbs, nominals, adverbs could all be medical events.

2.4.3 Medical event coreference resolution

Previous work in coreference resolution of medical entities performs coreference resolution on hospital discharge summaries by treating coreference resolution as a binary classification problem [He, 2007]. The author investigates critical features for coreference resolution for entities that fall into five medical semantic categories that commonly appear in discharge summaries.

The Unified Medical Language System (UMLS) [Bodenreider, 2004] knowledge sources include a large Metathesaurus of concepts and terms from many biomedical vocabularies and classifications; a Semantic Network of sensible relationships among the broad semantic types or categories to which all Metathesaurus concepts are assigned; and a lexicon which contains syntactic, morphological, and orthographic information for biomedical and

common words in the English language. The lexicon and its associated lexical resources are used to generate the indexes to the Metathesaurus and also have wide applicability in natural language processing applications in the biomedical domain.

Coreference resolution algorithms that are part of open source NLP tools like the Stanford Coreference Resolution System do not account for domain specific temporal attributes of clinical text. The MUC [Van Deemter and Kibble, 2000] and Automatic Content Extraction (ACE) [Doddington et al., 2004] corpora have been extensively used for training and testing coreference models. These corpora are mostly homogeneous consisting of documents only from the newswire domain.

To our knowledge, the only previous work that considered entity and event coreference resolution in the clinical domain is by He [2007], but limited to the medical domain and focused on just five semantic categories. We consider medical event coreferences where events are considered as coreferring if they are semantically and spatio-temporally similar. In other words, events that mean the same, resolve to the same occurrence and the same time point or duration, are considered as coreferring. In sync with this definition, medical event pairs that corefer may correspond to examples like {*chest pain, pain*}, {*episode, chest pain*}, {*heart attack, myocardial infarction*}, {*tumor, cancer*} and {*hypertension, hypertension*}. We explore the application of semi-supervised methods to the problem of pairwise medical event coreference resolution and show promising results.

Another important component of our timeline generation system tries to integrate information across structured and unstructured data sources. This allows us to use timestamped structured data to help learn from unstructured data.

2.5 Information Fusion

Integrating information across disparate data sources is an important knowledge extraction task. It helps generate a combined representation that may provide a better understanding of the underlying data. There are studies in different domains for fusing different data sources such as integrating genomic data into the EHR [Kho et al., 2013], combining cross-lingual resources for word alignment [Souza et al., 2013] and using structured and semi-structured data sources for question answering [Kalyanpur et al., 2012]. Integration across different data soucres including structured and ustructured data is essential to these applications. However, information fusion across data sources in the EHR is a relatively unexplored problem. We explore how structured information from the EHR can be used to better enable the process of timeline generation from unstructured clinical narratives. We observe a significant improvement in the accuracy of the timeline generated fron the unstructured data after using the process of information fusion.

In this chapter, we first traced the representation of events and time in linguistics and philsophy and noted its relevance in the context of medical events. We then discussed temporal relation learning using the Timebank and in the clinical domain, the types of coreference problems in natural language including that of medical event coreference resolution and finally information fusion. Next, we describe our dataset, annotations, and compare the schema to TimeML, in Chapter 3.

CHAPTER 3: ANNOTATING CLINICAL NARRATIVES FOR COREFERENCE RESOLUTION AND TEMPORAL REASONING

Addressing the natural language processing problems of coreference and temporal relation learning using machine learning methods requires annotated data for training models and evaluating results. This Chapter first defines a medical event, describes the annotation schema and inter-annotator agreement for events and their attributes.[7] Finally, the evaluation metric for the generated medical event timeline is outlined.

3.1 Introduction

As discussed in the previous chapter, the TimeBank corpus annotated using the TimeML specification, is a community shared resource for temporal relation learning in newswire text [Pustejovsky et al., 2003a]. In the medical domain, Zhou et al. [2006] defines annotations for temporal expressions found in discharge summaries. Savova et al. [2010a] propose how they are going to work towards temporal relation discovery with the long term goal of integrating temporal reasoning into the medical NLP system, cTAKES. The authors observe how off the shelf parsers don't work well with medical data as most of the parsers are trained on the Wall Street Journal, establishing the need for domain-specific corpora. They note, "For example, rash is typically an adjective in newswire but is a noun in clinical notes; erythema is not identified as a noun....." Although the authors propose using TimeML for

[7]Parts of this work have been published in AMIA 2012. P. Raghavan, E. Fosler-Lussier, and A. Lai, "Inter-Annotator Reliability of Medical Events, Coreferences and Temporal Relations in Clinical Narratives by Annotators with Varying Levels of Clinical Expertise," AMIA Annual Symposium, 2012.

tagging clinical narratives, they do not explain why or how these tags are the right choice. None of the annotation formats fully capture the requirements for temporal reasoning and coreference resolution within and across clinical narratives of a patient. We define an annotation format that supports these tasks and in turn enables creation of a longitudinal record of events over a patient's medical history. The proposed specification extends the TimeML [Pustejovsky et al., 2003a] annotation format to accommodate attributes specific to clinical text. The annotation specification that will help identify medical events, temporal expressions and temporal and coreference relations between medical events.

However, creating annotated clinical corpora with such detailed relationships is tedious, expensive and requires experts with domain knowledge. Many clinical narrative annotation efforts have used physicians, which can potentially be cost-prohibitive. Within this population, it can be difficult to find individuals willing to devote the time and effort to doing manual annotations. Thus, our annotation effort uses annotators that are current students or graduates from diverse clinical backgrounds with varying levels of clinical experience. We demonstrate that in spite of this diversity, the annotation agreement across the team of annotators is reasonably high.

3.2 Contributions

Our main contribution is defining a new tag and attributes for medical events. This is motivated by the observation that medical events do not have properties similar to events described in TimeML. Further, attributes like tense and aspect are not always useful for learning temporal relations between medical events since most clinical narratives are written in the past tense. Moreover, since our primary interest is in ordering medical events, there is a need to define special attributes for medical events including their temporal rank

as well as any coreference relations. The main contributions described in this chapter are as follows:

- We define an annotation schema with annotation elements and attributes defined for coreference resolution and temporal ordering of medical events.
- Measurement of inter-annotator reliability by determining inconsistencies across annotators by various annotators at various levels. We measure the consistency of identifying a word or phrase as a medical event and applying the same concept code to the event. We also measure inconsistencies in noting that medical events corefer and in noting temporal relations between events.
- Demonstration of high agreement across annotators with diverse clinical expertise. In this study, the annotators were current students and recent graduates from diverse medical and nursing backgrounds with varying levels of clinical experience. In spite of this diversity, we demonstrate that the annotation consistency across the team of annotators is high. We also describe the patterns of agreement between annotators from these different backgrounds.

3.3 Motivation

Timebank is the standard, widely used corpus for temporal relation learning in the computational linguistics community. However, 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. Medical events need not be verbs and are in fact in most cases noun phrases. In contrast, events in the news domain denote change in state, and are usually verbs. Thus, the temporal properties of medical events and events in the news domain are

MRN:12432 Inpatient Progress Note Name: Jack Payne
Date:04/02/2011

Today patient alert and spoke 1-2 word answers to xxxs.
Denied c/o.

Tm 100.8 ax HR 77 BP 141/90 RR 27 94% RA.

Heart RRR

Lungs clear anterior chest Abdomen +BS soft NT Assessment
and Plan: Extremities no edema

Fever: U/A clear BC no growth so far.

No obvious source except ?

sacral decubitus (present on admission).

MRI for xxx osteomyelitis pending.

Wound care nurse to see Monday.

Figure 3.1: Excerpt from a sanitized progress note.

quite different. For instance, tense and aspect are often used to temporally order events in Timebank. However, since many medical events are noun phrases, and most narratives are written in the past tense, tense becomes a less informative feature in clinical text. Further, differences in granularity of temporal expressions, and ways of expressing time with domain specific terminology adds to the difficulty in using the Timebank corpus for the clinical domain. Timebank annotations are also restricted to a single document. We intend to study cross narrative relationships between medical events. Finally, information about medical event coreferences is critical to many of our proposed tasks, especially when trying to reason with medical events across narratives. Such coreference information is not available in Timebank, and even if it were available, the differences between events in the news and clinical domain would deter the use of Timebank. Thus, to strengthen the motivation for generating a corpus of clinical narratives, we now examine the nature of text in a progress note (Figure 3.1) and a discharge summary (Figure 3.2).

MRN:12432 Discharge Summary Name: Jack Payne
Admission Date:03/29/2011 Discharge Date:04/10/2011

HISTORY OF PRESENT ILLNESS
Mr. Payne came in with a problem of a leaking G-tube.
HEENT anicteric The G-tube had been in place since 2008.
HOSPITAL COURSE
The G-tube was surgically removed at laparotomy
(not cooperative with deep breaths) with closure of the
stomach. After approximately 2 days of nothing by mouth
status we started bringing him around.
It took some time for his diet to be appreciated and he was
returned to an ECF (Extended care facility) with a regular
diet and wound care

Figure 3.2: Excerpt from a sanitized discharge summary.

Characteristics of Clinical Narratives: In order to automatically process the progress note, we would require knowledge of what each of the abbreviations and symbols mean. The note is written with the expectation that the reader (usually a physician or a nurse) has implicit knowledge about what is being described. Along with omitted implicit information, we can also observe short phrases, incomplete sentences and extensive domain specific terminology. Temporal expressions in this note include "to see on Monday," "Today patient alert." Let us also examine an excerpt from a sample discharge summary seen in Figure 3.2. As seen in the figure, the content in a discharge summary is usually grouped into different sections that describe the history of illness, hospital course, and plan after discharge. Discharge summaries tend to be more verbose, with mostly complete sentences. However, there is again considerable use of medical terminology, and multiple temporal references. Similarly, other types of clinical narratives also exhibit such characteristics to varying degrees, making it difficult to automatically extract information from them. In or-

der to perform supervised machine learning to infer a relative temporal order among events, there is a need for annotated data. Instances of events and temporal expressions need to be annotated to facilitate creation of features for machine learning of temporal relations in clinical text. Thus, we define an annotation specification that supports the following tasks:

- Identify medical events and temporal relations between medical events that occur in clinical narratives. For instance, chest pain *before* echocardiogram, echocardiogram *before* diastolic dysfunction, mitral valve prolapse *simultaneous* diastolic dysfunction.

- Identify overlapping mentions of medical events within and across all clinical narratives for a patient, and resolve medical event coreferences. Resolving medical event co-references requires identifying which medical events belong to the similar semantic categories, comparing when the medical events occurred, i.e., establishing, if possible, an overlap in time frame of the occurrence of the medical events.

- Learn temporal relations between medical events that occur within and across all clinical narratives for a patient. The temporal relations could be at different levels of granularity.

- Temporal reasoning to infer new temporal relations between medical events.

- Learn the temporal order of medical events across all clinical narratives of a patient.

The tasks described above support the final goal of being able to generate a chronology of unique medical events from across all clinical narratives for each patient. The corpus used for annotation and training our learning models is described next.

3.4 Annotating Clinical Narratives

We now describe our annotation specification for clinical narratives, while describing how much these annotations are similar to and differ from TimeML. To help understand the use of the annotations and their associated attributes, we discuss the apply relevant tags to the following example sentence: "The patient was admitted to the hospital 2 days after chest pain stopped."

3.4.1 Annotating Medical Events

As discussed in Chapter 2, Section 2.1.1, medical events include diseases and disorders, normal health situations like pregnancy that may affect the patient's health, as well as any treatments, procedures and drugs administered to the patient. Thus, the annotators were instructured to annotate any word or contiguous group of words found in a patient narrative that has a contextually relevant match in the UMLS as a medical event. These include the following.

- Any disease or disorder. These include medical conditions which are typically nouns. For example, heart attack, chest pain, hypothermia and *diastolic dysfunction.*
- Any treatment, test or procedure. These are again generally nouns. For example, echo-cardiogram and *cholesterol profile.*
- Any drugs administered. These are typically nouns which are names of drugs such as beta blockers and *niacin.*
- Any normal health condition that requires health care such as pregnancy.
- Any observations related to the patient that may affect his health care. These could be nouns or verbs based on context. For instance, smoking and *drug abuse.*

The general notion of an event in TimeML are that events are used as a cover term for situations that happen or occur. In the clinical domain, a medical event describes the patient's state, at a particular time point or time duration, as seen from a medical standpoint. In TimeML, events may be expressed by means of tensed or untensed verbs (example 1 and 2), nominalizations (example 3) , adjectives (example 4), or prepositional phrases (example 5). Some examples of such events in the medical domain are as follows[8]:

(1) The patient was *screened* for hepatitis.

(2) He was advised *to exercise*.

(3) She was *asked* to take a blood test.

(4) The patients cholesterol levels were *high*.

(5) She gets some dyspnea *on exertion* when she walks.

However, from the perspective of knowledge extraction for helping clinical applications, we are interested in medical concepts like *hepatitis, blood test, cholesterol levels*, and *dyspnea* in these sentences. These concepts correspond to the patient suffering from medical conditions like hepatitis or dyspnea, the cholesterol levels of the patient or a blood test being administered to the patient. These are referred to as medical events in the clinical community as they correspond to an occurrence of something the patient suffers from or is being treated with.

Events are semantically grouped into various classes in TimeML. A subset of classes that may occur in clinical narratives, include the following:

[8]*Note: Not all possible events have been tagged in the above examples.*

- Perception: This class represents physical perception of a medical event. E.g. see, watch, noted, listen
- Occurrence: This class represents something that happens or occurs in the world. E.g. loss of blood, suffered from a stroke
- Reporting: This class represents narration of an event, declaration. E.g. the patient / physician reported / explained

We expect perception and occurrence to be the most common classes occurring in clinical text.

While medical events could be of the general events class "occurrence", they are often noun phrases or nominals. Thus, we create a new tag called "Medical", with its own attributes. **The Medical Event Tag:** A medical event is a word or a contiguous group of words found in a clinical narrative that describes a medical condition affecting the patient's health. Some examples of medical events include *stroke, myocardial infarction, niacin, beta blockers* and *smoking*.

In TimeML, an event has one or more instances with certain attributes. These include attributes such as tense and aspect. However, most of the event instance attributes as defined by TimeML are not applicable to medical events. Let us now examine why this is the case.

Tense places temporal references along a conceptual time line. On the other hand, aspect encodes how a situation or action occurs in time. These features are useful as they help place a situation in time. Since tense is identified by inflections of the verb, it is possible to assign values to the tense attribute for events as defined in TimeML. This assignment is not possible for "medical events." This is because medical events are typically names of illnesses, treatments, tests and other medical conditions. With respect to the temporal

relation learning problem, in the example sentences above, we are interested in understanding when was the patient screened for hepatitis, when blood test was taken, when were the patient's cholesterol levels high, and when did the patient suffer from dyspnea. The answer to "when" needs to be answered relative to other medical events in the text. So our objective is to relate these medical entities with respect to time. The tense of generic events if appropriately associated with medical events occurring in the same sentence, this could help determine the relative time of occurrence of those medical events. However, a notable characteristic of clinical narratives is that they are usually written to capture what happened to the patient in his history, during the day, or during his hospital stay. Thus, most of the described events tend to be in the past tense. Therefore, tense of an event may not always be an informative feature in clinical text.

Taking all of this into consideration, we introduce new attributes for instances of tag "Medical." These attributes are specific to medical events and are useful features for temporal reasoning and co-reference resolution. While we annotate medical events with the newly introduced attributes, we would also like to annotate other event classes and use them wherever applicable. They may be useful, in combination with temporal signals, for standard medical events like admission and discharge (e.g. "after admission," "before admission," "on getting admitted," "during admission"). Details on how to tag event instances are presented in the next section. Similar to the TimeML format, we define the Backus-Naur Form (BNF) for the medical event tag. We begin by describing the BNF for the TimeML event tag.

BNF for Events: The BNF for events is shown in Figure 3.3. The document ID (docid) is the clinical narrative name followed by an integer indexing clinical narratives of a

attributes ::= docid mid class
docid :: = patient narrative name<integer>
mid ::= e<integer>
class ::= REPORTING | OCCURRENCE | PERCEPTION

Figure 3.3: Basic BNF for events

particular type for a given patient. For instance, a patient with could have Radiology1, Radiology2, Radiology3, HistoryPhysical1, DischargeSummary1, DischargeSummary2 etc. The medical event ID (mid) is the unique ID assigned to every event in a clinical narrative. A single patient could have multiple clinical narratives from radiology reports and progress notes to discharge summaries, either all documenting the progress of a particular medical condition or for various ailments over his medical history. Thus the combination of docid and mid helps uniquely identify an event in the patients history.

As mentioned earlier, class refers to the semantic categorization of a medical event. The events are highlighted in our example sentence, "The patient was admitted to the hospital 2 days after chest pain stopped."

- <EVENT docid=ds2 mid=me5 class=OCCURRENCE>*admitted*</EVENT> *chest pain*
- <EVENT docid=ds2 mid=me7 class=OCCURRENCE>*stopped*</EVENT>

Here, ds stands for discharge summary. Thus ds2 represents the second discharge summary of a patient. *Chest pain* is a medical event that has properties different from what is considered to be a TimeML event. Thus, we define a medical event tag which has additional attributes that are as follows:

- Semantic Type (semtype): Semantic type corresponds to the UMLS semantic type.

- Concept ID (cui) and Semantic Type: The CUI of a medical event represents the UMLS concept ID that is the closest contextual match for the medical event we are annotating.

- Polarity: The polarity of a medical event indicates whether the event actually occurred (positive polarity) or did not occur (negative polarity) as understood from the clinical text.

- Signal: This includes any function words that indicate a temporal relationship like *when, in, before* and *after*.

- Medical event coreferences (coref): The attribute coref helps track overlapping mentions of the same medical event within and across the clinical narratives for a particular patient.

- Time bins: The coarse time period in which the medical event occurred is referred to as a time-bin. We define the following time-bins associated with a medical event: *before admission (beforeadm), on admission (onadm), after admission (afteradm)* and *after discharge (afterdis)*. Admission and discharge are two events that almost always occur in every clinical narrative. Medical events could have occurred before the patient was admitted or during his stay in the hospital (in case of an in-patient narrative). Similarly, a medical event could have occurred before or after the patient is discharged from hospital. Knowledge of when a medical event occurred relative to the admission or discharge date could be leveraged for temporal inference. Hence, we have attributes beforeadm, onadm, afteradm, and afterdis, which take on boolean values indicating whether a medical event happened before or after admission and before or after discharge.

- Temporal expression (*tempex*): The ID of the temporal expression that is anchored to the medical event.

- Temporal anchors (*start* and *finish*): The *start* and *finish* attributes indicate the time when the medical event started and finished by associating corresponding temporal expressions IDs. For events where the duration is unclear or where the events are an occurrence at a specific time point, both the start and finish take the same value.

- Rank (*localrank* and *globalrank*): The local indicates the temporal rank of the medical event within the clinical narrative and global rank indicates the temporal rank of the medical event across all clinical narratives of the patient.

As the annotator reads through the patient narrative and encounters medical event occurrences that are the same as previously seen medical events, they get added to the coref attribute. This is determined by semantic and spatiotemporal similarity between two events. Thus, the coref attribute helps multiple occurrences of the same medical event in the text. Taking into account these additional attributes for medical events, the updated BNF for the medical event tag would be as shown in Figure 3.4.

Significance of Medical Event Coreference Attribute: The "coref" attribute is of significance because clinical narratives are written by clinicians at various points of the patient's hospital stay. History and Physical and Social Work Assessment reports are usually written during admission whereas Progress Notes are written for each day of the hospital stay. On the other hand, discharge summaries are written when the patient is discharged from the hospital. Though the purpose of each type of note is distinct, they all capture the course of events that occur during the patient's stay. Hence, there could potentially

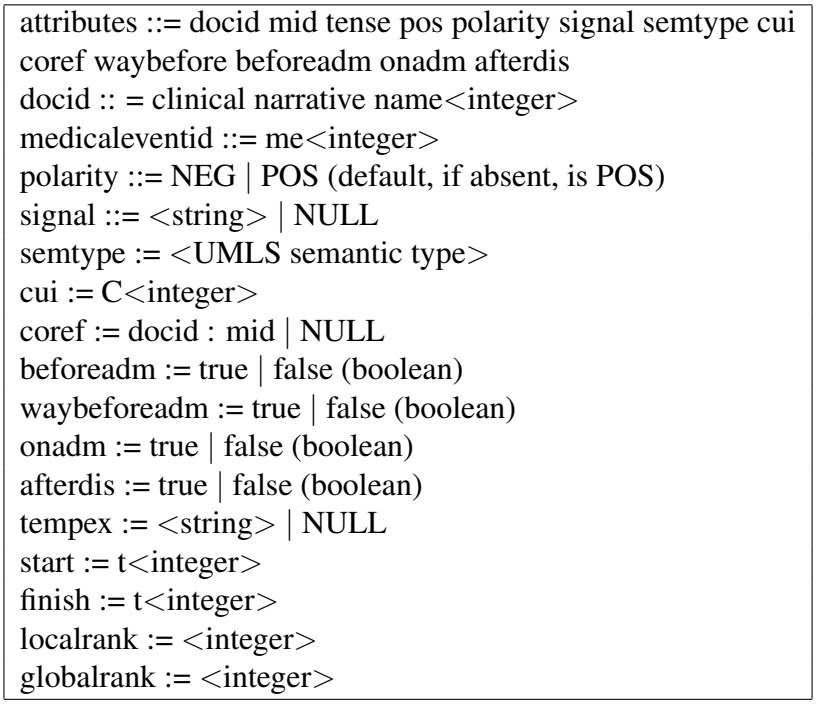


Figure 3.4: Medical Event BNF with additional attributes. This is an instance of the medical event start. If the start and finish are both known, there is another instance of the medical event with its own attribute values is created for the event with the attribute finish := t<integer> replacing start := t<integer>

be a considerable amount of overlapping entities across these narratives. The "coref" attribute along with the attribute "semcat" for the UMLS semantic category, helps identify two properties that help characterize overlap:

- Medical events that are semantically equivalent
- Medical events that take place at the same time point / time duration.

The instances of events are highlighted in our example sentence, "The patient was admitted to the hospital 2 days after chest pain stopped."

- <EVENT docid=ds2 mid=me1 tense=PAST pos=VERB polarity=POS>admitted </EVENT>
- <EVENT docid=ds2 mid=me2 polarity=POS semtype=Sign or Symptom cui=C0008031 coref=ds2:me15, hp1:me3 beforeadm=TRUE afterdis=FALSE tempex=t1 start=t1 finish=t1>chest pain </EVENT>
- <EVENT docid=ds2 mid=me3 tense=PAST pos=VERB polarity=POS> stopped</EVENT>

The event "admitted" is tagged with attributes "past tense" and "positive polarity." Similarly, "stopped" is tagged with attributes "past tense" and "positive polarity." The medical event "chest pain" is tagged with "polarity positive" since "chest pain" is not negated. The "semantic category" and "concept ID" for "chest pain" are "Sign or Symptom" and C0008031 respectively, as per UMLS meta-thesaurus. The "coref" attribute tells us that event with ID me15 from the patient's second "discharge summary" (ds2) and event with ID me3 from the patient's first "history and physical report" (hp1) are coreferring with the current event "chest pain."

attributes ::= docid tid type functionInDocument beginPoint endPoint
quant freq mod anchorTimeID
docid := clinical narrative name<integer>
tid ::= t<integer>
type ::= DATE | TIME | DURATION | SET
functionInDocument ::= ADMISSION TIME | DISCHARGE TIME | DOB | OTHER
beginPoint ::= tid
endPoint ::= tid
quant ::= every/ each etc.
freq ::= <integer> | 2 times etc.
value ::= duration | date
mod ::= BEFORE | AFTER | ON OR BEFORE | ON OR AFTER | LESS THAN | MORE THAN | EQUAL OR LESS | EQUAL OR MORE | START | MID | END | APPROX
eventAnchor ::= m<integer>

Figure 3.5: BNF for Temporal expressions

3.4.2 Annotating Temporal Expressions

We define a simplified version of the TIMEX tag from TimeML and adapt it to clinical narratives. Changes include attributes for time duration and rank. Additionally, we also adapt some attributes defined as part of the Temporal Constraint Structure that is used to implement TimeText [Zhou et al., 2006]. This includes attributes indicating the beginning and end of a time period and the medical events to which a temporal expression can be anchored. This tag is used to annotate any words or phrases that indicate the time point or time duration of an event. The simplified BNF for this tag is as follows: The document ID (docid) has the same interpretation as before. Temporal expression ID (tid) uniquely identifies a temporal expression within a narrative. The "type" of a temporal expression could be any one of the following:

- DATE: Describes a calendar time. For example, *Friday October 1 1999, yesterday* and *this summer*.

- TIME: Refers to a time of the day. For example, *eleven in the morning* and *last night*.

- DURATION: Describes a time duration. For example, *20 days, 3 hours* and *2 months*.

- SET: Describes a set of times. For example, *twice a week* and *every two days*.

- quant and freq: Used when a temporal expression is of the type SET to indicate temporal granularity and frequency.

- beginPoint and endPoint: Used when the start and end of a duration is anchored by another time expression.

The eventAnchor allows us to identify a temporal expression anchored to an event. These events are typically the standard admission or discharge events. For instance, "chest pain started just a day before admission." In this case, the temporal expression would be a day, the temporal signal would be before and the event anchor would be admission. However, temporal expressions could be anchored to medical events as well. The instances of temporal expressions are highlighted in our example sentence, "The patient was admitted to the hospital 2 days after chest pain stopped."

- <TIMEX docid=ds2 tid=t1 type=DURATION functionInDocument=OTHER eventAnchor=me2> 2 days </TIMEX>

In case of the temporal expression 2 days, the event type is "duration," and the event anchor is the medical event "chest pain" with id me2.

3.4.3 Temporal Nature of Clinical Text

The temporal relationship between medical events is varied and complicated. Zhou and Hripcsak [2007] identify six major categories of temporal expressions from a corpus of discharge summaries: "date and time," "relative date and time," "duration," "event-dependent temporal expression," "fuzzy time," and "recurring times." The study of temporal expressions in clinical text indicates that relative time (e.g., ever since the *episode* 2 days ago) may be more prevalent than absolute time (e.g., 06/03/2007). Further, temporal expressions may be fuzzy where "history of *cocaine use*" may imply that *cocaine use* started "2 years ago" or "10 years ago." All of this makes the problem of temporal relation learning from clinical text extremely challenging. However, addressing this problem is essential to many applications in the biomedical domain including patient recruitment for clinical trials. We briefly motivate the clinical trial recruitment task since we finally plan to evaluate the utility of the proposed methods for timeline generation in such a task.

3.4.4 Annotating Temporal Relations

The tag is used to annotate temporal relations between medical events. It is an important tag as it relates the previously defined tags in a meaningful manner to identify temporally related events. The temporal relations are adapted from Allen's temporal relations as defined in Chapter 2. The temporal relations are defined in terms of an interval calculus using thirteen mutually exclusive binary relations as different ways of relating two convex intervals. Using this calculus, given facts can be formalized and then used for automatic reasoning. The thirteen relations include before, after, meets (and its inverse e.g. event X meets Y and Y meets X), overlaps (and its inverse), starts (and its inverse), during (and its inverse), finishes (and its inverse), is equal to. The BNF for temporal relations is defined

attributes ::= trelid, eventid, relatedto, eventid, reltype
docid :: = clinical narrative name<integer>
tlinkid ::= trel<integer>
event ::= eid
relatedto event ::= eid
relationtype ::= BEFORE | AFTER | INCLUDES | IS INCLUDED | DURING
| DURING INV | SIMULTANEOUS | IAFTER | IBEFORE | IDENTITY | BEGINS
| ENDS | BEGUN BY | ENDED BY

Figure 3.6: BNF for temporal relations

as follows. The "trelid" attribute is used to uniquely identify a temporal relation within a document. The attributes "event" and "relatedto event" are used to specify the event IDs of the related events. The event relation type is specified with the "reltype" attribute. The temporal relation tag in case of our example sentence, "The patient was admitted to the hospital 2 days after chest pain stopped," would be as follows:

<TLINK docid=ds2 tlinkid=tlink1 event=e6 relatedto event = e5 relationtype=BEFORE>

In this example, event e6 is BEFORE event e5.

3.5 Comparison with TimeML and Analysis

We have adapted the TimeML[Pustejovsky et al., 2003a] annotation format to clinical narratives. The proposed changes include the following:

- A new class of events called "Medical" for annotating medical events.
- New attributes associated with instances of events with class "Medical." These include, medical event UMLS semantic type, UMLS concept ID, boolean attributes placing the medical event before / on admission / after admission / discharge, medical events that are coreferring, the "start time" and "end time" of a medical event

in order to learn medical event durations, rank of the event, both within and across clinical narratives for a patient.

- An "event anchor" for temporal expressions that help identify temporal expressions that are anchored to events.

Let us study how various tags are linked together with the help of the example sentence, "The patient was admitted to the hospital 2 days after chest pain stopped." In this sentence, we identified *admitted* and *stopped* as events and chest pain as a medical event. We also annotate "2 days" as a temporal expression and "after" as a temporal signal. The links between these annotations is as follows: The temporal expression 2 days is linked to the two medical events using the temporalanchor attribute of the medical events. The medical event chest pain is linked to the event admitted using the temporal relation tag. Let us also examine a few more sample sentences from a discharge summary and study their annotation as per the format described in this chapter.

"The patient *had been well* **until 4 months before** *admission*."

Here, medical events include "well" and "admission." "Until 4 months" is a temporal expression and "before" is a temporal signal. The temporal relations between events are as follows: "before" is anchored to "until 4 months," "until 4 months" is anchored to "well" and "well" is related to "admission" using relation type BEFORE

- <EVENT docid=ds1 eid=e1 class=OCCURRENCE>well</EVENT>
- <EVENT docid=ds1 eid=e2 class=OCCURRENCE>admission</EVENT>
- <TIMEX docid=ds1 tid=t1 type=DURATION functionInDocument=OTHER sid=s1 eventAnchor=e1:ei1> until 4 months </TIMEX>

- <TLINK docid=ds1 tlinkid=tlink1 eventinstance = e1 : ei1 relatedto eventinstance = e2 : ei1 relationtype=BEFORE>

"She was **again** *admitted* to the hospital where *acylovir given intravenously* for **10 days** had no effect."

Events include "admitted" and "effect." Medical events include "acylovir" and "intravenously" "for 10 days" is a temporal expression The temporal relations between events are as follows: "for 10 days" is anchored to "acylovir." "for 10 days" is anchored to "intravenously." "no effect" is anchored to "acylovir." "admitted" is related to "acylovir" using relation type BEFORE

- <EVENT docid=ds1 eid=e1 class=OCCURRENCE>admitted</EVENT>
- <EVENT docid=ds1 eid=e2 class=MEDICAL>acylovir</EVENT>
- <TIMEX docid=ds1 tid=t1 type=DURATION functionInDocument=OTHER sid=s1 eventanchor=e2:ei1> for 10 days </TIMEX>
- <TLINK docid=ds1 tlinkid=tlink1 eventinstance = e1 : ei1 relatedto eventinstance = e2 : ei1 relationtype=BEFORE>

3.6 Annotator Agreement

The corpus used to calculate agreement consists of three clinical notes from a chronic lymphocytic leukemia (CLL) patient's record. This patient was one of approximately 2060 CLL patient records we have collected over the last 10 years at The Ohio State University Wexner Medical Center (OSUWMC). The notes consisted of a discharge summary, radiology report and a history and physical report with an average of 600 words per narrative.

A team of 5 annotators with diverse backgrounds, but with some experience in understanding medical terminology, was hired to annotate the corpus. Our team had one medical student with a degree biomedical engineering, but no clinical experience (medstud); three recently graduated nurse practitioners with clinical experience gathered through the process of receiving their nurse practitioner degrees (np); and one graduate entry nurse practitioner student with some clinical experience and experience in working with biomedical documents (nstud). In order to achieve the level of detail we wanted in our annotations, each annotator required approximately one month's effort. This included clinical informatics IRB training, getting them familiar with the UMLS, explaining the motivation behind the task, having them read and understand the annotation guidelines and annotating a sample clinical narrative. The annotations efforts described in this paper were coded in Excel sheets.

An important aspect of annotating a large corpus is consistency. We measure consistency in terms of inter-annotator reliability. Inter-annotator agreement measures the consistency in annotating a particular concept across annotators. We measure inter-annotator reliability using Cohen's kappa statistic11. Kappa is interpreted as the proportion of agreement among raters after chance agreement has been removed. It can be expressed as follows:

$$Kappa = \frac{Proportion\,of\,observed\,agreement - chance\,agreement}{1 - chance\,agreement} \tag{3.1}$$

Chance agreement is estimated by the proportion of agreements that would be expected if the observer's ratings were completely random. Chance agreement increases as the variability of observed ratings decreases. The use of Kappa requires minimal assumptions about the underlying nature of the data. Three data collection conditions should be met: 1) The subjects to be rated are independent of each other, 2) the raters score the subjects in an independent fashion, and 3) the rating categories are mutually exclusive and exhaustive.

The flexibility of different forms of kappa is also a major advantage. Kappa is appropriate for nominal and ordinal data, where there are two or more raters per subject. Kappa can be calculated for each scale point or averaged into a generalized Kappa across the entire set of ratings.

We use the methods proposed by [Conger, 1980] to calculate agreement between multiple annotators. The author suggests a multiple-rater agreement statistic obtained by averaging all pairwise overall and chance-corrected probabilities.

For annotation of medical events, if the annotators marked a partially overlapping section of text as a medical event, we considered it to be an agreement. For medical event coreference, we considered agreement in a pairwise fashion. For example, for events A, B, and C, if annotator #1 identified events A and B corefer, and events B and C corefer, but annotator #2 only identified events B and C corefer, we count that as having 1 annotation in agreement. We also did not consider transitive closure. In the example given, if annotator #1 also identified events A and C coreferring and annotator #2 also identified events A and B as coreferring, which would count as having 2 annotations in agreement, and not 3. For temporal relations, we considered whether or not the same pairwise temporal relationship between events was identified. With medical events, we further analyzed whether or not the coders identified the medical events with the same UMLS CUIs.

Inter-annotator agreement metrics The total number of words in each clinical narrative (CN) is as follows: CN1= 454, CN2 = 612, CN3=386. Given the text in each narrative, the main unit of annotation is a medical event. Some examples of medical events in these clinical narratives include B-cell lymphoma, mass, physical examination, and beta blockers. We present statistics on the number of medical events, coreference pairs and temporal relations annotated by each annotator in the clinical narratives in Table 3.1.

The number of medical events, coreference pairs and temporal relations noted by each annotator in three different clinical narratives. We also present precision and recall metrics for each annotator measured against a reference annotator. In Tables 3.2 and 3.3, we present precision and recall values for medical event mentions, coreference pairs, and temporal relation pairs across the three narratives with the medical student (medstud) as the reference annotator. The other annotators are a nursing student (nstud) and three nurse practitioners (np1, np2, np3).

Precision and recall values for medical event mentions across the three narratives with (medstud) as the reference annotator.

	No. of Medical Events			Coreference Relations			Temporal Relations		
Annotator	**CN1**	**CN2**	**CN3**	**CN1**	**CN2**	**CN3**	**CN1**	**CN2**	**CN3**
medstud	65	81	53	15	19	12	15	19	12
nstud	58	70	58	8	10	7	8	10	7
np1	67	95	69	13	15	10	13	15	10
np2	52	87	70	12	16	10	12	16	10
np3	59	76	55	12	15	10	12	15	10

Table 3.1: The number of medical events, coreference pairs and temporal relations noted by each annotator in three different clinical narratives.

Annotator	**CN1**		**CN2**		**CN3**	
	P	**R**	**P**	**R**	**P**	**R**
nstud	94.8	84.6	89.3	87.5	88.2	80.4
np1	86.6	89.2	85.1	98.8	90.2	96.6
np2	96.1	76.9	90.8	97.5	88.6	95.2
np3	94.9	81.2	97.4	91.4	91.4	89.8

Table 3.2: Precision and recall percentages for medical event mentions across the three narratives with (medstud) as the reference annotator.

Annotator	CN1		CN2		CN3	
	P	**R**	**P**	**R**	**P**	**R**
nstud	87.7	98.3	86.4	85.2	84.1	79.6
np1	92.5	89.7	72.6	98.6	94.6	94.3
np2	92.3	82.8	75.3	95.7	85.3	92.4
np3	84.7	86.2	92	100	93.4	90.7

Table 3.3: Precision and recall values for coreference pairs across the three narratives with (medstud) as the reference annotator

In Table 3.5, we present the average pairwise Cohen's kappa for medical events, coreferences, temporal relations, and medical event concept unique identifiers across narratives CN1, CN2, and CN3. To further illustrate and clarify our results, we plot these values in Figure 3.7.

Annotator	CN1		CN2		CN3	
	P	**R**	**P**	**R**	**P**	**R**
nstud	96.9	94.0	98.7	85.2	92.7	82.4
np1	89.6	77.6	97.1	71.5	87.4	75.7
np2	92.3	71.6	96.5	88.4	86.8	85.8
np3	91.5	80.5	96.1	78.9	83.6	74.3

Table 3.4: Precision and recall values for temporal relation pairs across the three narratives with (medstud) as the reference annotator.

In looking at our results, we note that for medical events, the average kappa agreement between annotators from different backgrounds mostly varies between 0.80 and 0.85. The highest agreement is between np2 and np3 of 0.96. The lowest agreement was between nstud and np1 (kappa=0.78). For medical event coreference, the agreement is high between the medical student and the nurse practitioners, with kappa ranging from 0.81 and 0.92. The

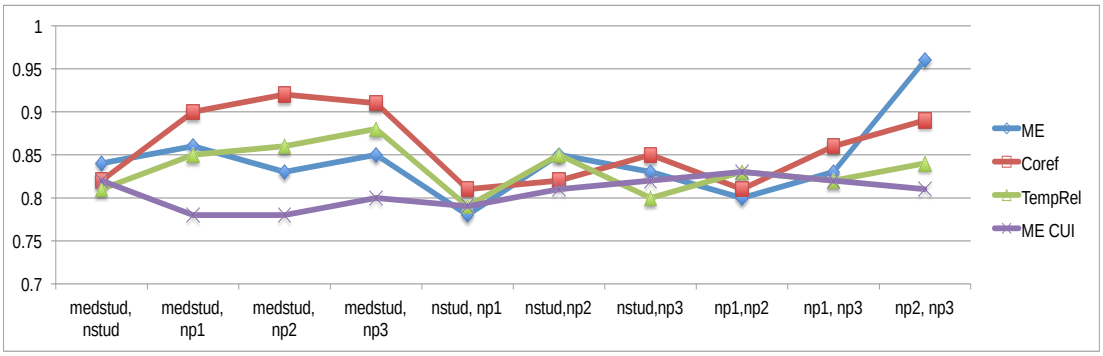


Figure 3.7: Pairwise Kappa agreement for medical events, coreference, temporal relations, and medical event CUIs. The pattern of agreement across the categories for different annotator pairs is more or less the same.

Annotator Pairs	**ME**	**Coref**	**TempRel**	**ME CUI**
medstud, nstud	0.84	0.82	0.81	0.82
medstud, np1	0.86	0.9	0.85	0.78
medstud, np2	0.83	0.92	0.86	0.78
medstud, np3	0.85	0.91	0.88	0.8
nstud, np1	0.78	0.81	0.79	0.79
nstud,np2	0.85	0.82	0.85	0.81
nstud,np3	0.83	0.85	0.8	0.82
np1,np2	0.8	0.81	0.83	0.83
np1, np3	0.83	0.86	0.82	0.82
np2, np3	0.96	0.89	0.84	0.81
Average kappa	0.843	0.859	0.833	0.806

Table 3.5: The average pair wise Cohen's kappa for medical events, coreference, temporal relations, and medical event concept unique identifiers across CN1, CN2 and CN3.

highest agreement of 0.92 is between medstud and np2. With temporal relations, the kappa agreement varies between 0.79 and 0.88 with the highest Kappa of 0.88 between medstud and np3. When looking at medical event CUIs, the agreement varies between 0.78 and 0.83.

The overall average inter-annotator kappa statistic for medical events, coreferences, temporal relations, and medical event concept unique identifiers was 0.843, 0.859, 0.833, and 0.806 (Table 3.5), respectively, all of which show excellent agreement. The average pairwise Cohen's kappa [Conger, 1980] is highest between when medstud is paired with other annotators (kappa=0.86). The average of the pairwise agreement among the nurse practitioners is 0.846, whereas the average of the pairwise Cohen's kappa between each nstud/np from the nursing group and medstud is 0.82. This is across all three categories medical events, coreferences, and temporal relations.

While overall agreement for coreferences is already high (0.859), it may be an underestimate. We did not consider transitive closure in our calculations. For example, if annotator #1 marked events A and B corefer, B and C corefer, and A and C corefer, but annotator #2 marked events A and B corefer and B and C corefer, we still consider there to be a missing annotation (A and C corefer).

3.7 Error Analysis

While the inter-annotator agreement for medical event CUIs was lower than for medical events, coreference, and temporal relations, agreement was still very high. Figure 3.7 indicates that the pattern of agreement across various annotations by different annotators with varying clinical expertise is more or less uniform.

In an analysis of the reason behind the discrepancies we discovered that in many cases there was either a discrepancy in the granularity to which the medical events were coded or whether or not clinical judgment was used in selecting the CUI. For example, all of our annotators marked "B-Cell CLL" as an event. The three NPs coded this term as "C0023434: Chronic Lymphocytic Leukemia." Both medstud and nstud coded this event as "C0475774: B-cell chronic lymphocytic leukemia variant." While both could be considered correct annotations for "B-Cell CLL," C0475774 is the more specific term. In another example, all of the annotators marked the phrase "white blood cell count of 10,000." For this situation, medstud selected "C0750426: white blood cell count increased," while nstud selected "C0023508: White Blood Cell count procedure." In contrast, all three NPs selected different CUIs, applying clinical judgment to the medical events. Np2 selected "C0860797: differential white blood cell count normal." Overall we found the medical student's (who did not have any real life clinic experience) annotations remained true to what was observed and could be inferred based on the data. However, the nursing student and the nurse practitioners often used clinical judgment to infer certain annotations that were not directly observed in the data. For instance, classifying something as an acute condition based on certain readings or values in the text.

3.8 Discussion

One limitation of our study is the small number of narratives. The main reason for this limitation is due to the large amount of effort required to annotate the narratives to the detail that we desired. Given that the 3 narratives used in the study required a month of effort for each annotator, we needed to begin having the annotators annotate non-overlapping narratives in order to increase the overall size of our gold standard. Another limitation

of our study is the lack of a physician annotator to compare their annotations. Given the amount of time required for our existing annotators to complete the annotations, having an additional physician annotator was not feasible. One reason for the time required to generate these annotations may be the lack of sophisticated annotation tools. Although we developed an annotation schema using Knowtator [Ogren, 2006] and trained our annotators to use this tool, the nursing and medical students always prefered using Excel sheets to record annotations.

3.9 Clinical Corpus and Timeline Evaluation

With the help of the annotation process described in the previous sections, we annotate a corpus of clinical data obtained from the The Ohio State University Wexner Medical Center. The dataset contains patient records for Chronic Lymphocytic Leukemia (CLL) and Methicillin-resistant Staphylococcus aureus (MRSA). The annotators annotate admission notes, history and physical reports, radiology reports, pathology reports and discharge summaries for 7 patients in the corpus. The average number of clinical narratives is 80 with an average of around 23 medical events per narrative. Table 3.6 shows the distribution of medical events across clinical narratives.

15% of the medical events corefer within and across the clinical narratives for each patient. The narratives are annotated with medical events and their attributes as per the format specified in Figure 3.4. The attributes also include time-bins, coreference relations between medical events, and temporal ranks for medical events both within and across narratives. It also includes the temporal starts and stops that associate each event with relevant temporal expressions. This allows representing medical events as an interval by splitting each medical event into a start and a stop whenever relevant temporal information

Patient	Narratives	Medical events
p1	5	125
p2	9	239
p3	20	488
p4	13	318
p5	8	220
p6	10	136
p7	15	297

Table 3.6: Distribution of medical events across clinical narratives for each patient

is available. For instance, if we know that *palpitations* started yesterday and stopped *2 days later*, we can represent *palpitations* as $palpitations_{start}$ and $palpitations_{stop}$. When specific information about the starts and stops are unavailable, and for certain machine learning methods (time-bin learning in Chapter 4 and coreference resolution in Chapter 5), we assume a point notation for the medical event. (considering that the event started and the stop is unknown). The annotation for temporal relations is generated as per the format specified in Figure 3.6.

Evaluating Timelines. Once the proposed system generates a medical event timeline, we need to evaluate it against the gold-standard timeline in our annotations. This is illustrated in Figure 3.8.

We measure the transformations required to obtain the reference sequence in the gold-standard from the one generated by our system. We adapt word error rate, which is an edit-distance measure popularly used in automatic speech recognition as an evaluation metric. Thus, we calculate the fewest modifications (edits) required to the system output so that it

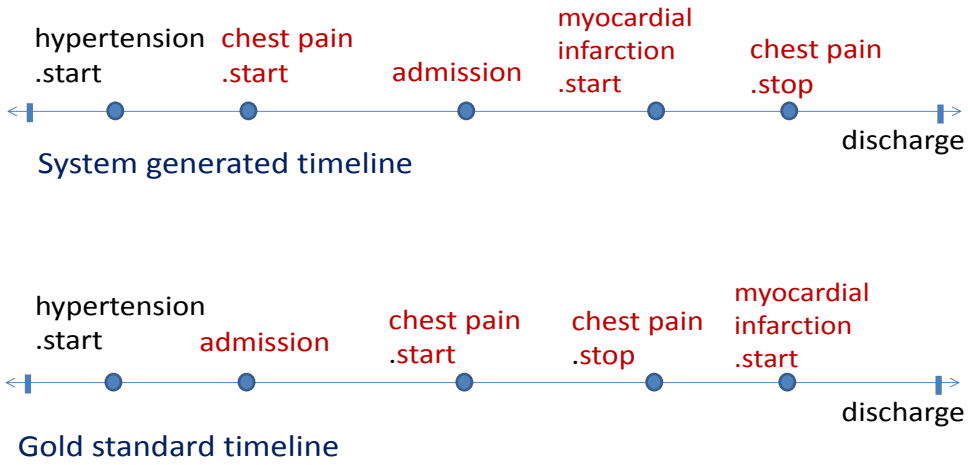


Figure 3.8: Example of a system generated timeline and the gold-standard timeline provided by the annotators.

is the same as the reference in the gold standard. The accuracy is given by

$$Accuracy = \frac{Total\ MEs - (I\ +\ S\ +\ D)}{Total\ MEs} \tag{3.2}$$

Here, "Total MEs" corresponds to the total number of medical events in the sequence, and insertions (I), substituitions (S) and deletions (D) correspond to the medical events required to the system generated timeline to obtain the gold-standard timeline in our annotations. In Figure 3.8, in going from the system generated output (sysout) to the gold standard output (gsout), we insert admission in position 2, replace admission with chest pain$_{stop}$ in position 4 and delete chest pain$_{stop}$ in position 6. Thus, we require 3 edit operations in going from sysout to gsout. Since the total number of medical events is 5, this gives us an accuracy of 40%.

3.10 Conclusion

In this chapter, we first established the need for an annotated corpus of clinical narratives by comparing the nature of clinical language to Timebank. We then discussed in detail the contents of certain types of clinical narratives like discharge summaries and radiology reports, and provided an annotation schema for annotating elements of these narratives. Finally, we described the metric used to evaluate the timeline generated by the methodology proposed in this dissertation. In the chapters that follow, we describe in detail various stages of the timeline generation process that leverage the annotations that were described in this chapter.

CHAPTER 4: COARSE INTRA-NARRATIVE TEMPORAL ORDERING OF MEDICAL EVENTS

Temporal relationships between medical events can be viewed at different granularities, where the medical events could be part of the same data source or across different data sources. Thus, in developing this framework for timeline generation, we begin at a higher level of temporal granularity and then drill down to a finer level. In the sections that follow, we describe a methodology for generating a coarse timeline of medical events within a narrative. This in turn can be used to derive an overall partially ordered timeline across narratives.[9]

4.1 Introduction

In the context of clinical text, the notion of time can be defined as follows: "A duration or relation of events expressed in terms of past, present, and future, and measured in units such as minutes, hours, days, months, or years." [10] These events describe medical conditions affecting the patient's health, or tests and procedures performed on a patient. Over the course of time, the EHR of the patient could have hundreds of such unstructured clinical narratives for various admissions and discharges, daily progress, lab tests, physical review, etc. Sample excerpts from two different clinical narratives (cn1 and cn2) of the same patient, generated over time, are shown in Figures 4.1 and 4.2.

[9]This work has been published in BioNLP 2012. P. Raghavan, E. Fosler-Lussier, and A. Lai, "Temporal Classification of Medical Events," BioNLP 2012.

[10]Stedman's Medical Dictionary: http://www.medilexicon.com

HISTORY PHYSICAL DATE: **09/01/2007**
NAME: Smith Daniel T MR#: XXX-XX-XXXX
ATTENDING PHYSICIAN: John Payne MD DOB: 03/10/1940
HISTORY OF PRESENT ILLNESS
The patient is a 67-year-old Caucasian male with a history of paresis secondary to back injury who is bedridden status post colostomy and PEG tube who was brought by EMS with a history of fever. The patient gives a history of fever on and off associated with chills for the last 1 month. He does give a history of decubitus ulcer on the back but his main complaint is fever associated with epigastric discomfort.
PAST MEDICAL HISTORY
Significant for polymicrobial infection in the blood as well as in the urine in July 2007 history of back injury with paraparesis. He is status post PEG tube and colostomy tube.
REVIEW OF SYSTEMS
Positive for decubitus ulcer. No cough. There is fever. No shortness of breath.
PHYSICAL EXAMINATION
On physical exam the patient is a debilitated malnourished gentleman in mild distress. Abdomen showed PEG tube with discharging pus and there are multiple scars one in the midline. It had a healing wound. Bowel sounds were present. Extremities revealed pain and atrophied muscles in the lower extremities with decubitus ulcer which had a transparent bandage in the decubitus area which was stage 2-3. CNS - The patient is alert and awake x3. There was good power in both upper extremities. Cranial nerves II-XII grossly intact.

Figure 4.1: Excerpt from a de-identified clinical narrative (cn1) [2007]

There has been a lot of interest in building timelines of medical events from across such unstructured patient narratives [Jung et al., 2011; Zhou and Hripcsak, 2007]. An important characteristic of a clinical narrative is that the medical events in the same narrative are more or less semantically related by narrative discourse structure. However, medical events in the narrative are not ordered chronologically. The clinical narrative structure moves back and forth in time and is not always temporally coherent (as seen in Figure 4.1 and 4.2). Thus, creating a timeline from longitudinal clinical text requires learning Allen's temporal relations [Allen, 1981] such as *before, simultaneous, includes, overlaps, begins, ends* and their inverses between medical events found within and across patient narratives. However,

learning temporal relations for fine-grained temporal ordering of medical events in clinical text is challenging: the temporal cues typically found in clinical text may not always be sufficient for this task. Instead, as a first step towards temporal ordering, we learn a coarser temporal ordering of medical events by learning to assign them to coarsely defined temporal classes that we call time-bins. Time-bins like *way-before-admission, before-admission, on-admission, after-admission, after-discharge* are defined around an anchor date that is mostly likely to occur in every clinical narrative such as the admission date or the date of creation of the narrative.

HISTORY PHYSICAL DATE: 06/17/2009
NAME: Smith Bob MR#: XXX-XX-XXXX
ATTENDING PHYSICIAN: Bill Payne MD DOB: 02/28/1960

He is a 48-year-old African American gentleman with a history of hypertension and cocaine use. He has hidradenitis of both axilla resected. The patient is MRSA positive on IV antibiotics at the present time. The patient's physical condition is excellent but he had MRSA in the axilla for hidradenitis that was devastating. The wounds now are very large but he is wound vac and being changed to alginate. Both axilla show major wounds of 20-25 cm in diameter and 4-5 cm deep in overall size and he has excoriations on his chest from the tape. The plan is to change him from vac to alginate and see him in a week.

Figure 4.2: Excerpt from another de-identified clinical narrative (cn2)[later in 2007]

4.2 Contributions

The main innovation in this chapter is the assignment of medical events to time-bins centered around a reference date. The assignment of medical events to time-bins is done with the help of features based on narrative structure and explicit temporal expressions. This allows us to label a sequence of medical events from each clinical narrative with a highly probable sequence of time-bins using Conditional Random Fields (CRFs). The learned time-bins can be used as an informative temporal feature for tasks such as medical event coreference resolution and fine-grained temporal ordering of medical events as demonstrated in Chapters 5 and 6. Moreover, the coarse temporal orderings of medical events within each narrative, can be partially combined with the help of explicit dates in each narrative, like the admission and discharge date, to create a comprehensive coarse partial ordered timeline of medical events across the patient's history. Such a timeline by itself is useful to maybe useful clinical tasks like clinical decision making where the temporal constraints that need to be resolved are coarse.

4.3 Related Work

Prior work in machine learning of temporal relations on the WSJ-based Timebank corpus [Pustejovsky et al., 2003a] include Mani et al. [2006]; Chambers et al. [2007]; Verhagen et al. [2009], who experimented with pairwise classification for learning temporal relations between event pairs. However, as described in Chapter 3, it is difficult to directly adopt Timebank for temporal reasoning in clinical text. Previous attempts at learning temporal relations between medical events in clinical text include Jung et al. [2011]; Zhou et al. [2006]; Bramsen et al. [2006]. A comprehensive survey of temporal reasoning in medical data is provided by Zhou and Hripcsak [2007]. Gaizauskas et al. [2006] learn the temporal

relations *before, after, is_included* between events from a corpus of clinical text much like the event-event relation TLINK learning in Timebank [Pustejovsky et al., 2003a]. Bramsen et al. [2006] characterize temporal organization of discharge summaries in terms of temporal segments and their ordering, where a temporal segment to be a fragment of text that does not exhibit changes in temporal focus. They learn temporal relations between segments using a supervised classifier. The clinical text corpora used in these studies are not freely available. Before getting into fine-grained temporal ordering, in this task, we define coarse time-bins and classify medical events into one of the time-bins. Our task varies from prior work in the following aspects: Instead of pairwise classification of events into multiple granular temporal relations, we use coarse granularity time-bins and classify each instance of a medical event and assign it to a time-bin. This is a relatively easier task; learning time-bins with high accuracy can be then used to inform fine-grained temporal relation learning models. Using a sequence tagging model allows us to capture temporal progression from narrative order of medical events in clinical text (see temporal features in section 4.4.3 for an example).

4.4 Assigning Medical Events to Time-bins

We adress the problem of assigning medical events to time-bins using a sequence learning approach. In order to understand why sequence learning model is appropriate for this task, and what features help the machine learning model, we first describe the medical event representation used for this task.

4.4.1 Medical event representation

In order to keep the task of classifying medical events into coarse time-bins relatively easy to learn, we use a time-point notation for representing medical events. Each mention

of a medical event is assigned to a time-bin without taking into consideration whether it denotes the beginning or end of that event. We also do not differentiate between coreferences of the same medical event. Thus, if *chest pain* is mentioned in the past medical history and the same *chest pain* continues to persist in the after admission time-bin, the two different mentions of chest pain get anchored to different time-bins. Similarly, *cocaine use* started in the history of the patient and *cocaine abuse* still persists. We assign the two different mentions of this medical event into different time-bins.

4.4.2 Time-bins

We learn to classify medical events into one of the following time-bins: *way before admission, before admission, on admission, after admission, after discharge*. The time-bin *way before admission* is intended to capture all medical events that happened in the past medical history of the patient but are not mentioned as being directly related to the present illness. *Before admission* captures events that occurred before admission and are related to the present illness. *On admission* captures medical events that occur on the day of admission. *After admission* captures medical events that occur between admission and discharge (during the hospital stay or clinic visit). Finally, medical events that are supposed to occur in the future after the patient is discharged belong to the class *after discharge*.

Further, the time duration of each time-bin varies based on the patient. For instance, the hospital stay of a patient could be 4 days or 1 month or a year. This makes it very difficult to define exact time-bins based on the intuitions described above. In order to make the problem consistent across different patients, we restrict *way before admission* to events that happened more than a year ago and *before admission* to events that occurred in

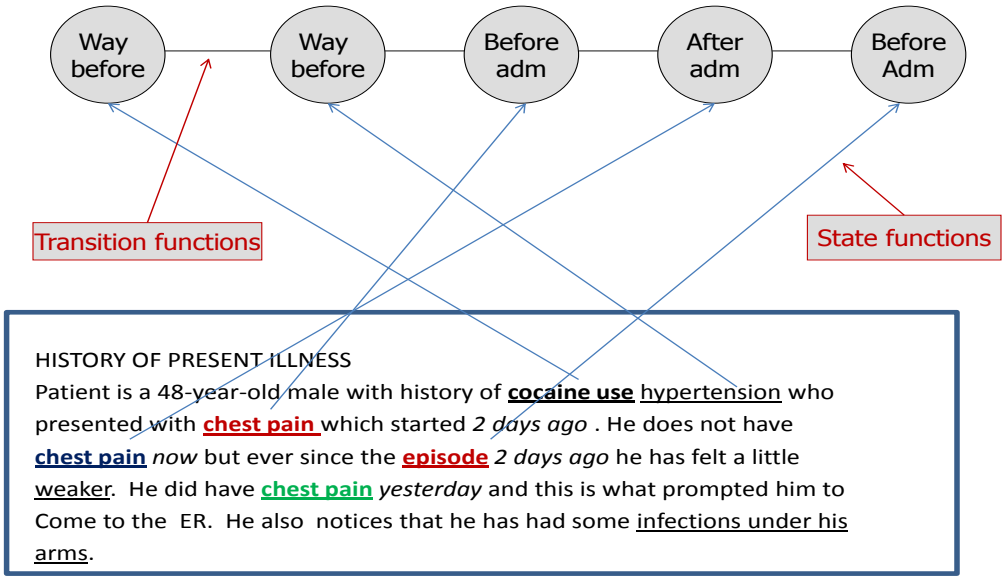


Figure 4.3: Linear chain CRF used to assign time-bin label sequence to a medical event sequence

the same year before admission. If it is unclear as to when in the past the medical event occurred, we assume it happened *way before admission*.

For the task proposed in this paper, an observation sequence is composed of medical events in the order in which they appear in a clinical narrative, and the state sequence is the corresponding label sequence of time-bins. This is illustrated in Figure 4.3. Each label in the label sequence could be any one of the time-bins *way before admission (wba), before admission (ba), on admission (a), after admission (aa), after discharge (ad)*. Thus, given a sequence of medical events in narrative order we learn a corresponding label sequence of time-bins {wba, b, a, aa, ad}.

The probability of time-bin (label) sequence y, given a medical event (input) sequence x, is given by,

$$P(Y|X) = \exp \sum_{i} (S(x, y, i) + T(x, y, i)) \tag{4.1}$$

where i is the medical event index and S and T are the state and transition features respectively. State features S consider the label of a single medical event and are defined as,

$$S(x, y, i) = \sum_{j} \lambda_j s_j(y, x, i) \tag{4.2}$$

Transition features consider the mutual dependence of labels y_{i-1} and y_i (dependence between the time-bins of the current and previous medical event in the sequence) and are given by,

$$T(x, y, i) = \sum_{k} \mu_k t_k(y_{i-1}, y_i, x, i) \tag{4.3}$$

where s_j and t_k are the state and transition feature functions. Above, s_j is a state feature function, and λ_j is its associated weight and t_j is a transition function, and μ_j is its associated weight. In contrast to the state function, the transition function takes as input the current label as well as the previous label, in addition to the data. The mutual dependence between the time-bins of the current and previous medical events is observed frequently in sections of the text describing the history of the patient. Around 40% of the medical events in gold standard corpus demonstrate such dependencies.

4.4.3 Feature Space

We extract features from medical event sequences found in each clinical narrative. The extracted feature-set captures narrative structure in terms of the narrative type, sections, section transitions, and position in document. The medical event and the context in which it is mentioned is captured with the help of lexical features. The temporal features resolve

temporal references and associate medical events with temporal expressions wherever possible.

Section-based features. Determining the document-level structure of a clinical narrative is useful in mapping medical events to time-bins. This can be achieved by identifying different sections in different types of clinical narratives and relating them to different time-bins. Commonly found sections in discharge summaries and history and physical reports include: "past medical history," "history of present illness," "findings on admission," "physical examination," "review of systems," "impression," and "assessment/plan."Some clinical notes like cn2 in Figure 4.2 may not have any section information.

The combined feature representing the type of clinical narrative along with the sections can be informative. Section transitions may also indicate a temporal pattern for medical events mentioned across those sections. For e.g., "past medical history" (*way before admission*), followed by "history of present illness" (*way before admission*), followed by "findings on admission" (on admission), followed by "physical examination" (after admission), followed by "assessment/plan" (*discharge*). Medical events in different types of sections may also exhibit different temporal patterns. A "history of present illness" section may start with diseases and diagnoses 30 years ago and then proceed to talk about them in the context of a medical condition that happened few years ago and finally describe the patient's condition on admission.

In addition to the section information, we also use other features extracted from the clinical narrative structure such as the position of the medical concept in the section and in the narrative.

Lexical features. Bigrams are pairs of words that occur in close proximity to each other, and in a particular order. The bigrams preceding the medical event in the narrative can

be useful in determining when it occurred. For instance, "**history of** *cocaine use and hypertension*," "**presents with** *chest pain*," "**have** *chest pain*," "**since the** *episode*," etc. If the preceding bigram contains a verb, we also extract the tense of the verb as a feature. However, tense is not always helpful in learning the time of occurrence of a medical event. Consider the following line from cn2 in Figure 4.2, "He has *hidradenitis of both axilla resected*." Though "has" is in present tense, the medical event has actually occurred in the history and is only being observed and noted now. Additionally, we also explicitly include the preceding bigrams and the tense of verb for the previous and next medical event as a feature for the current medical event.

Every medical event that occurs above a certain frequency threshold in all the clinical narratives of a particular patient is also represented as a binary feature. More frequent medical events tend to occur in the history of the patient, for example, *cocaine use*. We use a threshold of 3 in our experiments. The medical event frequency in also calculated in combination with other features such as the type of clinical narrative and section type.

Dictionary features. We map each medical event to the closest concept in the UMLS Metathesaurus and extract its semantic category. The semantic categories in UMLS include "Finding," "Disease or Syndrome," "Therapeutic or Preventative procedure," "Congenital abnormality," and "Pathologic Function." The intuition behind this is that medical events associated with certain semantic categories may be more likely to occur within certain time-bins. For instance, a medical event classified as "Congenital abnormality" may be more likely to occur *way before admission*.

Temporal features. Temporal features are derived from any explicit dates that are in the same sentence as the medical concept. The gold-standard corpus contains annotations for temporal anchors for events. Although there are no explicit dates in cn1 and cn2, there

may be narratives where there are mentions of dates such as *fever* on June 7^{th}, 2007. In some cases, there may also be indirect references to dates, which tell us when the medical event occurred. The reference date with respect to which the indirect temporal reference is made depends on the type of note. In case of history and physical notes, the reference date is usually the admission date. For instance, in *chest pain which started 2 days ago*, this would mean *chest pain* which started 2 days before admission. Since the admission date is 06/03/2007 (3^{rd} June 2007), *chest pain* would have started on 06/01/2007 (1^{st} June 2007). Similarly, 3 to 4 months ago resolves to February 2007 or March 2007 and 2 to 3 weeks ago resolves to first or second week of May 2007. Whenever the exact date is fuzzy, we assume the date that is farthest from the reference date as accurate. So in case of these examples, February 2007 and first week of May 2007 are assumed to be correct. We also calculate the difference between admission date and these dates associated with medical events. Another fuzzy temporal expression is "history of," where history could mean any time frame before admission. We assume that any medical event mentioned along with "history of" has occurred *way before admission*.

4.5 Experiments

We use the gold-standard corpus described in Chapter 3, Section 3.9 for our experiments. We conduct two sets of experiments with the clinical narratives in this corpus: 1) Medical event, time-bin experiments using hand-tagged features from the corpus and 2) Medical event, time-bin experiments using automatically extracted features from the corpus.

We first conducted experiments using the hand-tagged features in our corpus. We extracted the features described in the previous sections and used 10-fold cross validation.

We use the Mallet[11] implementation of CRFs and MaxEnt. CRFs are trained by Limited-Memory Broyden-Fletcher-Goldfarb-Shanno (BFGS) for our experiments. The per-class accuracy values of both sequence tagging using CRFs and using a MaxEnt model are indicated in Table 4.1.

When modeled as a multi-class classification task using MaxEnt, we get an average precision of 81.2% and average recall of 71.4% whereas using CRFs we obtain an average precision of 89.4% and average recall of 79.2%. In order to determine the utility of temporal features, we do a feature ablation study with the temporal features removed. In this case the average precision of the CRF is 79.5% and average recall is 67.2%. Similarly, when we remove the section-based features, the average precision of the CRF is 82.7% and average recall is 72.3%. The section-based features seems to impact the precision of the *on admission* and *after admission* time-bins the most.

We compare our approach for classifying medical events to time-bins with the following rule-based baseline. We assign medical events to time-bins based on the type of narrative, any explicit dates and section in which they occur. Each section is associated with a pre-defined time-bin. In the case of the sections in cn1, any medical event under "history of present illness" is *before admission*, "review of systems" is *after admission* and "assessment/plan" is *discharge*. If the narrative has a "past medical history" or a similar section, the events mentioned under it would be assigned to *way before admission* However, this baseline does not work for clinical narratives like 4.2 that do not have section information. This model gives us an average precision of 58.02% and average recall of 60.26% across the 5 time-bins. Per-class predictions for the baseline are shown in Table 4.1.

[11] http://mallet.cs.umass.edu

Class (time-bin)	Section baseline		MaxEnt		CRF	
	P	R	P	R	P	R
way before admission	56.3	61.4	72.4	63.5	79.8	66.7
before admission	60.2	57.5	83.4	80.8	92.0	92.4
on admission	63.8	59.1	76.6	72.1	87.5	75.2
after admission	57.5	68.2	88.6	82.1	93.6	99.1
after discharge	52.3	55.1	85.2	58.7	94.3	62.5

Table 4.1: Time-bin predictions by the section baseline method and per-class precision (P) and recall (R) for medical events, time-bins using hand-tagged extracted features.

The most common false positives for the *before admission* class are medical events belonging to *on admission*. This may be due to lack of temporal features to indicate that the event happened on the same day as admission. Frequently, medical events that belong to the *aa*, *ba* and *wa* time-bin get classified as *after discharge*. One of the reasons for this could be misleading section information in case of historical medical events mentioned in the assessment/plan section.

Next, we conduct experiments using automatically extracted features. This is done as follows. The medical events are extracted using MetaMap Aronson [2001], which recognizes medical concepts and codes them using UMLS. Based on this UMLS code, we can extract the semantic category associated with the code. Compared to the 1854 medical events marked by the annotators, MetaMap identifies 1257 medical events, which are a subset of the 1854. The UMLS coding by the annotators is more contextually relevant and precise.

We use a rule-based algorithm to identify and extract document structure based features such as sections from clinical narratives. In case of the lexical features, we extract bigrams and calculate the tense of the verb preceding the medical event using the Stanford NLP

Gold-standard Features		
	P	R
medical event	81.2	71.4
CRF	89.4	79.2
CRF (no temp. feats)	79.5	67.2
CRF (no section feats)	82.7	72.3
Automatic Features		
	P	R
medical event	74.3	66.5
CRF	79.6	69.7
Baseline (P;R)	58.0	60.3

Table 4.2: Overall Result Summary: Average precision (P) and recall (R) with manually annotated gold-standard features, automatically extracted features and the baseline.

software.[12] The temporal features are extracted with the help of TimeText developed by Zhou and Hripcsak [2007] that automatically annotates temporal expressions in clinical text. However, it is not able to capture many of the implicit temporal references. Following this, a temporal expression is linked to a medical event if it occurs in the same sentence as the medical event.

The average precision and recall of the MaxEnt model using automatically extracted features is 74.3% and 66.5% respectively. Sequence tagging using CRFs gives us an average precision and recall of 79.6% and 69.7% respectively. Although the results are not as good as using hand-tagged features, they are certainly promising. One reason for the loss in accuracy could be because the automatically calculated temporal features are not as precise as the hand-tagged ones. These results are summarized in Table 4.2.

[12] http://nlp.stanford.edu/software

4.6 Discussion

Consider the clinical narratives cn1 and cn2 in Figures 4.1 and 4.2. The medical events assigned to time-bins allow us to derive a coarse temporal order between medical events within and across the longitudinal medical history of the patient. Since we learn time-bins centered around admission in each narrative and we also know the admission date and perhaps the discharge dates in cn1 and cn2, we can derive a coarse partial order across the medical events in cn1 and cn2 (Figure 4.4). Even if the discharge date is not known, we still know that the admission date $A1$ of cn1 is 6/03/2007 and $A2$ of cn2 is 06/17/2007. Thus, $A2 > A1$, and all the time-bins in cn2 that are on or after admission would have happened after $A2$. The overall partial ordering can now be mined for coarse temporal patterns between symtoms, diseases, medications and tests for applications like adverse drug reaction mining and clinical decision making.

The simplified medical event "point" representation allows learning of a highly accurate coarse temporal relations, using gold standard annotations for medical events, as seen in Table 4.2. However, even if medical event annotations are unavailable, we demonstrate that the same methodology could be applied after extracting medical events using MetaMap [Aronson, 2001] and temporal expressions using TimeText [Zhou et al., 2006]. Shown in Table 4.2, are the results for assigning medical events to time-bins using automatic tools for annotating medical events and temporal expressions. We observe that the sequence tagger doesn't perform as well as it does with the gold-standard annotations for these elements. This is because MetaMap has certain limitations in terms of mapping medical events in text to UMLS concepts. While annotators map phrases in the text to contexually relevant UMLS concepts, MetaMap provides a list of possible UMLS concepts for a word or a phrase, which may not always be contexually relevant. Further, if we depend only on

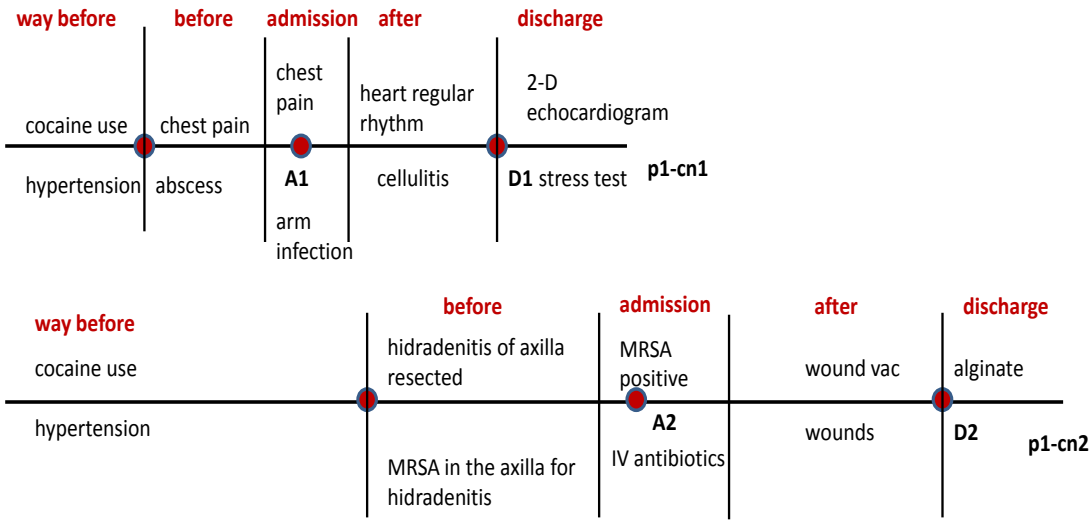


Figure 4.4: Medical events in clinical narratives cn1 and cn2 for patient p1 assigned to time-bins. A1 is the admission date in cn1 and D1 is the discharge date. Similarly A2 is the admission date in cn2 and D2 is the discharge date. Thus, we have, $A1 < D1, D1 < A2, A2 < D2$

TimeText for temporal cues, and associate the cues to medical events based on some heuristics, it increases the margin of error. This is because the correct temporal expression may or may not get anchored to a particular medical event. Inspite of these limitations, the automatic feature provide a reasonable accuracy of 79.6% precision and 69.7% recall. This is promising as with newer tools for medical concept extraction, such as HealthTermFinder[13] by Columbia, the gap between medical concept extraction with gold-standard vs. automatic feature extraction may be bridged.

[13] http://projects.dbmi.columbia.edu/nlp/healthtermfinder

4.7 Conclusion

We investigate the task of classifying medical events in clinical narratives to coarse time-bins. We describe document structure based, lexical and temporal features in clinical text and explain how these feature are useful in time-binning medical events. The extracted feature-set when used in a sequence tagging framework with CRFs gives us high accuracy when compared with a section-based baseline or a MaxEnt model. We also experimented with hand-tagged vs. automatically extracted features for this task and observe that while automatically extracted features show promising results, they are not as good as using hand-tagged features for this task.

The learned time-bins can be used as an informative feature for tasks such as medical event coreference resolution (Chapter 5) and fine-grained temporal ordering of medical events (Chapter 6). The time-bins also allow us to generate a coarse partially ordered timeline of medical events across the patient's history.

The next logical step would be to deepen the granularity of temporal relation learning and order medical events within time-bin in each clinical narrative. The characteristic of information redundancy in clinical narratives implies that there may be multiple references to the same medical event in the same time-bin (provided the classification has be learned accuractely). Resolving these references and identifying coreferring medical events in time-bins helps associate medical events with more temporal features (if multiple mentions of the same medical events co-occur with temporal expressions that help resolve their time of occurence) and perform better fine-grained temporal ordering. To this end, we now address the problem of medical event coreference resolution in the next chapter.

CHAPTER 5: COREFERENCE RESOLUTION IN CLINICAL TEXT

Information redundancy is a fundamental concept that is essential to automated information integration, relationship learning between entities, and inference. In the context of the electronic health record, redundant information arises both within and across clinical data sources. The tendency to copy and paste an old clinical note and edit parts of it whenever a new note is generated, gives rise to multiple mentions of the same medical event across notes. Moreover, the tendency to summarize past information in the context of newer medical events also results in multiple mentions of the medical event both within and across clinical narratives. The ability to resolve multiple mentions of the same medical event not only helps identify unqiue medical events in the patient's history, but also acts as a useful anchor in infereing relationships between medical events across narratives.[14]

5.1 Introduction

Coreference resolution in clinical text refers to the problem of identifying all medical events that refer to the same medical event. Medical events are a cover-term for medical concepts including entities, events or states associated with the patient's medical condition and healthcare. These include medical conditions, drugs administered, diseases, procedures

[14]This work has been previously published in NAACL 2012. P. Raghavan, E. Fosler-Lussier, and A. Lai, "Exploring Semi-Supervised Coreference Resolution of Medical Concepts using Semantic and Temporal Features," North American Association for Computational Linguistics Annual Meeting - Human Language Technologies Conference (NAACL HLT), 2012.

and lab tests, as well as normal health situations like pregnancy affecting the patient's health.

The large number of clinical narratives generated per patient, adds to the complexity of the challenge. Consider the example of a patient suffering from *chronic lymphocytic leukemia* (CLL). Such a patient may have been admitted to the hospital or have visited the clinic numerous times over the years. Every hospital stay leads to the generation of clinical narratives documenting patient history, medical conditions on admission, progress during the hospital stay, discharge and possibly an assessment plan after discharge. Redundant medical events may be found across various clinical narratives describing the patient's medical history, or when a physician describes lab and radiology results in the context of the present illness. Extracting and unambiguously resolving such clinical references to the same medical condition, diagnosis or procedure is extremely important in processing clinical text for various clinical applications.

Machine learning models for addressing this problem require gold-standard annotations for medical event coreferences for training and evaluation purposes. However, obtaining coreference annotations within and across clinical narratives is a tedious and time-consuming task. Moreover, since the annotations are credible only when marked by medical domain experts (physicians, medical or nursing students), this places more constraints on the annotation process. One way to address this problem is to try and leverage limited annotated data to train models for coreference resolution in a semi-supervised manner. Thus, in this chapter, we explore the application of certain semi-supervised resolution models for coreference resolution and compare its performance to using a supervised learning model.

5.2 Contributions

We investigate the task of resolving references to the same medical event in the clinical narratives of a patient using supervised and semi-supervised methods. Our main contributions are as follows:

- Since manual coreference annotation of patient narratives is a slow and expensive process and publicly available datasets are difficult to acquire (at the time of this work), we study the application of semi-supervised methods, co-training and using expectation constraints with posterior regularization, to medical event coreference resolution.

- We work with the hypothesis that if two medical events have the same meaning and have occurred at the same time, there is a very high probability that they corefer. Based on this hypothesis, we explain extraction of semantic and temporal feature sets that are effectively used for medical event coreference resolution.

- We demonstrate that the semi-supervised methods perform comparably with supervised learning for pairwise medical event coreference using a MaxEnt classifier, with the help of corpora created from the New England Journal of Medicine (NEJM) and clinical narratives obtained from the Ohio State University Wexner Medical Center.

5.3 Related Work

Coreference resolution is a well-studied problem in computational linguistics [Ng, 2010; Raghunathan et al., 2010a; Soon et al., 2001]. There has been recent interest in coreference resolution in the clinical domain with standard supervised approaches to noun phrase coreference resoltuion [Soon et al., 2001] being applied to medical events and anaphora

[He, 2007; Zheng et al., 2012]. Medical NLP systems like Mayo's cTakes [Savova et al., 2010b], IBM's MedKAT,[15] and MedLEE [Chiang et al., 2010], have components specifically trained or designed for the clinical domain, to support tasks such as named entity recognition. However, other than cTakes, which recently introduced a module for coreference resolution, none of the other systems provide solutions for this problem. Recently, the i2b2 challenge[16] on coreference resolution examined coreference resolution in clinical data. The problem addressed in our paper is similar to the task described in the i2b2 challenge.[17] However, the participating systems are not available publicly for comparison.

A disadvantage of supervised methods is the need for an unknown amount of annotated training data for optimal performance. We investigate the applicability of two weakly supervised methods, co-training [Blum and Mitchell, 1998] and posterior regularization [Ganchev et al., 2010] to the task of medical event coreference resolution using semantic and temporal views. We annotate a corpus of clinical narratives to tag medical events, temporal relations, and coreference information. We use this corpus as a gold standard to evaluate the proposed approach to resolving coreferences between medical events in clinical text. Creating annotated clinical corpora is tedious, time consuming, and costly, as it requires experts with medical domain knowledge. Thus, the ability to train semi-supervised models with limited labeled data for medical event coreference resolution would be of tremendous value to the clinical community.

To summarize, we study the problem of intra and cross-narrative coreference resolution on longitudinal patient data using relatedness between medical events in terms of semantics

[15] https://cabig-kc.nci.nih.gov/Vocab/KC/index.php/OHNLP

[16] https://www.i2b2.org/NLP/Coreference/

[17] https://www.i2b2.org/NLP/Coreference/assets/CoreferenceGuidelines.pdf

and time. Further, we importantly demonstrate that this task gives us reasonable results even when modeled as a semi-supervised problem.

5.4 Medical Event Coreference Resolution

Problem Formulation. Consider a corpus of clinical narratives, where multiple clinical narratives are associated with each patient. If P_i, $i \in \{1, 2, ..., n\}$ where n is the number of patients in corpus, then for each P_i, we have a set of associated clinical narratives. Each clinical narrative in turn has a set of medical events. Thus, each P_i has a set of associated medical events, $M = \{M_1, M_2, M_3, ..\}$ that occur within each clinical narrative as well as across clinical narratives for that P_i. We study the problem of medical event coreference resolution of all medical events in M for each P_i. The pipeline consisting of semantic and temporal feature extraction and application of semi-supervised learning algorithms is illustrated in Figure 5.1. We describe in detail the components of the pipeline in the following sections.

5.4.1 Semantic and Temporal Features

We extract features based on semantic and temporal relatedness for each pair of medical events. Semantic relatedness measures closeness between medical events in terms of their meaning. This is quantified by measuring distance between medical events in the UMLS Metathesaurus graph structure [Xiang et al., 2011]. Temporal relatedness measures the closeness between medical events in terms of when they occurred. This is achieved by first learning to assign every medical event to a time-bin, and then using the time-bin as a feature for learning to resolve coreferences. Extracting semantic and temporal features helps identify conditionally independent views of the data for co-training classifiers. As

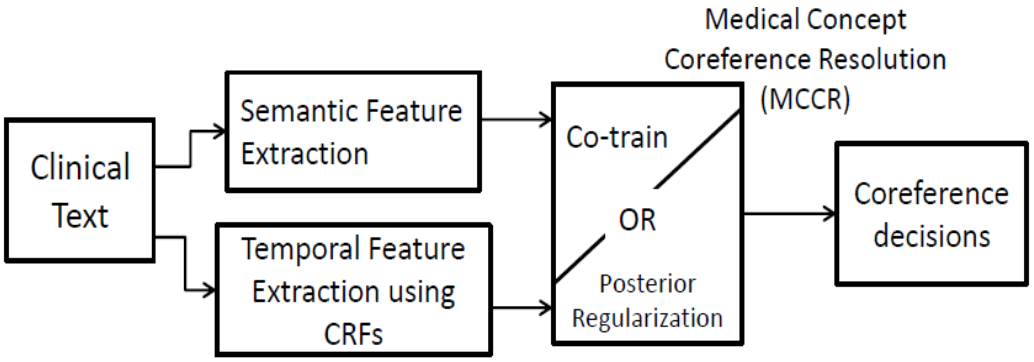


Figure 5.1: Medical event coreference resolution pipeline: Extract semantic and temporal features from clinical text to train MaxEnt classifiers using 1) Co-training or 2) Posterior Regularization

previously noted by [Nigam and Ghani, 2000], it is hard to identify conditionally independent views for real-data problems. However, we believe there are no natural dependencies between the semantic and temporal feature sets. While semantic features help identify synonymous medical events, that alone may not guarantee coreference. Medical events that are similar in meaning, but dissimilar in terms of their time of occurrence, most probably do not corefer. Similarly, medical events that occur during the same time duration but are dissimilar in terms of meaning, most probably do not corefer.

Semantic Relatedness. We leverage the UMLS to derive a semantic relatedness score between medical events. The UMLS codifies concepts found in various medical vocabularies (e.g., ICD[18] and SNOmedical eventD-CT[19]) and includes relationships between various concepts. The medical events and their relationships are modeled in a graph structure.

We use the k-Neighborhood decentralization method (kDLS) [Xiang et al., 2011] to index and transitively traverse associated relations between concept unique identifiers (CUIs) in the UMLS graph. The UMLS uses semantic relations to mark the available links between

[18] http://www.cdc.gov/nchs/icd.htm

[19] http://www.ihtsdo.org/snomed-ct/

two concepts. Around 2,404,937 CUIs and 15,333,246 links between them are seen in the full UMLS graph structure. The kDLS method is shown to outperform both breadth-first and depth-first search in terms of speed and various other measures in finding important information, such as reachability, distance, and a summary of paths, between two concepts in the UMLS graph structure. The relation between two concepts M_j (denoted by x) and M_k (denoted by y) is measured as follows.

$$R(x,y) = \sum_{p \in D_{(x,y)}} \frac{1}{\gamma^{length(p)-1}} + \sum_{q \in D_{(y,x)}} \frac{1}{\gamma^{length(q)-1}}$$

where $D(x,y)$ is the set of paths from x to y and $D(y,x)$ is the set of paths from y to x obtained using the kDLS method, excluding paths with length equal to 1. In order to make the measurement between medical events unbiased against the available links in the UMLS that directly connect them, the paths with length 1 (direct connection between the two concepts in the UMLS) between them are not counted. Each path's contribution to the relation score $R(x,y)$ is determined by its length and γ. γ is varied between 1 to 50; if γ is set to 1, then all paths contribute equally to R irrespective of their lengths. When γ increases, more weight will be placed on the short paths as opposed to the long paths. [Xiang et al., 2011] observe several fold enrichment values when γ is varied between 5 and 15.

Besides traversing the UMLS graph structure using the kDLS method to obtain a similarity score between medical events, we also measure similarity between medical events by taking into account the surrounding context. We do so by measuring the KL-divergence between the sentences to which the medical events belong. In order to avoid the possibility of an empty set when calculating the intersection of the probability distributions, we use a smoothing method that makes the probability distributions sum to 1 [Brigitte, 2003].

Another important semantic feature is the type of relation between the medical events. This feature is calculated by first computing the stemmed word overlap between the medical events and deriving features based on exact and partial matches between the word stems of the medical events. If there is no exact or partial match between the concepts, we query the UMLS to check if the stem of one of the medical events occurs in the UMLS definition or atoms of the other medical event. An atom is the smallest unit of naming within the UMLS. A medical event in UMLS represents a single meaning and contains all atoms in the UMLS that express that meaning in any way, whether formal or casual, verbose or abbreviated. All the atoms within a concept are synonymous. Besides the described features, we also include the UMLS semantic category of each medical event and the WordNet[20] similarity score between sentences containing the medical event.

Temporal Relatedness. Clinical text is frequently characterized by temporal expressions co-occurring with medical events [Zhou and Hripcsak, 2007]. For instance, *two days ago*, fever started *4 days before* rash, *July 10th, 2010* etc. The ability to associate medical events with temporal expressions helps order medical events and determine potential temporal overlap between them. This in turn could be a powerful discriminatory feature in medical event coreference resolution. Consider the medical event *chest pain* that occurs multiple times in a clinical narrative. If these mentions of *chest pain* have occurred at the same time, there is a possibility that they all refer to the same instance of the medical event *chest pain*.

Instead of relying on implicit temporal references that may or may not be evident from the clinical narrative, we focus on temporal expressions that are found in most clinical narratives. We do so by leveraging structural properties of clinical narratives such as section information and explicit temporal information such as admission and discharge dates, to

[20] http://wordnet.princeton.edu/

learn to assign medical events to time periods we refer to as time-bins (Chapter 2). The list of features extracted for the task of medical event coreference resolution include the following:

- Verb pattern in the sentence in which the medical event occurs.
- Last verb before the medical event in the same sentence.
- Type of clinical narrative.
- Section under which the medical event is mentioned.
- Position of the medical event.
- Dates that fall in the same sentence as the medical event.
- Difference between admission date and the date in the same sentence as the clinical narrative.
- The learned time-bin of each medical event. We also derive features based on the overlapping in time-bins for the medical event pair and the nature of time-bin (past, present, future).
- Difference in verb patterns in the sentences of the medical event pair.
- Difference in dates between the medical event pair.
- UMLS relatedness score between the medical event pair and all the UMLS related and other features described previously in the semantic relatedness section.

5.5 Weakly Supervised Learning

Co-training. We co-train two MaxEnt classifiers, one each on the semantic features f_s and temporal features f_t of the data, to classify pairs of medical events as *corefer* or *no-corefer* in a semi-supervised fashion. We use the co-training algorithm proposed by [Blum and Mitchell, 1998].

The assumption here is that each feature set contains sufficient information to train a model for classification of medical events. Consider the concept pair, {*renal inflammation, posterior uveitis*} that corefer. The semantic view for this concept pair may not strongly indicate coreference. The "UMLS relation type" feature indicates that the two concepts are not similar in meaning. However, both concepts are mapped to the same time-bin *after admission*. Thus, the time-bin along with features extracted based on explicit temporal expressions co-occurring with the medical events indicate a coreference between the pair of medical events. Similarly, the semantic view is confident about the coreference of certain medical event pairs which do not occur in the same time-bin. The classifiers trained on each view complement each other in the learning process. Thus, we can leverage the predictions made by each classifier on the unlabeled dataset to augment the training data of both classifiers.

The co-training algorithm is shown in Figure 5.2. We set a threshold for an unlabeled sample to be added into the labeled pool. An unlabeled sample is labeled in a particular iteration, if *classifier confidence* $>$ *1/number of labels*. In the next iteration, randomly pick a subset of unlabeled samples and label all samples in this subset. This could include samples that have already been labeled in previous iterations. A label is assigned in a subsequent iteration if: the sample was previously labeled OR if *classifier confidence* $>$ *threshold*. The parameters in this algorithm are the number of iterations, the pool size of examples selected from the unlabeled set in each iteration and the number of labeled examples added at each iteration to the labeled data pool. Similar to Blum and Mitchell [1998], we update the pool size by $2p + 2n$ in each iteration, where p is the number of medical pairs that corefer and n is the number of medical event pairs that do not corefer.

Function coTrain

Repeat till all unlabeled data is labeled.

1. Train classifier c_1 on t_{fs} to obtain model m_1
2. Train classifier c_2 on t_{ft} to obtain model m_2
3. Use m_1 to classify a subset of unlabeled data and update the training data as,
   t_{fs}.subset = $\{u_{subset1}$, *predicted label*$\}$
   iff classifier confidence $>$ *1/number of labels*
4. Use m_2 to classify a subset of unlabeled data and update the training data as,
   t_{ft}.subset = $\{u_{subset2}$, *predicted label*$\}$
   iff classifier confidence $>$ *1/number of labels*
5. $t_{fs} = t_{fs} + t_{ft}$.subset + $\{u_{subset1}$, *predicted label*$\}$
6. $t_{ft} = t_{ft} + t_{fs}$.subset + $\{u_{subset2}$, *predicted label*$\}$

Figure 5.2: Co-training [Blum and Mitchell, 1998] for the binary pairwise classification task of medical event coreference resolution.

c = classifier, u = unlabeled data.

$u_{subset1}$, $u_{subset2}$ = subsets of unlabeled data.

$u_{subset1}$ and $u_{subset2}$ are mutually exclusive.

$F = \{f_s, f_t\}$ is the features space divided into conditionally independent semantic and temporal feature sets.

$t_{fs} = \{f_s, l\}$ training data consisting of semantic features of a medical event pair along with class label.

$t_{ft} = \{f_t, l\}$ training data consisting of temporal features of a medical event pair along with class label.

Posterior Regularization. The next semi-supervised method applied to medical event coreference resolution is MaxEnt with posterior regularization using expectation constraints [Ganchev et al., 2010]. This method incorporates prior knowledge directly on the output variables during learning. The prior knowledge is expressed as inequalities on the expected value under the posterior distribution of user-defined constraint features. Thus, posterior regularization incorporates side-information into unsupervised estimation in the form of constraints on the model's posteriors. It is similar to the EM algorithm during learning, but solves a problem similar to MaxEnt inside the E-Step to enforce the constraints.

Posterior regularization is used to derive a multi-view learning algorithm while specifying constraints that the models should agree on the label distribution. We train MaxEnt models based on two views of the data, semantic and temporal. This method starts by considering the setting of complete agreement where there is a common desired output for the two models and each of the two views is sufficiently rich to predict labels accurately. The search is restricted to model pairs p_1, p_2 that satisfy $p_1(y|x) \approx p_2(y|x)$, where p_1 and p_2 each define a distribution over labels. The product distribution $p_1(y_1)p_2(y_2)$ is considered and constraint features are defined such that the proposal distribution $q(y_1, y_2)$ will have the same marginal for y_1 and y_2. There is one constraint feature defined for each label y given by, $\phi_y(y_1, y_2) = \delta(y_1 = y)\delta(y_2 = y)$, where $\delta(.)$ is the 0-1 indicator function. The constraint set $Q = q : Eq[\phi] = 0$ requires that the marginals over the two output variables are identical $q(y_1) = q(y_2)$. An agreement between two models is defined as $agree(p_1, p_2) = argmin\ KL(q(y_1, y_2)||p_1(y_1)p_2(y_2))$ | Eq [ϕ] = 0.

In the semantic feature set, we convert the following feature (described in Section 5.4.1) into expectation constraints. The type of relation between the pair of medical events, is derived from matching the word stems and querying the UMLS definition and atoms of

the medical events. Based on the relation between the medical events (i.e., partial match, complete match, UMLS definition match, UMLS atom match, and no match), we indicate the probability of label distribution coref and no-coref. If the relation turns out to be no match, there is a high probability that the medical events do not corefer. In the temporal feature set, we convert the features based on time-bins of the medical events in the pair into expectation constraints.

5.6 Experiments

We use two annotated clinical corpora for our experiments. The first is a small corpus of 6 chronic lymphocytic leukemia (CLL) case reports extracted from the New England Journal of Medicine (NEJM) annotated with 722 medical events. The other corpus is the annotator-generated gold-standard corpus described in Chapter 3, Section 3.9. The clinical narratives include discharge summaries, radiology and pathology reports. The NEJM case report is similar to a clinical discharge summary, however it is far more logically coherent and less noisy as it is consists of carefully written journal articles.

The first step involves extraction of semantic and temporal features for the annotated medical events, as described in Section 5.4.1 from both corpora. The semantic relatedness scores are computed using the kDLS [Xiang et al., 2011] method to calculate the relationship between concepts in the UMLS with value of γ set to 7. The type of relation between medical events is derived by matching word stems in each medical event using the Lucene[21] implementation of the Porter stemming algorithm. We query the UMLS Metathesaurus (UMLS 2011AB) for finding a match between medical event and the UMLS definition or

[21] http://lucene.apache.org/

UMLS atoms. The WordNet similarity score is computed using Java API for WordNet Searching (JAWS).[22]

Explicit temporal expressions annotated in the corpora are included in our temporal feature set. Medical events in the NEJM are mostly described temporally relative to the patient's admission. Temporal expressions like "3 weeks before admission" are common. Hence, we use a algorithm parses case reports and identifies the temporal expressions anchored to admission. All medical events following such a temporal expression are anchored to it until a new temporal expression is encountered. Over 88% of the medical event-temporal expression associations done with the algorithm above is accurate when compared with the NEJM gold standard.

We use the learned time-bins (Chapter 4) as a temporal feature. The percentage of medical events that fall under time-bins "way before admission" and "on admission" are less than 5%, affecting the learning accuracy of those classes. When modeled as a multi-class classification task using MaxEnt, we achieve 86% accuracy.

5.7 Results and Discussion

Class	NEJM		Clinical Narratives	
	Precision	Recall	Precision	Recall
coref	79.24	94.53	74.81	88.33
no-coref	86.71	90.62	83.92	94.86

Table 5.1: Supervised learning for medical event coreference resolution.

[22] http://lyle.smu.edu/~tspell/jaws/

Class	NEJM		Clinical Narratives	
Co-train	Precision	Recall	Precision	Recall
coref	70.32	82.54	69.26	87.31
no-coref	82.54	84.85	71.15	89.44
PR	Precision	Recall	Precision	Recall
coref	76.63	90.41	74.81	84.25
no-coref	80.35	89.21	78.93	87.46

Table 5.2: Co-training and posterior regularization (PR) for medical event coreference resolution using semantic and temporal feature sets.

We perform the following experiments for medical event coreference resolution: (i) Supervised learning with a MaxEnt classifier, using the combined semantic and temporal feature set, (ii) Co-training two MaxEnt models, (iii) Training MaxEnt models with using posterior regularization.

We use the MaxEnt classifier available in Mallet for (i) and (ii) and the the Mallet implementation of MaxEnt models with posterior regularization for (iii).

From all the candidate pairs in the clinical narrative corpus, 1025 pairs corefer. We randomly sample the no-coref instances to reduce the bias towards negative instances. The results for all 3 experiments for both corpora is shown in Tables 5.1, 5.2. We also train-test a supervised MaxEnt classifier on a 60-40 split of the entire corpus. This gives us a precision of 74.81% and 88.33% recall (coref) for the binary classification task of pairwise medical event coreference resolution in the clinical narratives corpus. In the both the semi-supervised experiments, we use an initial labeled pool size of 30 where 12 medical event pairs that corefer (p) and 18 that do not corefer (n). The growth size of each iteration of co-training is $2p+2n$. At each iteration, confidently labeled examples are added to the training set from the previous iteration. The co-training algorithm is run until all unlabeled instances

become labeled. The parameters in the posterior regularization implementation include the regularization penalty for each step and the number of iterations. We use the default values (maxIterations=100, pGaussianPriorVariance=0.1, qGaussianPriorVariance=1000) suggested on the Mallet toolkit page [Bellare et al., 2009]. Co-training two MaxEnt models based on independent semantic and temporal views of the data results in 69.26% precision and 87.31% recall (coref), whereas training MaxEnt models with expectation constraints gives us 74.81% precision and 84.25% recall (coref), on the corpus of clinical narratives. Posterior regularization does better than co-training and the performance of both the semi-supervised methods is comparable to the supervised classifier trained on a 60-40 split of the corpus. Thus, our results indicate that the use of semantic and temporal features is effective for medical event coreference resolution in clinical text. It is clear from the co-training and posterior regularization results that treating medical event coreference resolution as a semi-supervised problem works well as demonstrated through our experiments.

5.8 Conclusions

We investigated the task of medical event coreference resolution in clinical text using supervised and semi-supervised learning methods. We create annotated corpora of clinical text with case reports from the NEJM and narratives obtained from OSUWMC. We work with the hypothesis that determining semantic and temporal similarity between medical events helps resolve coreferences. In order to test this hypothesis, we describe the process of semantic and temporal feature extraction from clinical text. We demonstrate the effectiveness of the extracted features in a supervised binary classification task for medical event coreference resolution with MaxEnt classifiers (using the combined feature set) as well as using semi-supervised methods of co-training MaxEnt classifiers and training

MaxEnt models using posterior regularization (using two independent views of the data - semantic view and temporal view). Thus, we show that medical event coreference resolution can be performed using semi-supervised learning with semantic and temporal views of the data.

Resolving coreferences is critical to the intra- and cross-narrative temporal ordering tasks. Chapter 6 uses coreference information as a feature in learning temporal relations within each clinical narrative. Coreference implicitly entails the *simultaneous* temporal relation and no-coreference entails relationships like *before, after*. Frequently, coreference information is the only useful indicator that helps determine cross-narrative temporal ordering as seen in Chapter 7.

CHAPTER 6: INTRA-NARRATIVE TEMPORAL ORDERING

Reasoning about temporal relationships in clinical text could be done at different levels of temporal granularity. We began by learning temporal relations within a clinical narratives by learning to assign medical events to coarse time-bins in Chapter 4. Now, with the help of these learned time-bins and coreference information (Chapter 5), we enable fine-grained temporal relation learning between medical events within the same clinical narrative. In doing so, we also demonstrate the need for novel NLP methods that are better suited to addressing problems in the clinical domain, by comparing our methods on both a corpus of clinical narratives and Timebank [Pustejovsky et al., 2003a].[23]

6.1 Introduction

There has been considerable research on learning temporal relations between events in natural language. Most learning problems try to classify event pairs as related by one of Allen's temporal relations [Allen, 1981], i.e., *before, simultaneous, includes/during, overlaps, begins/starts, ends/finishes* and their inverses [Mani et al., 2006]. The Timebank corpus, widely used for temporal relation learning, consists of newswire text annotated for events, temporal expressions, and temporal relations between events using TimeML [Pustejovsky et al., 2003a]. In Timebank, the notion of an "event" primarily consists of

[23]This work has been published in ACL 2012. P. Raghavan, E. Fosler-Lussier, and A. Lai, "Learning to Temporally Order Medical Events in Clinical Text," (Short Paper) Association for Computational Linguistics Annual Meeting (ACL), 2012.

verbs or phrases that denote change in state. This varies from the notion of a "medical event" which mainly consists of noun phrases and nominals.

We study the problem of learning temporal relations between medical events in clinical text. The idea of a medical "event" in clinical text is very different from events in Timebank. Medical events are temporally-associated concepts in clinical text that describe a medical condition affecting the patient's health, or procedures performed on a patient. Learning to temporally order events in clinical text is fundamental to understanding patient narratives and key to applications such as longitudinal studies, question answering, document summarization and information retrieval with temporal constraints. We propose learning temporal relations between medical events found in clinical narratives by learning to rank them. This is achieved by representing medical events as time durations with starts and stops and ranking them based on their proximity to the admission date.[24] This implicitly allows us to learn all of Allen's temporal relations between medical events.

6.2 Contributions

Researchers have successfully demonstrated how temporal relations between events can be learned from such a corpus in a multi-class classification framework using features of events like tense, aspect and part of speech [Mani et al., 2006; Chambers et al., 2007]. However, there may be a need to rethink how we learn temporal relations between events in different domains. Timebank, its features, and established learning techniques like classification, may not work optimally in many real-world problems where temporal relation learning is of great importance. We establish the need to rethink the methods and resources used in temporal relation learning, as we demonstrate that the resources widely used for

[24]The admission date is the only explicit date always present in each clinical narrative.

learning temporal relations in newswire text do not work on clinical text. When we model the temporal ordering problem in clinical text as a ranking problem, we empirically show that it outperforms classification; we perform similar experiments with Timebank and observe the opposite conclusion (classification outperforms ranking).

Moreover, the fine-grained intra-narrative temporal ordering along with explicit dates in the narrative can be used to generate a partially ordered timeline across all clinical narratives for a patient. This timeline may be sufficient to help clinical applications that do not require fine-grained resolution of cross-narrative temporal relationships.

6.3 Related Work

The Timebank corpus provides hand-tagged features, including tense, aspect, modality, polarity and event class. There have been significant efforts in machine learning of temporal relations between events using these features and a wide range of other features extracted from the Timebank corpus [Mani et al., 2006; Chambers and Jurafsky, 2008; Lapata and Lascarides, 2006]. The SemEval/TempEval [Verhagen et al., 2009] challenges have often focused on temporal relation learning between different types of events from Timebank. Zhou and Hripcsak [2007] provide a comprehensive survey of temporal reasoning with clinical data. There has also been some work in generating annotated corpora of clinical text for temporal relation learning [Roberts et al., 2008; Savova et al., 2009]. However, none of these corpora are freely available. Zhou et al. [2006] propose a Temporal Constraint Structure (TCS) for medical events in discharge summaries. They use rule-based methods to induce this structure.

We demonstrate the need to rethink resources, features and methods of learning temporal relations between events in different domains with the help of experiments in learning

temporal relations in clinical text. Specifically, we observe that we get better results in learning to rank chains of medical events to derive temporal relations (and their inverses) than training a classifier for the same task.

The problem of learning to rank from examples has gained significant interest in the machine learning community, with important similarities and differences with the problems of regression and classification [Joachims et al., 2007]. The joint cumulative distribution of many variables arises in problems of learning to rank objects in information retrieval and various other domains. To the best of our understanding, there have been no previous attempts to learn temporal relations between events using a ranking approach.

6.4 Learning Temporal Relations using Ranking

We model the temporal relation learning problem as a ranking task, where the rank of a medical event corresponds to its relative temporal order in the clinical narrative. To enable the use of a ranking model, it is first important to represent medical events appropriately. Given the point-based notation for medical events used in Chapter 4, if we use a ranking-based model to rank each point (medical event) by its relative time of occurrence, we will be able to learn only the temporal relations *before, after, simultaneous* between medical events. In order to enable the learning all of Allen's temporal relations, i.e. *before, after, starts with, finishes with, during, overlaps* (and their inverses) between medical events using a ranking model, we describe an alternate interval-based event representation.

6.4.1 Representation of Medical Events

Clinical narratives contain unstructured text describing various medical events including conditions, diagnoses and tests in the history of a patient, along with some information on when they occurred. Much of the temporal information in clinical text is implicit and

embedded in relative temporal relations between medical events. Medical events are temporally related both qualitatively (e.g., *paresis before colostomy*) and quantitatively (e.g. *chills 1 month before admission*). Relative time may be more prevalent than absolute time (e.g., *last 1 month, post colostomy* rather than *on July 2007*). Temporal expressions may also be fuzzy where *history* may refer to an event *1 year ago* or *3 months ago*. The relationship between medical events and time is complicated. Medical events could be recurring or continuous vs. discrete date or time, such as *fever* vs. *blood in urine*. Some are long lasting vs. short-lived, such as *cancer, leukemia* vs. *palpitations*.

We represent medical events of any type in terms of their time duration. The idea of time duration based representation for medical events is in the same spirit as the Temporal Constraint Structure(TCS) [Zhou et al., 2006]. We break every medical event *me* into *me.start* and *me.stop* (in essence making every event 2 events). Given the ranking of all starts and stops, we can now compose every one of Allen's temporal relations [Allen, 1981]. If it is clear from context that only the start or stop of a medical event can be determined, then only that is considered. For instance, *"history of paresis secondary to back injury who is bedridden status post colostomy"* indicates the start of *paresis* is in the past history of the patient prior to *colostomy*. Its exact date or stop time (if any) may not be clear from a single narrative. We only know about *paresis.start* relative to other medical events and may not be able to determine *paresis.stop*. For recurring and continuous events like *chills* and *fever*, if the time period of recurrence is continuous (*last 1 month*), we consider it to be the time duration of the event. If not continuous, we consider separate instances of the medical event. For medical events that are associated with a fixed date or time, the start and stop are assumed to be the same (e.g., *polymicrobial infection in the blood as well as in the urine* in July 2007). In case of negated events like *no cough*, we consider *cough* as the

medical event with a negative polarity. Its start and stop time are assumed to be the same. Polarity allows us to identify events that actually occurred in the patient's history.

6.4.2 Data Characteristcs and Feature Generation

Given a patient with multiple clinical narratives, our objective is to induce a partial temporal ordering of all medical events in each clinical narrative based on their proximity to a reference date (admission).

Data Characteristics. The training data consists of medical event chains, where each chain consists of an instance of the start or stop of a medical event belonging to the same clinical narrative along with a rank. The assumption is that the medical events in the same narrative are more or less semantically related by virtue of narrative discourse structure and are hence considered part of the same medical event chain. The rank assigned to an instance indicates the temporal order of the event instance in the chain. Multiple medical events could occupy the same rank. Based on the rank of the starts and stops of event instances relative to other event instances, the temporal relations between them can be derived as indicated in Figure 6.1. Our corpus for ranking consisted of 91 clinical narratives obtained from the medical center and annotated with medical events, temporal expressions, relations and event chains.

Thus, we extracted 91 medical event chains across 7 patients. The distribution of medical events across event chains and chains across patients (p) is as as follows. p1 had 5 chains with 125 medical events, p2 had 9 chains with 90 medical events, p3 had 20 chains with 488 medical events, p4 had 13 chains with 318 medical events, p5 had 8 chains with 220 medical events, p6 had 10 chains with 136 medical events, and p7 had 15 chains with 297 medical events.

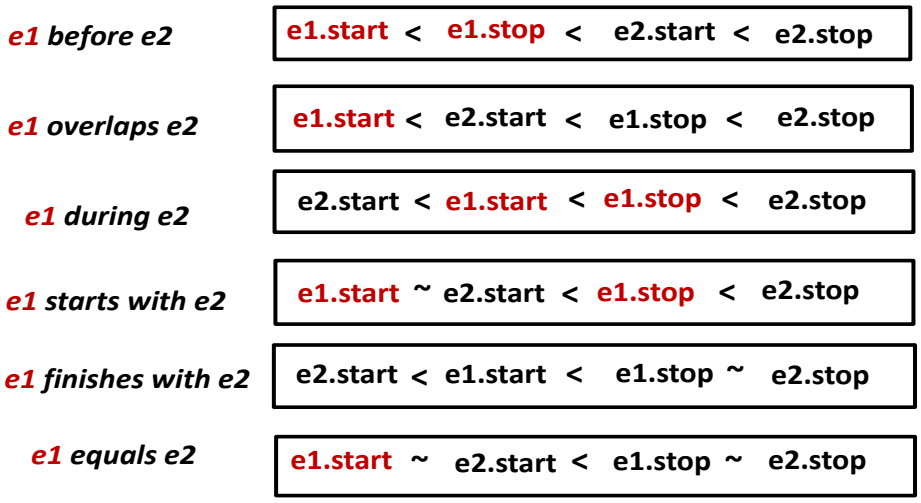


Figure 6.1: The start/ stop notation allows learning temporal relations between events by ranking the starts and stops using *before* ($<$), *after* ($>$) and *simultaneous*($\sim$) relations. This also maps to learning pairwise ranking constraints between the medical events.

Feature Generation. We construct a vector of features, from the manually annotated corpus, for each medical event instance. Although there is no real query in our set up, the admission date for each chain can be thought of as the query "date" and the medical events are ordered based on how close or far they are from each other and the admission date. The features extracted for each medical event include the the type of clinical narrative, section information, medical event polarity, position of the medical concept in the narrative and verb pattern. We extract temporal expressions linked to the medical event like *history, before admission, past, during examination, on discharge, after discharge, on admission.* Temporal references to specific times like *next day, previously* are resolved and included in the feature set. We also extract features from each temporal expression indicating its closeness to the admission date. Differences between each explicit date in the narrative is

also extracted. The UMLS [Bodenreider, 2004] semantic category of each medical concept is also included based on the intuition that medical events of a certain semantic group may occur closer to admission. We tried using features like the tense of medical event or the verb preceding the medical event (if any), part-of-speech (POS) tag in ranking. We found no improvement in accuracy upon their inclusion.

A list of the features used are as follows.

- Verb pattern in the sentence in which the medical concept occurs.
- Last verb before the medical concept in the same sentence.
- Type of clinical narrative.
- Section under which the medical concept is mentioned.
- Section under which the medical concept is mentioned.
- Position of the medical concept within the section.
- Sentence number of the medical concept in the entire clinical narrative
- Temporal expressions linked to the medical event. Examples include *history, before admission, past, during examination, on discharge, after discharge, on admission.* Temporal references to specific times like *next day, previously* are resolved and included in the feature set. We also extract features from each temporal expression indicating its closeness to the admission date.
- Dates that fall in the same sentence as the medical concept.
- Difference between admission date and the date in the same sentence as the medical concept. This is the only query (admission date) dependent feature.
- UMLS semantic category of the medical event. The intuition behind this feature is that medical events of a certain semantic group may occur closer to admission.

Importantly, the learned time-bins for each medical event (described in Chapter 4) are also used to derive a set of features for each medical event. These features include binary feature for each pair of medical events indicating whether they belong to the same time-bin, rank assigned to the time-bin (*way-before admission* $<$ *before admission* $<$ *on-admission* $<$ *after-admission* $<$ *after-discharge*).

The learned coreference information is also used as a feature in order to bias the ranking model towards assigning the same rank to events that corefer, and different ranks to events that do not corefer.

6.5 Ranking Model, Experiments and Results

We ran ranking experiments using SVM-rank [Joachims, 2006], and based on the ranking score assigned to each start/stop instance, we derive the relative temporal order of medical events in a chain.[25] This in turn allows us to infer temporal relations between all medical events in a chain.

Ranking Model. SVMRank optimizes the area under a ROC curve [Marrocco et al., 2008]. The ROC curve is determined by the true positive rate vs. the false positive rate for varying values of the prediction threshold. This ranking model is then utilized to rank entities in a new sample set. Given independently and identically distributed training samples *S* of size *n* containing entities *e* (as a feature vector of m attributes) with their target ranking *r** the learner will build a ranking model to minimize the ranking error. In linear SVM, this is equivalent to finding the weight vector so that the maximum number of the following inequalities is satisfied.

$$\forall(e_i, e_j) \in r_n^* : \bar{w}.e_i > \bar{w}.e_j \tag{6.1}$$

[25] In evaluating *simultaneous*, ± 0.05 difference in ranking score of starts/stops of medical events is counted as a match.

where $(e_i, e_j) \in r_n^*$ if e_i has been ranked higher than e_j based on the target rank r_n^*

The weight vector is first computed and then used for ranking a new sample of n new entities.

Results. The first baseline (baseline1) for this task is the temporal order of medical events as they occur in the narrative (natural reading order). We also generate a rule-based baseline (baseline2) for intra-narrative temporal ordering. We use a deterministic algorithm, based on the regular expression-based TimeText tagger [Zhou et al., 2006] to annotate temporal expressions and then anchor them to medical events using the following procedure:

(i) Parse the document in natural reading order.

(ii) On encountering a temporal expression, anchor all following medical events to this temporal expression until you encounter the next temporal expression.

(iii) Repeat (i) and (ii) until the end of the narrative.

We now adjust the temporal order in baseline1 based on any explicit and relative temporal expressions that the medical events are anchored to using the above procedure. These temporal expressions include dates, and expressions that are relative to these dates like *yesterday*, *2 days ago* and *last week*. The natural reading order based baseline1 gives us an accuracy of 47.3%. This improves to 58.9% accuracy using the rule-based baseline2.

We then ran SVM-rank on the corpus of clinical narratives described in Section 6.4.2. The ranking error on the test set is 28.2%. On introducing the time-bin feature, the ranking error drops to 16.8%. The overall accuracy of ranking medical events on including the time-bin feature is 80.2%. Each learned relation is now compared with the pairwise classification of temporal relations between medical events. We train a SVM classifier [Joachims, 1999]

Relation	Clinical Text		Timebank	
	Ranking	Classifier	Ranking	Classifier
begins	81.2	73.3	52.6	58.8
ends	76.3	69.8	61.3	82.8
simultenous	85.4	71.3	50.2	56.6
includes	83.7	74.2	59.6	60.7
before	88.3	77.1	61.3	70.4

Table 6.1: Per-class accuracy (%) for ranking, classification on clinical text and Timebank. We merge class ibefore into before.

with an RBF kernel for pairwise classification of temporal relations. The average classification accuracy for clinical text using the same feature set is 71.3%. We used Timebank (v1.1) for evaluation, 186 newswire documents with 3345 event pairs. We traverse transitive relations between events in Timebank, increasing the number of event-event links to 6750 and create chains of related events to be ranked. Classification works better on Timebank, resulting in an overall accuracy of 63.8%, but ranking gives only 55.4% accuracy. All classification and ranking results from cross validation are presented in Table 6.1.

6.6 Discussion

In ranking, the objective of learning is formalized as minimizing the fraction of swapped pairs over all rankings. This model is well suited to the features that are available in clinical text. The assumption that all medical events in a clinical narrative are temporally related allows us to totally order events within each narrative. This may work because a clinical narrative usually has a single protagonist, the patient. This assumption, along with the availability of a fixed reference date in each narrative, allows us to effectively extract features that work in ranking medical events. However, this assumption is not true in

newswire text: there tend to be multiple protagonists, and it may be possible to totally order only events that are linked to the same protagonist. Ranking implicitly allows us to learn the transitive relations between medical events in the chain. Ranking medical event starts/ stops captures relations like *includes* and *begins* much better than classification, primarily because of the date difference and time-bin difference features. However, the hand-tagged features available in Timebank are not suited for this kind of model. The features work well with classification but are not sufficiently informative to learn time durations using our proposed event representation in a ranking model. Features like "tense" that are used for temporal relation learning in Timebank are not very useful in medical event ordering. Tense is a temporal linguistic quality expressing the time at, or during which a state or action denoted by a verb occurs. In most cases, medical events are not verbs (e.g., *colostomy*). Even if we consider verbs co-occurring with medical events, they are not always accurately reflective of the medical events' temporal nature. Moreover, in discharge summaries, almost all medical events or co-occurring verbs are in the past tense (before the discharge date). This is complicated by the fact that the reference time / medical event with respect to which the tense of the verb is expressed is not always clear. Based on the type of clinical narrative, when it was generated, the reference date for the tense of the verb could be in the patient's history, admission, discharge, or an intermediate date between admission and discharge. For similar reasons, features like POS and aspect are not very informative in ordering medical events. Moreover, a feature like aspect requires annotators with not only a clinical background but also some expert knowledge in linguistics, which is not feasible.

6.7 Conclusions

Representing and reasoning with temporal information in unstructured text is crucial to the field of natural language processing and biomedical informatics. We presented a study on learning to rank medical events. Temporally ordering medical events allows us to induce a partial order of medical events over the patient's history. We noted many differences between learning temporal relations in clinical text and Timebank. The ranking experiments on clinical text yield better performance than classification, whereas the performance is the exact opposite in Timebank. Based on experiments in two very different domains, we demonstrate the need to rethink the resources and methods for temporal relation learning.

On applying the ranking methods described in this chapter to temporally order medical events, we obtain sequences of temporally ordered medical events corresponding to each clinical narrative. These sequences when combined with the explicit dates like admission, discharge dates, and note creation date can be used to create a partially ordered timeline across all narratives of the patient. However, certain applications like multiple document summarization and clinical trial recruitment may require reasoning about fine-grained temporal relationships across clinical narratives. Thus, we investigate the problem of cross-narrative temporal ordering of medical events in the next chapter.

CHAPTER 7: CROSS-NARRATIVE TEMPORAL ORDERING OF MEDICAL EVENTS

Cross-narrative temporal ordering of medical events is essential to the task of generating a comprehensive timeline over a patient's history. However, reasoning about cross-document relationships is always challenging in any domain. As Radev [2000] points out, clusters of topically related documents, containing events that have evolved over time, are a challenge for natural language processing. For applications like cross-document summarization, combining sentences written by different sources are not likely to read coherently. In fact, the lack of logical continuity and context in the text across documents is a huge barrier in effectively reasoning about cross-document relationships between events. In this chapter, we describe the process of cross-narrative temporal reasoning using coreference and temporal relations between medical events to combine medical event sequences obtained using the ranking methodology described in Chapter 6.[26]

7.1 Introduction

Discourse structure, logical flow of sentences, and context play a large part in ordering medical events based on temporal relations within a clinical narrative. Cross-narrative temporal relation ordering is a challenging task as it is difficult to learn temporal relations among medical events which are not part of the logically coherent discourse of a

[26]This work has been published in ACL 2014. P. Raghavan, E. Fosler-Lussier, N. Elhadad, and A. Lai, "Cross-narrative Temporal Ordering of Medical Events," Association for Computational Linguistics Annual Meeting (ACL), 2014.

single narrative. Resolving cross-narrative temporal relationships between medical events is essential to the task of generating an event timeline from across unstructured clinical narratives such as admission notes, radiology reports, history and physical reports and discharge summaries. Such a timeline has multiple applications in clinical trial recruitment [Luo et al., 2011], medical document summarization [Bramsen et al., 2006; Reichert et al., 2010] and clinical decision making [Demner-Fushman et al., 2009].

Given multiple temporally ordered medical event sequences generated from clinical narratives in a patient record, how can we combine the events to create a timeline across all the narratives? The tendency to copy and paste text and summarize past information in newly generated clinical narratives leads to multiple mentions of the same medical event across narratives [Cohen et al., 2013]. These cross-narrative coreferences act as important anchors for reasoning with information across narratives. We leverage cross-narrative coreference information along with confident cross-narrative temporal relation predictions and learn to align and temporally order medical event sequences across longitudinal clinical narratives. We model the problem as a sequence alignment task and propose solving this using two approaches. First, we use weighted finite state machines to represent medical events sequences, thus enabling composition and search to obtain the most probable combined sequence of medical events. As a contrast, we adapt dynamic programming algorithms [Needleman et al., 1970; Smith and Waterman, 1981] used to produce global and local alignments for aligning sequences of medical events across narratives. We also compare the proposed methods with an Integer Linear Programming (ILP) based method for timeline construction [Do et al., 2012]. The cross-narrative coreference and temporal relation scores used in both of these approaches are learned from a corpus of patient narratives from a large medical center.

7.2 Contributions

Learning of cross-document event-event relations is a relatively unexplored area. This work explores this task in the context of learning temporal and coreference relations between medical events across clinical narratives.

The main contribution of this work is a general framework that allows aligning multiple event sequences using cascaded weighted finite state transducers (WFSTs) with the help of efficient composition and decoding. Moreover, we demonstrate that this method can be used for more accurate multiple sequence alignment of medical events when compared to dynamic programming or other ILP-based methods proposed in literature.

7.3 Related Work

In the areas of summarization and text-to-text generation, there has been prior work on several ordering strategies to order pieces of information extracted from different input documents [Barzilay et al., 2002; Lapata, 2003; Bollegala et al., 2010]. In this chapter, we focus on temporal ordering of information, as discussed next.

Recent state-of-the art research has focused on the problem of temporal relation learning within the same document, and in many cases within the same sentence [Mani et al., 2006; Verhagen et al., 2009; Lapata and Lascarides, 2006]. Chambers and Jurafsky [2009] describe a process to induce a partially ordered set of events related by a common protagonist by using an unsupervised distributional method to learn relations between events sharing coreferring arguments, followed by temporal classification to induce partial order. The task was carried out on the Timebank newswire corpus, but was limited to an intra-document setting. More recently, Do et al. [2012] proposed an ILP-based method to combine the outputs of an event-interval and an event-event classifier for timeline construction

on the ACE 2005 corpus. However, this approach is also restricted to events within documents and requires annotations for event intervals. We empirically compare our methods for timeline creation from longitudinal clinical narratives to such an ILP-based approach in Section 7.8. While a lot of this work has been done in the news domain, there is also some recent work in rule-based algorithms [Zhou et al., 2006] and machine learning applied to temporal relations between medical events in clinical text [Roberts et al., 2008]. Clinical narratives are written in a distinct sub-language with domain specific terminology and temporal characteristics, making them markedly different from newswire text.

There is limited prior work in learning relations across documents. Ji and Grishman [2008] extended the one sense per discourse idea [Yarowsky, 1995] to multiple topically related documents and propagate consistent event arguments across sentences and documents. Barzilay and McKeown [2005] propose a text-to-text generation technique for synthesizing common information across documents using sentence fusion. This involves multisequence dependency tree alignment to identify phrases conveying similar information and statistical generation to combine common phrases into a sentence. Along with syntactic features, they combine knowledge from resources like WordNet to find similar sentences. In case of clinical narratives and medical event alignment, the objective is to identify a unique sequence of temporally ordered medical events from across longitudinal clinical data.

To the best of our knowledge, there is no prior work on cross-document alignment of event sequences. Multiple sequence alignment is a problem that arises in a variety of domains including gene/protein alignments in bioinformatics [Notredame, 2002], word alignments in machine translation [Kumar and Byrne, 2003], and sentence alignments for

summarization [Lacatusu et al., 2004]. Dynamic programming algorithms have been popularly leveraged to produce pairwise and global genetic alignments, where edit distance based metrics are used to compute the cost of insertions, deletions and substitutions. We use dynamic programming to compute the best alignment, given the temporal and coreference information between medical events across these sequences. More importantly, we propose a cascaded WFST-based framework for cross-document temporal ordering of medical event sequences. Composition and search operations can be used to build a single transducer that integrates these components, directly mapping from input states to desired outputs, and obtain the best alignment [Mohri et al., 2000]. In natural language processing, WFSTs have seen varied applications in machine translation [Kumar and Byrne, 2003], morphology [Sproat, 2006], named entity recognition [Krstev et al., 2011], and biological sequence alignment/ generation [Whelan et al., 2010] among others. We demonstrate that the WFST-based approach outperforms popularly used dynamic programming algorithms for multiple sequence alignment.

7.4 Problem Description

Medical events are temporally-associated concepts in clinical text that describe a medical condition affecting the patient's health, or procedures performed on a patient. We represent medical events by splitting each event into a start and a stop. When there is insufficient information to discern the start or stop of an event, it is represented as a single concept. If only the start is known then the stop is set to $+\infty$, whereas when only the stop is known , the start is set to the date of birth of the patient.[27] Temporal relations exist between the start and stop of events as shown in Figure 7.1. Following the work of [Raghavan et al.,

[27]Patient date of birth, admission / discharge date are usually available in the metadata associated with a clinical narrative.

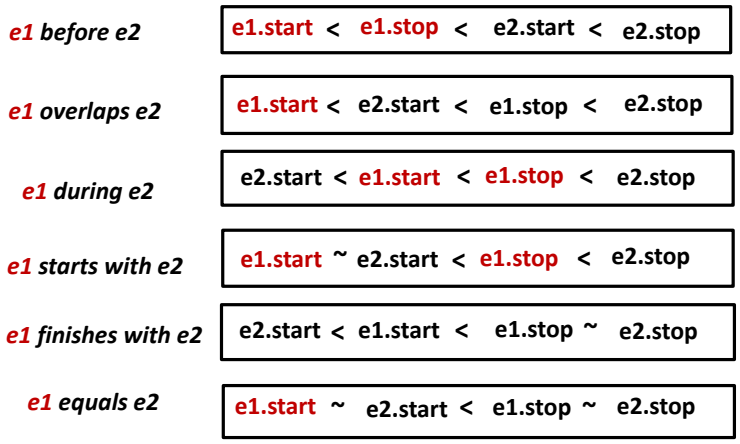


Figure 7.1: Medical event representation mapped to temporal relations. $\sim$ indicates simultaneity between the events. $e1_{start} = e2_{start}$ and $e1_{stop} = e2_{stop}$, when $e1$ and $e2$ corefer.

2012], such a representation allows us to temporally order the event starts and stops within each clinical narrative by learning to rank them in relative order of time. The problem definition is as follows:

Input: Sequences of temporally ordered medical event starts and stops. This corresponds to N_1, N_2, and N_3 in Figure 7.2. Each sequence corresponds to a clinical narrative. The total number of sequences correspond to the number of clinical narratives of a patient.

Problem: Combine medical events across these sequences to generate a timeline i.e., a single comprehensive sequence of medical events over all clinical narratives of the patient.

Expected Output: In the example shown in Figure 7.2, the output would be as follows: Timeline (N_1, N_2, N_3)= $\{$cocaine use$_{start}$ $<$ hypertension$_{start}$ $=$ hypertension$_{start}$ $<$ admission1 $<$ chest pain$_{start}$ $\sim$ palpitations$_{start}$ $<$ chest pain$_{stop}$ $<$ heart attack$_{start}$ $=$ myocardial infarction$_{start}$ $<$ admission2 $<$ infection$_{start}$ $<$ MRSA$_{start}$ $<$ admission3 $<$ wounds$_{start}\}$.

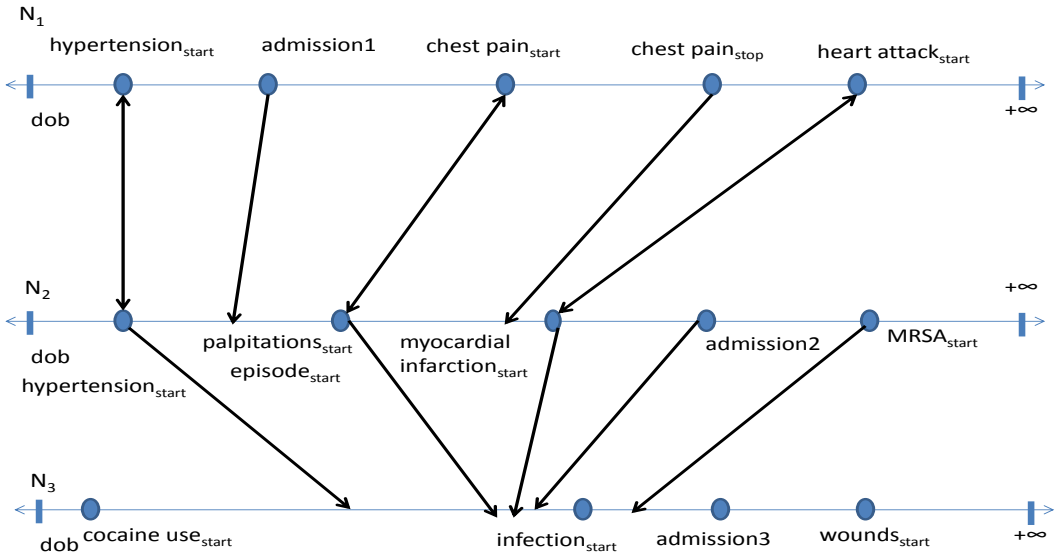


Figure 7.2: Given temporally ordered medical event sequences, N_1, N_2, N_3, we address the task of combining events across these sequences by merging or ordering them to create a single comprehensive timeline.

The goal of multiple sequence alignment is to find an alignment that maximizes an overall alignment score. Thus, in order to align event sequences, we need to compute scores corresponding to cross-narrative medical event coreference resolution and cross-narrative temporal relations.

7.5 Cross-Narrative Coreference Resolution and Temporal Relation Learning

The first approach to learning a temporal ordering of medical events across all clinical narratives is to consider all pairs of events across all narratives and learn to classify them as sharing one of Allen's temporal relations [Allen, 1981] using a single learning model. Alternatively, a ranking approach, similar to the one used to generate intra-narrative temporal

ordering, can also be extended to the cross-narrative case. However, the features related to narrative structure and relative and implicit temporal expressions used for temporal ordering within a clinical narrative may not be applicable across narratives. For instance, a history and physical report may have sections like "past medical history," "history of present illness," "assessment and plan," and a certain logical pattern to the flow of text within and across these sections. Further, temporal cues like "thereafter," "subsequently," follow from the context around an event mention. The absence of such features in the cross-narrative case does not allow such a model to generate accurate temporal relation predictions.

Thus, for use in our sequence alignment models, we train two independent classifiers for medical event coreference and temporal relation learning across narratives. We train a classifier to resolve cross-narrative coreferences by extracting semantic and temporal relatedness feature sets for each pair of medical concepts. Extracting these feature sets helps us train a classifier to predict medical event coreferences using the methods described in Chapter 5. Another classifier is then trained to classify pairs of medical event starts and stops across narratives as sharing temporal relations, i.e., {before, after, overlaps}. The learned cross-narrative coreference predictions can then be used along with confident temporal relation predictions to derive a joint probability to enable cross-narrative temporal ordering.

Sequence alignment algorithms have been developed and popularly used in bioinformatics. However, multiple sequence alignment (MSA) has been shown to be NP complete [Wang and Jiang, 1994] and various heuristic algorithms [Notredame, 2002] have been proposed to solve this problem. We propose a novel WFST-based representation that enables accurate decoding for MSA when compared to popularly used dynamic programming

algorithms [Needleman et al., 1970; Smith and Waterman, 1981] or other state of the art methods [Do et al., 2012].

In the problem of aligning events across multiple narrative sequences, we want to align temporally ordered medical events corresponding to clinical narratives of a patient. Unlike problems in biological sequence alignment where the symbols to be aligned across sequences are restricted to a fixed set, our symbol set is not fixed or certain because the symbols correspond to medical events in clinical narratives. Moreover, we cannot have fixed scores for symbol transformations since our transformations correspond to coreference and temporal relations between the medical events across sequences. The computation of these scores is described next.

7.5.1 Scoring Scheme

Let us assume a and b are medical events in the first clinical narrative and have been temporally ordered so $a < b$. Similarly, x and y are medical events in the second clinical narrative such that $x < y$. There exists a match or an alignment between a pair of medical events, across the sequences, in the following cases:

1. If the medical events are simultaneous and coreferring, denoted as $a = x$.

2. If the medical events are simultaneous and non-coreferring, denoted as $a \sim x$.

3. If the a medical event from one sequence is before a medical event from another sequence, denoted as $a < x$.

4. If the a medical event from one sequence is after a medical event from another sequence, denoted as $a > x$.

We now illustrate how the scores for candidate aligned sequences are computed using the learned cross-narrative coreference and temporal probabilities for the following three scenarios:

- The medical events across sequences are simultaneous and corefer as illustrated in Figure 7.3. For example, let the medical events $chestpain_{start}$, $chestpain_{stop}$, $epsiode_{start}$ and $epsiode_{stop}$ be denoted by a, b, x and y respectively. if $a_{start} = x_{start} < b_{stop} = y_{stop}$, then we compute the probability of the candidate sequence as *P(a simult x | a coref x) P(a coref x)*
- Some medical events across sequences are simultaneous but do not corefer as illustrated in Figure 7.4. For example, let the medical events $chestpain_{start}$, $chestpain_{stop}$, $palpitations_{start}$ and $palpitations_{stop}$ be denoted by a, b, x, y respectively. if $a_{start} \sim x_{start} < b_{stop} < y_{stop}$, then we compute the probability of the candidate sequence as, *P(a simult x | a no-coref x)* $\times$ *P(x before b | a no-coref x)* $\times$ *P(b before y | a no-coref x) P(a no-coref x)*
- The medical events across sequences are not simultaneous and do not corefer as illustrated in Figure 7.5. For example, let the medical events $hypertension_{start}$, $palpitations_{start}$, $infection_{start}$ and $MRSA_{start}$ be denoted by a, b, x, y respectively. if $a_{start} < x_{start} < b_{start} < y_{start}$, then we compute the probability of the candidate sequence as, *P(a before x | a no-coref x) P(a no-coref x)* $\times$ *P(x before b | x no-coref b) P(x no-coref b)* $\times$ *P(b before y | b no-coref y) P(b no-coref y)*

Thus, the coreference and temporal relation scores can be leveraged for aligning sequences of medical events. These scores are used in both the WFST-based representation and decoding, as well as for dynamic programming.

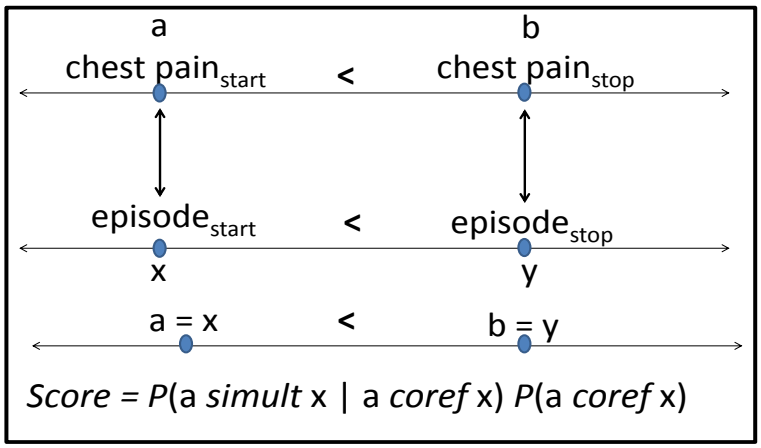


Figure 7.3: Score computation for aligning events across temporally ordered event sequences chest pain$_{start}$ < chest pain$_{stop}$ and episode$_{start}$ < episode$_{stop}$, where events across the sequences occur simultaneously and corefer.

7.5.2 Alignment using a Weighted Finite State Representation

A weighted finite-state transducer (WFST) is an automaton in which each transition between states is associated with an input symbol, an output symbol, and a weight [Mohri et al., 2002]. WFSTs can be used to efficiently represent and combine sequences of medical events based coreference and temporal relation information. The WFST representation gives us the ability to talk about the global joint probability derived from coreference and temporal relation scores described in Section 7.5.1. It allows us to build a weighted lattice of sequences that can be searched for the most probable sequence of medical events from across all clinical narratives of a patient.

We use unweighted FSAs to represent the input described in Section 7.4, i.e. temporally ordered sequences of medical events corresponding to clinical narratives. This corresponds to N_1 and N_2 in Figure 7.6.

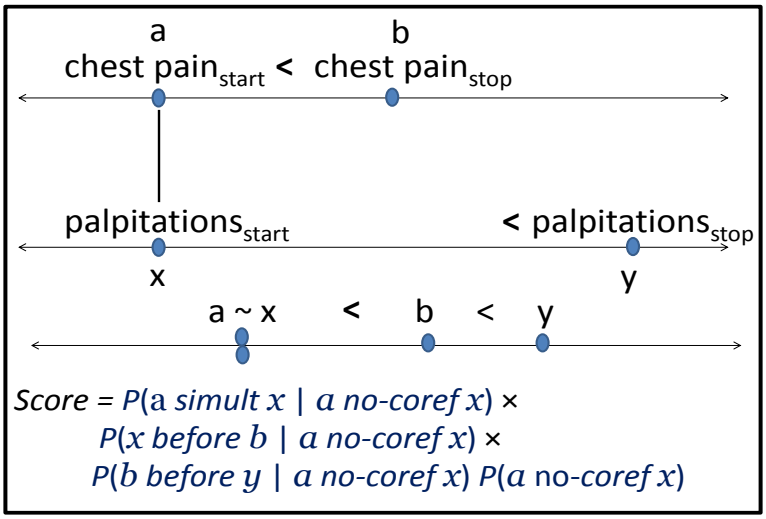


Figure 7.4: Score computation for aligning events across temporally ordered event sequences chest pain$_{start}$ < chest pain$_{stop}$ and palpitations$_{start}$ < palpitations$_{stop}$, where some events across the sequences occur simultaneously but do not corefer.

Based on whether we want to align the sequences purely based on coreference scores or both coreference and temporal relation scores, the arc weights for the WFST can be determined. M_{12}^{c} is a WFST that maps input symbols from N_1 to output symbols in N_2 and is weighted by the probability of coreference or no-coreference between medical events across N_1 and N_2 (shown in Figure 7.6). The representation in WFST M_{12}^{c+t} shown in Figure 7.7 allows us to align N_1 and N_2 based on both coreference as well as temporal relation probabilities. The WFST has ϵ transitions to accommodate insertion and deletion of medical events when combining the sequences. Deletions correspond to the case when an event in the first sequence does not map to any event in the second sequence; similarly insertions correspond to the case where an event in the second sequence does not map to any event in the first sequence. The WFST composition operation allows the outputs of one WFST to be fed to the inputs of a second WFST or FSA. Thus, we build our final machine

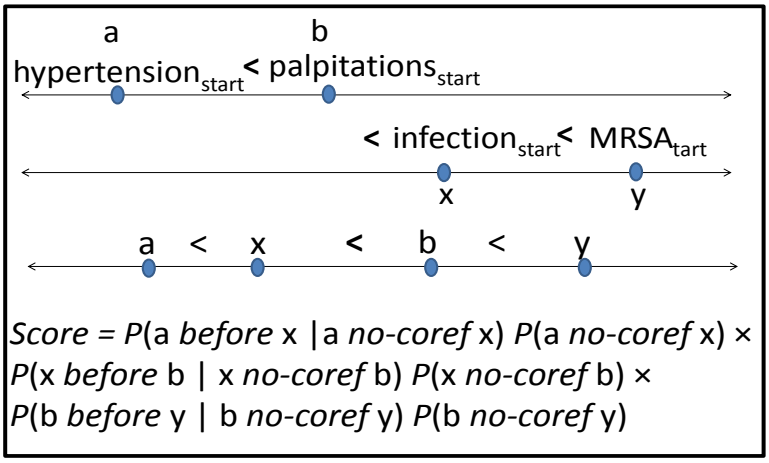


Figure 7.5: Score computation for aligning events across temporally ordered event sequences hypertension$_{start}$ < palpitations$_{start}$ and infection$_{start}$ < MRSA$_{start}$, where events across the sequences do not occur simultaneously and do not corefer.

by composing the three sub-machines as,

$$D = N_1 \circ M_{12}^{i} \circ N_2. \tag{7.1}$$

where $i = c$ or $i = c + t$. This gives us a combined weighted graph by mapping the output symbols of the first medical event sequence to the input symbols of the second medical event sequence. The scores on the decoding graph are derived from only the coreference probabilities if $i = c$ and both coreference and temporal relation probabilities if $i = c + t$.

In the medical event sequence alignment problem, we want to align multiple sequences of medical events that correspond to multiple clinical narratives of a patient. Since we want to now combine all narrative chains belonging to the same patient, the composition cascade to build the final combined sequence will be as,

$$D_f = N_1 \circ M_{12}^{i} \circ N_2 \circ M_{23}^{i} \circ N_3 \circ M_{34}^{i} ... \circ N_n \tag{7.2}$$

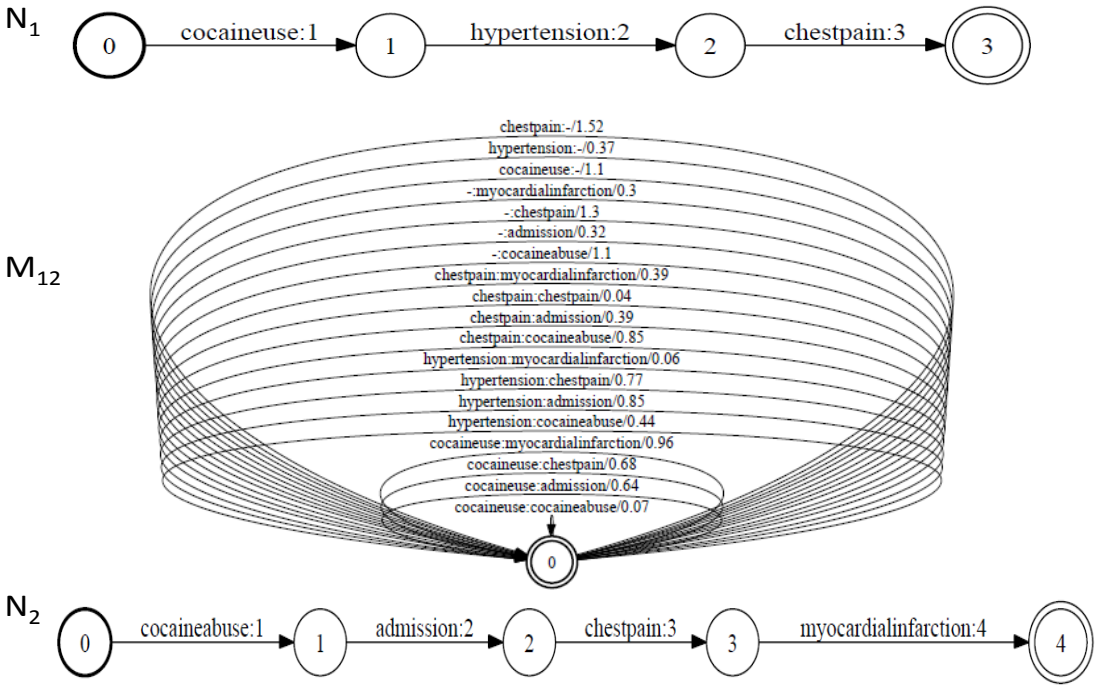


Figure 7.6: N_1 and N_2 are medical event sequences represented using FSAs. M_{12}^c maps medical events across N_1 and N_2 and is weighted only by the probability of coreference between events across N_1 and N_2.

where $i = c$ or $i = c + t$ and n is the number of medical event sequences corresponding to clinical narratives of a patient. During composition we retain intermediate paths like M_{23}^i utilizing the ability to do lazy composition [Shu, 2006] in order to facilitate beam search through the multi-alignment. The best hypothesis corresponds to the highest scoring path which can be obtained using shortest path algorithms like Djikstra's or Viterbi beam search algorithm. The best path corresponds to the best alignment across all medical event sequences based on the joint probability of cross-narrative medical event coreferences and temporal relations across the narrative sequences.

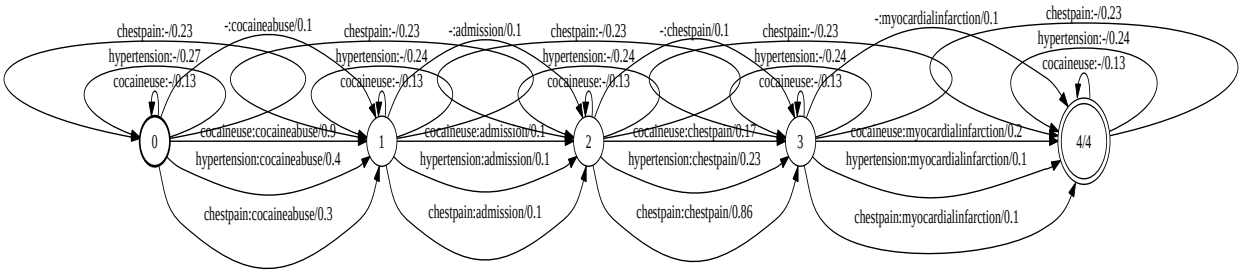


Figure 7.7: M_{12}^{c+t} is a WFST representation used for mapping medical events between N_1 and N_2 (from Figure 7.6) and is weighted by both the coreference and temporal relation probabilities

7.6 Narrative Sequence Alignment for Cross-narrative Temporal Ordering

The complexity of decoding increases exponentially with the number of narrative sequences in the composition, and exact decoding becomes infeasible. One solution to this problem is to do the alignment greedily pairwise, starting from the most recent medical event sequences, finding the best path, and iteratively moving on to the next sequence, and proceeding until the oldest medical event sequence. The disadvantage of such a method is that it does not take into account constraints between medical events across multiple event sequences and may lead to a less accurate solution.

An alternative method is to use lazy composition to perform more efficient composition as it allows practical memory usage. We also use beam search to make for an efficient approximation to the best-path computation [Mohri et al., 2002]. This allows for accommodating constraints from across multiple sequences and generates a more accurate best path. Thus, this method generates more accurate alignments when we have more than two sequences to be aligned.

For instance, say $a, b \in N_1, x, y \in N_2$, and $m, n \in N_3$ are temporally medical event sequences corresponding to narratives N_1, N_2 and N_3. Based on the learned pairwise temporal relations, if we have the following constraints $a < x$, $m > x$, $m < a$. Aligning N_1 and N_2 greedily pairwise may give us the best combined sequence as $a, x, b, y \in N_{12}$. Now in aligning N_{12} with N_3, we won't be able to accommodate $m > x$ and $m < a$. However, performing a beam search over the composed WFST in equation 7.2 allows us to accommodate such constraints across multiple sequences.

The complexity of composing two transducers is $O(V_1 V_2 D_1 (log D_2 + M_2))$ where each edge from the first sequence matches every edge in the second sequence and V_i is the number of states, D_i is the maximum out-degree and M_i maximum multiplicity for the i^{th} FST. The complexity of the shortest path computation is $O(V log V + E)$ [Mohri et al., 2002].

We also use popular dynamic programming algorithms [Needleman et al., 1970; Smith and Waterman, 1981] for sequence alignment of medical events across narratives and compare it to the WFST-based representation and decoding.

7.6.1 Pairwise Alignment using Dynamic Programming

Dynamic programming algorithms have been popularly used for sequence alignment in a variety of problems in computational biology, speech recognition and other domains [Myers and Habiner, 1981; Notredame, 2002]. We adapt two such dynamic programming algorithms for sequence alignment: 1) Global alignment using the Needleman Wunsch algorithm (NW) [Needleman et al., 1970] and 2) Local alignment using the Smith-Waterman algorithm (SW) [Smith and Waterman, 1981]. NW allows us to align all events in one sequence with all events in another sequence. However, a drawback of NW is that short

and highly similar sequences maybe missed because they get overweighted by the rest of the sequence. NW is suitable when the two sequences are of similar length with significant degree of similarity throughout. On the other hand, SW gives the longest sub-sequence pair that yields maximum degree of similarity between the two original sequences. It does not force all events in a sequence to align with another sequence. SW is useful in aligning sequences that differ in length and have short patches of similarity. The time complexity of these methods for sequences of length m and n are $O(mn)$.

The scoring scheme described earlier is used to update the scoring matrix for dynamic programming. In order to accommodate the temporal relations before and after, we insert a null symbol after every medical event in each sequence in the scoring matrix. A vertical or horizontal gap arises when cases 1, 2, 3 and 4 in Section 7.5.1 are not true. If the medical events are not simultaneous, not before or not after, the medical events will not align. Thus, the value of each cell in the scoring matrix is determined by computing the maximum score at each position $C(i, j)$ as,

$$\max\{(C(i-1, j-1) + S_{ij}), (C(i, j-1) + w), (C(i-1, j) + w)\} \tag{7.3}$$

$$\text{where, } S_{ij} = \max\{P(i = j), P(i < j), P(i > j)\},$$
$$\text{and } w = \max\{(1 - P(i = j)), (1 - P(i < j)), (1 - P(i > j))\}.$$

Here, $C(i-1, j-1)$ corresponds to a match, whereas $C(i-1, j)$ and $C(i-1, j)$ correspond to a gaps in sequence one and two. The first row and column C(i, 0) and C(0, j) are set to 0.

In the case of the SW algorithm, the negative scoring matrix cells are set to zero, thus making the positively scoring local alignments visible. Backtracking starts at the highest

scoring matrix cell and proceeds until a cell with score zero is encountered, yielding the highest scoring local alignment.

The time and space complexity grows exponentially with the number of sequences to be aligned and finding the global optimum has been shown to be a NP-complete problem. The time complexity of aligning N sequences of length L is $O(2^N L^N)$ [Wang and Jiang, 1994]. Thus, for MSA using dynamic programming, we use a heuristic method where we combine pairwise alignments in, an iterative manner, starting with the latest narrative and progressing towards the oldest narrative.

7.7 Experiments and Evaluation

Corpus Description. We gathered a gold standard set of seven patients (80 clinical narratives overall) with manual annotation of all medical events mentioned in the narratives, coreferences, and medical event sequence information (described in Chapter 3, Section 3.9. The annotation agreement across annotators is high, with 89.5% agreement corresponding to inter-annotator Cohen's kappa statistic of 0.86 [Conger, 1980]. The types of clinical narratives included discharge summaries, history and physical reports, radiology reports and pathology reports. The distribution of the number of medical event sequences and medical events across patients is shown in Table 7.1.

Evaluation Metric. The accuracy of the timeline generated by our proposed methodology is calculated as the number of transformations required to obtain the reference sequence in the annotated gold-standard from the one generated by our system. Transformations are measured in terms of the minimum edit distance, insertions, deletions, and substitutions of medical events (described in Chapter 3, Section 3.9).

	p1	p2	p3	p4	p5	p6	p7	
No. of Narrative Sequences	5	9	20	13	8	10	15	
No. of Medical events	125	239	488	318	220	136	297	
	% Accuracy							% Avg.
WFST (MSA)[c+t]	**76.1**	**73.2**	**81.2**	**83.5**	**76.4**	**82.5**	**79.7**	78.9
WFST (Pairwise)[c+t]	70.4	67.1	73.5	74.1	61.8	75.5	62.9	69.3
SW (Pairwise)[c+t]	71.2	69.7	75.5	75.6	66.3	77.4	68.3	72.1
NW (Pairwise)[c+t]	68.1	66.3	72.1	74.4	61.1	75.5	63.6	68.7
WFST (MSA)[c]	**68.5**	**65.3**	**72.3**	**74.4**	**67.2**	**71.3**	**69.1**	69.7
WFST (Pairwise)[c]	61.2	63.3	61.9	60.4	59.8	64.8	60.5	61.7
SW (Pairwise)[c]	60.3	63.7	68.2	62.3	58.6	66.7	60.2	62.8
NW (Pairwise)[c]	56.6	60.1	59.3	65.6	54.7	63.1	58.2	59.6

Table 7.1: The distribution of medical events across narrative sequences and sequences across patients and multiple sequence alignment results for the WFST-based framework, and dynamic programming using just coreference scores [c] and using coreference as well as temporal relation scores [c+t].

Experiments and Results. We first use a ranking model to obtain sequences of temporally ordered medical events corresponding to clinical narratives (intra-narrative) based on the method of [Raghavan et al., 2012]. The overall accuracy of ranking medical events using leave-one-out cross validation is 80.2%. These sequences serve as the input to the problem of cross-narrative sequence alignment.

We first run a baseline experiment for the task of cross-narrative temporal ordering, where we use the note creation and admission/ discharge dates to create an ordering across all sequences. This generates a timeline with 54.1% accuracy.

The cross-narrative coreference and temporal relation pairwise classification models are trained using a Maximum entropy classifier. The coreference resolution performs with 71.5% precision and 82.3% recall. The temporal relation classifier performs with 60.2%

precision and 76.3% recall. The learned pairwise coreference and temporal relation probabilities are now used to derive the score for the WFST and dynamic programming approaches.

WFST representation and decoding. We build finite-state machines using the open source OpenFST library.[28] We use a tropical semi-ring that is weighted using the negative log-likelihood of the computed scores. OpenFST provides tools that can search for the highest scoring sequences accepted by the machine, and can sample from high-scoring sequences probabilistically, by treating the scores of each transition within the machine as a negative log probability. As shown in Equation 7.2, the decoding process to compute the most likely combined medical event sequence can be defined as searching for the best path in the combined graph representation. The best path is the one that minimizes the total weight on a path (since the arcs are negative log probabilities). In searching for the best path, the beam size is set to 8. The accuracy of the WFST-based representation and beam search across all sequences using the coreference and temporal relation scores to obtain the combined aligned sequence is 78.9%.

Dynamic Programming. We use the NW and SW algorithms described in Section 7.6.1 to produce local and global alignments respectively. We use the scoring scheme described in Section 7.5.1 to update the cost matrix for dynamic programming and implement the algorithms as described in Section 7.6.1. The overall accuracy of sequence alignment with both coreference and temporal relation scores using Needleman-Wunsch is 68.7% whereas Smith-Waterman gives an accuracy of 72.1%. In case of aligning just two sequences, both methods yield the same results. The accuracy of cross-narrative MSA for each patient, for each method, using cross validation, is shown in Table 7.1. Results

[28] www.openfst.org

indicate that the WFST-based method outperforms the dynamic programming approach for multi-sequence alignment (statistical significance $p<0.05$). Morever, the results using both coreference and temporal realtion scores for alignment outperform using only coreference scores for alignment using all approaches. This indicates that cross-narrative temporal relations are important for accurately aligning medical event sequences across narratives.

7.8 Discussion

The accuracy of alignments across multiple medical event sequences is affected by the error induced by the coreference and temporal relation scores. Often, insufficient temporal cues leads to misclassification of events incorrectly as sharing the "overlaps" temporal relation and often as coreferring. This induces errors in the score calculation and hence the alignments. Thus, there is no clear trend with respect to the number of medical events and narratives of a patient (Table 7.1.) and the alignment accuracy as it depends on the learned coreference and temporal relation probabilities used to calculate the score. Additionally, we also implemented the ILP method for timeline construction proposed in Do et al. [2012] that also allows combining the output of classifiers subject to some constraints. We derive intervals from event starts and stops and learn two perceptron classifiers for classifying the temporal relations between events and assigning events to intervals. The classifier probabilities are then used to solve the optimization problem using the lpsolve solver.[29] We also use intra-document coreference information to resolve coreference before performing the global optimization. We observe that in case of MSA, the optimal solution using ILP is

[29] http://lpsolve.sourceforge.net/5.5

still intractable as the number of constraints increases exponentially with the number of sequences. In this case, aligning the medical event sequences pairwise in an iterative manner gives us an overall average accuracy of 68.2% similar to dynamic programming.

7.9 Conclusion

We propose a novel framework for aligning medical event sequences across clinical narratives based on coreference and temporal relation information using cascaded WFSTs. FSTs provide a convenient and flexible framework to model sequences of temporally ordered medical events and compose them into a combined graph representation. Decoding this graph allows us to jointly maximize coreference as well as temporal relation probabilities to derive a timeline of the most likely temporal ordering of medical events. This approach to aligning multiple sequences of medical events significantly outperforms other approaches such as dynamic programming. Moreover, we demonstrate the importance of learning temporal relations for the task timeline generation from across multiple clinical narratives by empirically proving that decoding using both coreference and temporal relation scores is far more accurate than decoding with only coreference scores.

At the end of the decoding process using the WFST framework, we obtain the highest scoring sequence of medical events that corresponds to the timeline of medical events across the patient's history. This timeline can now be leveraged to help various clinical applications including clinical trial recruitment, information retrieval with temporal constraints, multi-document summarization and clinical decision making. The accuracy of this overall timeline is limited by the temporal cues available in the unstructured clinical text. Often these cues may be sparse, implicit and hard to decipher and anchor to a medical event. One way to address this is to leverage temporal cues from the structured data in

the patient's electronic health record to help the models that extract information from the unstructured portion of the patient's health record. In the next chapter, we explore the different ways in which structured data can be useful in combination with unstructured data to the process of medical event timeline generation.

CHAPTER 8: INFORMATION FUSION

Electronic health records (EHRs) capture patient information using structured controlled vocabularies and unstructured narrative text. There is a wealth of temporal information in the structured and semi-structured data in the EHR. The ability to fuse information across structured data and unstructured clinical narratives allows generating a more accurate and complete timeline of medical events from a patient's EHR. This is because information fusion across these data sources enables the use of temporal information from structured data to improve the process of temporal relation learning from unstructured data.

In this chapter, we address an important question of leveraging the structured data to help information extraction from unstructured data in the EHR. Given the large amount of timestamped medical events in the structured data for a particular disease, can we leverage this to improve machine learning models for timeline generation from the unstructured data?

Next, we explore the problem of information fusion across structured and unstructured data in the EHR and investigate whether it helps in the process of using the medical event timeline for clinical trial recruitment.

8.1 Introduction

The overall purpose of our research has been to generate an improved longitudinal health record, which contains a comprehensive clinical summary of patient problems and treatments that are appropriately identified and organized in time. The electronic health record is composed of multiple data sources that are often redundant [Wrenn et al., 2010]

or inconsistent [Hripcsak et al., 2009], stored in uncoordinated unstructured clinical narratives and structured data. Structured data typically encodes lab results, encounters and medication lists, while unstructured data captures the physician's interpretation of the patient's condition, prognosis, and response to therapeutic intervention. An excerpt from structured lists of encounters and procedures are seen in Figure 8.1.

We propose methods to integrate structured (e.g. laboratory results, encounters, discharge conditions, medication lists, demographics, patient birth and death information) and unstructured information (e.g. discharge summaries, history & physical reports, admission notes, radiology reports) to help generate a comprehensive medical event timeline. The proposed methods for information fusion includes learning a temporal model from the structured data and using it as a prior in the machine learning models for temporal relation learning (Chapter 6) and coreference resolution (Chapter 5) from the unstructured data. Since the generated timeline merges information from multiple data sources, the resulting information is more accurate than would be possible if these sources were used individually [Dasarathy, 2001].

Further, we explore the utility of the generated timeline in clinical trial recruitment. We perform an empirical study to validate the argument and show that structured data alone is insufficient in resolving eligibility criteria for recruiting patients onto clinical trials for chronic lymphocytic leukemia (CLL) and prostate cancer. Unstructured data is essential to solving 59% of the CLL trial criteria and 77% of the prostate cancer trial criteria. More specifically, for resolving eligibility criteria with temporal constraints, we show the need for temporal reasoning and information integration with medical events within and across unstructured clinical narratives and structured data.

MRN	Start Time	Stop Time	Encounter
100002222	24-AUG-07	31-AUG-07	Lymphsrc unsp xtrndl org
100002222	01-SEP-07	30-SEP-07	Screen mammogram NE
100002222	07-SEP-07	07-SEP-07	Gastroduodenal dis NEC
100002222	17-NOV-07	19-NOV-07	Swelling in head & neck

MRN	Start Time	Stop Time	Procedure
100002222	24-AUG-07	31-AUG-07	Lipid Panel
100002222	24-AUG-07	31-AUG-07	Total Bilirubin
100002222	24-AUG-07	31-AUG-07	Direct Bilirubin
100002222	24-AUG-07	31-AUG-07	Glycosylated Hemaoglobin(A1C)
100002222	01-SEP-07	30-SEP-07	Bilateral Screening Mammography

Figure 8.1: Sample excerpt from structured of encounters and procedures (medical events) for a patient. Each medical event has an associated start and stop timestamp.

8.2 Contributions

Structured data in the EHR has timestamped medical events in lists of discharge conditions, encounters, medications and lab reports. Many of these medical events are referenced in the descriptions provided by physicians in unstructured clinical notes. Information fusion across these data sources can be done in two ways.

i Leverage the entire structured dataset (across all patients) for a disease (say chronic lymphocytic leukemia dataset (CLL)) to estimate a temporal model that can be used to inform the learning models from unstructured data.

ii Match medical concepts across structured and unstructured data and use it for improved inference in clinical applications including resolving eligibility criteria for clinical trails.

Some more specific contributions include the following.

- The timestamps in the structured data are considered ground truth, and can used to correct the time of occurrence of medical events in the unstructured data.

- Information fusion with structured data helps introduce detailed meta-data about certain medical events in the timeline. This in turn is very useful in practical applications where we need to query the timeline for information. For instance, a medical event "blood test" in the unstructured data timeline, may get mapped to 50 medical events in the structured data describing various parameters measured and calculated during the blood test such as bilirubin levels, RBC / WBC counts etc.

- We empirically evaluate the commonly assumed hypothesis that unstructured clinical text processing is required, and that structured data alone is insufficient to accurately resolve eligibility criteria. This is done with the help of a clinical trial use case.

- We experimentally demonstrate the need for cross-narrative temporal reasoning and information fusion across structured and unstructured data in solving certain temporal eligibility criteria

8.3 Related Work

The recent decade has seen considerable research in the natural language processing of unstructured clinical text [Chiang et al., 2010; Aronson, 2001; Savova et al., 2010b]. Demner-Fushman et al. [2009] discuss how successful processing of clinical narratives is the key to overall success of automated clinical decision support systems. They stress the importance of medical concepts with the help of named entity recognition and learning relations between those named entities are important for better understanding clinical narrative text. Wang et al. [2009] propose a framework for automated pharmacovigilance by applying NLP and association statistics on comprehensive unstructured clinical data from the EHR. They argue that previous algorithms have focused on coded and structured data, and therefore miss important clinical data relevant to this task. Medical NLP systems like

Mayo's cTakes [Savova et al., 2010a] and MedLEE [Chiang et al., 2010] have components specifically trained or designed for information extraction from multiple clinical text.

To the best of our understanding there is no prior work on information fusion across structured and unstructured data in the EHR with the objective of improving information extraction from unstructured data. We address this problem by estimating a temporal model with the timestamped medical concepts in the structured data and using this as a prior in learning temporal relations from unstructured data.

There has been some work on modeling temporal knowledge in eligibility criteria to help effective clinical text processing [Ross et al., 2010; Boland et al., 2012]. Ross et al. [2010] observe that temporal features were present in 40% of clinical trial criteria analyzed as part of their study, where the type of temporal expression in the criteria ranged from well-specified to loosely-specified. Similarly, there have been considerable efforts, including rule-based algorithms, temporal annotation of clinical corpora, and machine learning methods, towards learning temporal relations and generating timelines of medical events from unstructured clinical text [Zhou and Hripcsak, 2007; Sun et al., 2013]. Zhou et al. [2006] extract temporal relations between medical events in discharge summaries. The CLEF project [Savova et al., 2010a] uses a pairwise supervised classification approach to learn temporal relations between medical events within the same narrative. While temporal information has been studied in the intra-document context, there is not much prior work in cross-narrative temporal relation learning and information fusion. Carlo et al. [2010] attempt to align medical problems in structured and unstructured EHR data using UMLS by studying the information overlap between structured ICD-9 diagnoses and unstructured discharge summaries. They conclude that this is a non-trivial task with the need for better methods to detect correlating structured and unstructured data before aligning them.

Köpcke et al. [2013] compare the eligibility criteria defined in trial protocols with patient data contained in the EHR in multi-site trials to determine the extent of available data compared with the eligibility criteria of randomly selected clinical trials. However, their study is restricted to structured data in the EHR. In spite of the large body of recent work in processing structured and unstructured clinical narratives for temporal reasoning, and other NLP tasks, there are no prior studies that empirically evaluate the usefulness of structured vs. unstructured data for a clinical task. We perform an empirical analysis of CLL and prostate cancer patient records and evaluate the performance of structured and unstructured data in resolving clinical trial eligibility criteria. We specifically focus on criteria with temporal constraints and illustrate the need for unstructured clinical narrative analysis including cross-narrative temporal reasoning and information fusion.

8.4 Information Fusion across Structured and Unstructured Data

A patient's EHR has structured and unstructured data. In the last few chapters of this dissertation, we have proposed applying supervised machine learning models to learn temporal and coreference relations between medical events. However, such supervised learning models depend on accurate gold-standard annotations for training and evaluation purposes. Annotating clinical narratives for fine-grained relationships between medical events is a tedious and time-consuming task as described in Chapters 3 and 5. Further, the annotations need to be marked by medical domain experts to ensure correct interpretation of the clinical sub-language. This puts limitations on the amount and type of annotations that can be generated within a reasonable amount of time. Moreover, some clinical narratives may not even have sufficient temporal expressions co-occurring with medical events. This makes it difficult to accurately predict their relative temporal order and place them on a timeline.

In this section, we address these problems with the help of structured data using the the methodology described in Section 8.5.

8.5 Temporal Model from Structured data

Our strategy for information fusion tries to address a bigger problem of the lack of expert knowledge. Much of the knowledge required for clinical systems to perform better is expert medical domain knowledge. Often, physicians make decisions simply because they "know" certain facts either from years of experience or through intensive medical training. For instance, a physician knows that it is highly probable that the event *MRSA* will be followed by *antibiotics*. In order to replicate such knowledge one may need to build a very vast knowledge base with complex causal, temporal and semantic relationships at a patient, disease level, capturing the kind of knowledge that a physician may have. While the UMLS meta-thesaurus is an ontology that attempts to do this, it is not nearly as comprehensive enough. Moreover, it does not capture causal or temporal relationships. In this section, we explore how structured data can be used as a possible substitute for expert knowledge about temporal relationships between medical events.

Structured data contains timestamped medical events as part of lists of discharge conditions, encounters, medications and lab values. Given a disease dataset, there may be thousands of such lists across all patients. For instance, across all patients, the structured portion of our CLL dataset has 144512 timestamped procedures, 30083 timestamped encounters and 10731 diagnoses. The timestamped medical events in these lists can be used to estimate a temporal model to help timeline generation.

The proposed model is similar to a language model that models the probability of a sequence of words. Instead, this model models the probability of a sequence of medical

events using a probability distribution. Thus, the temporal model tries to capture certain temporal properties of a disease domain. The probability of observing medical events $me_1, .., me_m$ is computed as follows.

$$P(me_1, \ldots, me_m) = \prod_{i=1}^{m} P(me_i | me_1, \ldots, me_{i-1}) \approx \prod_{i=1}^{m} P(me_i | me_{i-(n-1)}, \ldots, me_{i-1}) \tag{8.1}$$

Here, the probability of observing the i^{th} medical event me_i in the context of the preceding $i-1$ medical events can be approximated by the probability of observing it in the shortened context of the preceding $n-1$ medical events (n^{th} order Markov property). Since we have access to limited structured data, we set n to 2. However, given access to larger amounts of structured data for a particular disease, it may be possible to estimate larger order temporal models. The maximum likelihood estimate when $n = 2$ is given as follows.

$$P(me_i | me_{i-1}) = \frac{count(me_{i-1} me_i)}{count(me_{i-1})} \tag{8.2}$$

This would work well for frequent events. However, since we have limited structured data, and also given that not every medical event documented in the unstructured data may be present in the structured data, there may be zero count medical events. We perform a Laplace (add 1) smoothing to address the zero count problem.

Now, we can estimate the probability of certain medical events occurring after certain other medical events from the structured data (Equation 8.2). This probability can be used as a temporal feature in the temporal relation learning and coreference resolution models. If medical event sequence is highly probable, then it can even be introduced as a hard constraint in learning temporal relations from unstructured data. The results for the final timeline (cross-narrative temporal ordering) using the probability estimated from the temporal model on the structured data is shown in Table 8.1. We see there is a significant

improvement in accuracy in temporal ordering in unstructured data on introducing probabilities from the temporal model as features.

	p1	**p2**	**p3**	**p4**	**p5**	**p6**	**p7**
Needleman-Wunsch	70.3	69.2	74.1	80.4	63.5	78.1	69.2
Smith-Waterman	73.1	73.5	78.5	83.8	68.7	79.6	74.5
WFST beam search	79.4	77.1	85.2	86.1	77.1	84.2	86.7

Table 8.1: Cross-narrative temporal ordering from unstructured data using the probability from the temporal model estimated from the structured data as a feature

8.6 How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?

Clinical trial recruitment may be semi-automated through information extraction from the EHR. Clinical trials have eligibility criteria that describe characteristics and constraints that help determine if a patient qualifies for a trial. Typically, clinicians and trial recruitment coordinators identify potential clinical trial patients from characteristics described in their medical history and match them against the eligibility criteria for individual trials. This standard model of clinical trial recruitment is rife with errors. If the clinical staff is unfamiliar with a particular trial or if there are competing trials, an eligible patient may be overlooked. On the other extreme, the clinical trials staff may be asked to pre-screen patients who are clearly not candidates. This information mismatch has the potential to be streamlined. Generating automated queries corresponding to eligibility criteria and querying patient records from the EHR in order to identify qualifying patients provides an efficient and agnostic approach to clinical trials recruitment. The pertinent question then is

whether structured data, being easier to automatically process and understand, has sufficient information to resolve these eligibility criteria, or if there is a need to extract and reason with medical concepts in unstructured clinical narratives.

Researchers have often emphasized the importance of using clinical narratives for clinical decision support [Demner-Fushman et al., 2009], information retrieval [Tange et al., 1998], question answering [Kalyanpur et al., 2012] and automated clinical trial recruitment [Köpcke et al., 2013]. Unstructured data in clinical narratives captures important decisions and relationships between medical concepts including causal (symptom caused disease), consequential (why a drug or treatment was administered) and temporal (symptom before disease / treatment). Furthermore, Rosenbloom et al. [2011] suggest that clinical notes containing naturalistic prose have been more accurate and reliable for identifying patients with given diseases, and more understandable to healthcare providers reviewing patient records. However, to the best of our knowledge, there are no prior empirical studies that evaluate the usefulness of structured vs. unstructured data considering their advantages and limitations for a clinical task. In this paper, we study two datasets of structured and unstructured data with patients suffering from chronic lymphatic leukemia (CLL) and prostate cancer obtained from The Ohio State University Wexner Medical Center. Given a set of eligibility criteria from corresponding clinical trials, we evaluate the number of criteria that can be resolved using information from just the structured data and the number of criteria that require information extraction from and reasoning with unstructured clinical narratives and data.[30]

[30]This work has been published in AMIA CRI 2014. P. Raghavan, J. Chen, E. Fosler-Lussier, and A. Lai, "How essential are unstructured clinical narratives and information fusion to clinical trial recruitment?" AMIA Joint Summits on Translational Science, 2014.

8.6.1 Data Description

The EHR data used in this study consists of medical records for 2060 CLL patients and 1808 prostate cancer patients. The CLL dataset contains 95 different types of unstructured reports including discharge summaries, history and physical reports, specialty reports such as wound care, operative notes, OB/GYN and psych evaluations, social work assessment, referral letters and progress notes. It also consists of radiology reports, pathology reports and cardiology reports. The total number of unstructured clinical narratives in the CLL dataset is 100704. The structured data consists of lab reports, procedures list, diagnoses list and encounters list. The prostate cancer dataset consists of 2652 oncology reports, 1582 pathology reports, 6606 radiology reports as part of unstructured data. The structured data in this dataset includes a discharge medications list (30178 medications), laboratory values (939 values), and a medications list (141932 medications). The clinical trials dataset consists of a set of top 100 clinical trials each, as defined by clinicaltrials.gov, for both CLL and prostate cancer.

8.6.2 Methodology

Medical concept extraction - We annotated the clinical trial criteria datasets with medical concepts, concept unique identifiers (CUIs) and semantic types using MetaMap [Aronson, 2001]. We then extracted criteria containing the following semantic types: Disease or Syndrome, Laboratory or Test Result, Procedure, Sign or Symptom, and Pharmacological Substance. The criteria containing the Temporal Concept semantic type were labeled as temporal eligibility criteria. Similarly, we also annotated both patient datasets with medical concepts and the semantic types mentioned previously. Matching medical concepts across clinical trials and patient datasets - In order to evaluate the degree of overlap between the

clinical trials dataset and structured and unstructured data in the medical records dataset, we compute the Match between medical concepts across these datasets. The match functions are computed across the datasets as follows. 1) UMLS CUI Match where an exact CUI match is computed and 2) Phrase Match where we compute a match between medical concepts (textual fragment identified as the medical concept). Thus we have,

- Match(CUI in the trial dataset, CUI in structured data)
- Match(CUI in the trial dataset, CUI in unstructured data)
- Match(Phrase in the trial dataset, medical concept in the structured data)
- Match(Phrase in the trial dataset, medical concept in the unstructured data)

These match functions are computed for two levels of analysis - (i) medical concept-level, where we compare all the medical concepts in the trials dataset against the structured and unstructured data, and (ii) eligibility criteria level, where we compare all the medical concepts in each criterion against the structured and unstructured data. The medical concept-level match helps analyze the number and type of medical concepts typically found in the structured and unstructured datasets when solving clinical trial eligibility criteria. As shown in the algorithm below, we compute the match between all medical concepts in the clinical trials dataset and the structured data. If there are no matching concepts found in the structured data, we then compute a match with the unstructured data.

The eligibility criteria-level match helps us analyze the number of criteria that can be solved by structured data, unstructured data or both. In order to evaluate the need for temporal reasoning and information fusion and constrain the number of eligibility criteria, we restricted the eligibility criteria-level analysis to criteria with temporal constraints. We

compare each eligibility criterion against both structured data and unstructured data to determine if the concepts in the criterion require only structured data, only unstructured data or both datasets together for resolution, as shown in the algorithm below.

The algorithm first compares all medical concepts in the eligibility criterion against all medical concepts in the structured data. If all the concepts in the criterion are found in the structured data, we conclude that the criterion may be resolved using the structured data. We then do a similar comparison for unstructured data and if all concepts in the criterion are found in the unstructured data, we conclude that the criterion may be resolved using the unstructured data.

Information fusion. In the case where all the concepts in the criterion are found in both the structured as well as the unstructured data, we conclude that the criterion can be solved using either the structured or the unstructured data. However, the criterion may also require both structured as well as unstructured data for resolution. Taking this into consideration, we define information fusion as follows. Given medical concepts $m_1, ..., m_n$ in a clinical trial criterion, if S_k is a set of k concepts that match the structured data and U_j is a set of j concepts that match the unstructured data, where $k, j > 0$ and $k, j < n$. Now there are two possibilities.

i $L = Sk \cap U_j$ is not empty. Here, L concepts match both structured and unstructured data.

ii $L = S_k \cap U_j$ is empty. Here, L concepts match the structured data and the remainder j concepts match the unstructured data. So S_k and U_j are disjoint.

Temporal reasoning in unstructured data - For subset of criteria that require unstructured data for resolution, we further analyze the temporal constraints in the criteria and attempt

to answer the following questions. How many temporal constraints can be solved using coarse temporal reasoning within each clinical narrative? How many temporal constraints require more granular temporal ordering within each clinical narrative? How many temporal constraints require cross-narrative temporal reasoning? In order to answer these questions, we run a CRF-based time-bin tagger (Chapter 4) and learn to associate the medical events within each narrative with one of the coarse time-bins: "way before admission, before admission, admission, after admission, discharge." The time-bin tagger was trained on different patient records not part of this dataset. We also perform fine-grained temporally ordering by learning to rank medical concepts within a clinical narrative by their order of occurrence (Chapter 6). This gives us both a coarse ordering and a fine-grained ordering of medical concepts within each clinical narrative. These intra-narrative temporal orderings are then combined with the admission and discharge dates across narratives to generate a cross-document partially ordered timeline of medical concepts for each patient.

8.6.3 Results

The methodology is empirically evaluated by calculating the extent of match between the eligibility criteria dataset and the structured and unstructured datasets. The medical concept-level match results between the trials datasets, consisting of all eligibility criteria, and the structured and unstructured data are shown in Table 8.2. The CLL trials dataset has 2167 medical concepts and the prostate cancer dataset has 1019 medical concepts. The CLL trials have a total of 1720 eligibility criteria, while the prostate cancer trials have 1325 eligibility criteria, containing diseases, procedures, tests, symptoms and medications. We observe that more than half of the medical concepts in the CLL and prostate patient

data were only found in the unstructured data. The most frequent medical concept semantic types found in the unstructured datasets include Finding, Sign or Symptom, Disease or Syndrome, whereas the most frequent medical concept semantic type in the structured data includes Laboratory Test or Procedure, Pharmacological Substance and Disease or Syndrome. If the structured data has diagnoses and encounters lists, there tend to be overlapping Disease or Syndrome type concepts across the structured data and unstructured clinical narratives.

	CLL		Prostate Cancer	
	CUI	Medical Concept	CUI	Medical Concept
Structured Data Match	23%	29%	11%	19%
Unstructured Data Match	61%	68%	48%	57%

Table 8.2: Medical Concept-level Analysis on CLL and Prostate Cancer Trials and Patient Records

354 of the eligibility criteria in the CLL trials and 297 of the eligibility criteria in the prostate cancer trials have temporal constraints. Table 8.3 shows results from matching temporal clinical trial eligibility criteria against structured and unstructured data. In both patient datasets, matching the textual fragment identified as the medical concept gives us a higher match percentage than trying to match CUIs. Importantly, the dependence on unstructured data for resolution of temporal eligibility criteria is higher than structured data. There is especially a huge gap between the structured and unstructured data match in the case of prostate cancer, where structured data only contributes to the resolution of 9% of the criteria.

	CLL		Prostate Cancer	
	CUI	Medical Concept	CUI	Medical Concept
Structured Data Match	35%	37%	9%	9%
Unstructured Data Match	53%	59%	75%	77%

Table 8.3: Eligibility Criteria-level Analysis on CLL and Prostate Cancer Trials and Patient Records

We observed that from the temporal criteria requiring unstructured data for resolution, frequently intra-narrative temporal reasoning was sufficient for resolving temporal constraints. The learned time-bins, along with the admission and discharge dates on each narrative, were useful in assigning medical concepts to coarse time-periods and in resolving 41% of the eligibility criteria that required an unstructured data match. For instance, the constraint, "patients with a distant history (greater than 6 months before study entry) of venous thromboembolic disease are eligible," requires mapping of venous thromboembolic disease to a time-bin way before time. Whereas "clinically significant bleeding event within the last 3 months, unrelated to trauma, or underlying condition that would be expected to result in a bleeding diathesis" required fine-grained temporal ordering of medical concepts. Further, as shown in Table 8.4, from the criteria that required unstructured data for resolution, 33% and 35% required cross-narrative temporal reasoning in the CLL and prostate cancer dataset respectively. A criteria such as, "fever $>$ 100.5 F for 2 weeks without evidence of infection," requires extracting the fact that fever lasted for 2 weeks by examining multiple mentions of fever across history and physical reports and discharge summaries to determine when fever started and stopped. This additionally requires the ability to perform coreference resolution across clinical narratives. Criteria requiring information from both structured and unstructured data (information fusion) were determined based on the

presence of the medical concepts in the criteria across these data sources. For instance, "if they have achieved stable blood pressure (bp) on a regimen of over 2 drugs after 6-8 weeks of therapy." The value of "bp" can be obtained from the structured data, however the nuanced relationship information about the drug regimen that was prescribed to stabilize "bp," along with its time duration, requires time-bin learning and cross-narrative temporal reasoning. We observed that while a large percentage of CLL criteria required fusion, the lower number of prostate cancer criteria is mainly due to limited structured data available for prostate cancer.

	CLL	Prostate Cancer
Cross-Narrative Temporal Reasoning	33%	35%
Information fusion$L = S_k \cap U_j$ is not empty	24%	3%
Information fusion $L = S_k \cap U_j$ is empty	17%	1%

Table 8.4: Eligibility Criteria that require Cross-narrative Temporal Reasoning and Information Fusion for resolution

8.6.4 Discussion

We studied two datasets of patients, CLL and prostate cancer, and evaluated the usefulness of structured vs. unstructured data in recruiting for corresponding clinical trials. We observed that the type of structured data, its granularity, and the information available vary across patient datasets. While the CLL patient dataset has detailed structured data in the form of diagnoses lists, encounters list, procedures and lab values, the prostate cancer dataset has limited structured data mostly consisting of medication lists and lab values. More fundamentally, the data heterogeneity reflects the underlying tumor heterogeneity at multiple levels. These levels include: (i) patient referral patterns, (ii) patterns of disease

treatment, and (iii) differences in disease stages. At The OSU James Cancer Hospital, the majority of prostate cancer patients tend to be referrals from community oncologists or urologists after failure of first and second line therapies. In contrast, CLL patients are mostly evaluated from time of diagnosis and thus their entire case history is within the OSU system. Secondly, laboratory values for prostate cancer patients are often drawn at their local laboratory and subsequently faxed to their oncologist at OSU. These labs are not directly accessible and are found in the unstructured component of the medical record. In stark contrast, CLL labs are nearly universally drawn at OSU. These tumor type differences would help explain our findings that prostate cancer requires the use of the unstructured data more frequently. The end result is that prior treatment history for prostate cancer patients who are seen at a later stage will have their disease course and treatment course summarized in the unstructured narrative. CLL patients are captured at an earlier stage and therefore their disease course and treatment history is more easily obtained from the structured text. This tumor type heterogeneity is reflected in the diagnosis codes that are available. In the case of CLL, these codes are useful in checking eligibility criteria that check for the presence or absence of a medical condition can be resolved easily from the structured data using these lists. In case of prostate cancer, this data is not as complete. Tumor heterogeneity aside, structured data may also fail if the medical concept is at a finer level of granularity than what is required for an exact match. In such cases, examining the unstructured data for additional information, or additional processing to check for related higher level concepts for medical events in the structured data may help better resolve the eligibility criteria.

8.6.5 Conclusion

We performed an empirical evaluation of clinical trial eligibility criteria resolution using structured and unstructured patient datasets from CLL and prostate cancer. We observed that unstructured data is essential to resolving eligibility criteria in 59% of the CLL trial criteria and 77% of the prostate cancer trials. We also demonstrated the need for cross-document temporal relation learning and information fusion across structured and unstructured data sources. Although structured data is useful in resolving certain criteria, it is limited by information granularity and structured data type. Thus, structured data is best used for first pass filtering of EHR data in eliminating a criterion based on the presence or absence of a certain lab test or diagnoses, prior to a more nuanced second pass using unstructured data. Moreover, improving the coverage of the structured data in the EHR would improve its ability to be used as a clinical trial recruitment tool.

CHAPTER 9: CONCLUSIONS AND FUTURE WORK

Extracting and reasoning with events and time expressions in natural language has become an active area of research in the computational linguistics. Specifically, this is an important topic in clinical informatics as time is an important feature of longitudinal clinical text. The increased availability of electronic health records over the past decade has given researchers the opportunity to extract and reason with clinical variables in structured and unstructured patient records to help support clinical applications.

In this dissertation, we presented a novel framework for medical event timeline generation from a patient's electronic health record. The framework (illustrated in Chapter 1, Figure 1.2) consists of models for intra- and cross-narrative temporal ordering of medical events from unstructured clinical narratives and a module for information fusion across structured and unstructured data. This is the first end-to-end framework for timeline generation from clinical narratives. Importantly, through a series of experiments to address the problem of temporal relation learning and coreference resolution, we demonstrate how we can leverage clinical domain heuristics in training machine learning models for information extraction from unstructured clinical text.

9.1 Summary of Work and Contributions

We first address the problem of intra-narrative temporal ordering in Chapters 4 and 6. The main contribution in addressing this problem is leveraging narrative structure and sub-language characteristics in training models for learning coarse as well as fine-grained

temporal relations between medical events in the same clinical narrative. We demonstrate our approach through experiments on an actual dataset of unstructured and structured patient data obtained from the EHR at The Ohio State University Wexner Medical Center. The patient narratives are annotated with medical events, temporal relations and coreference information to enable training and evaluation of machine learning models.

Time-bin learning. In Chapter 4, we addressed the problem of learning to assign medical events to coarsely defined time-bins using a sequence tagging approach with a linear-chain Conditional Random Field. We demonstrated through experiments that this outperforms a MaxEnt model that does not use any sequence information.

The main contribution here is that the learned time-bins (along with explicit dates like admission, discharge) can be used to infer a coarse partially ordered timeline of medical events that may be useful to clinical applications with coarser temporal constraints. The time-bins also serve as a useful temporal feature for coreference resolution and fine-grained temporal ordering of medical events.

Semi-supervised coreference resolution. Since information redundancy is a characteristic of clinical narratives, we propose methods for coreference resolution of medical events in Chapter 5. Taking into consideration how tedious and time-consuming the task of obtaining expert annotations for these tasks is, we explore semi-supervised methods, co-training and posterior regularization, for medical event coreference resolution.

Importantly, we empirically demonstrate that posterior regularization does almost as well as supervised learning (on a 60/40 split of the data) for this task. This contribution is of great value to the community, where owing to the difficult nature of the data and the task, annotations are hard to obtain for supervised learning. The learned coreferences also serve as a useful feature in learning fine-grained temporal relations.

Ranking for temporal ordering. In Chapter 6, we learn all of Allen's temporal relations between the medical events "starts" and "stops" by learning to rank them in relative order of occurrence using SVM-rank. We demonstrate through experiments that this method outperforms a pairwise classification approach to learning temporal relations between medical events. We observe that the opposite is true of these methods on the Timebank corpus of newswire text.

An important contribution here is highlighting the differences between domains in learning to temporally order events. State-of-the-art methods that are proven to perform well on standard community-shared corpora (usually newswire text) do not always perform as well on other real-world data. This finding has important implications for styles of data representation and resources used for temporal relation learning: clinical narratives may have different language attributes corresponding to temporal ordering relative to Timebank, implying that the field may need to look at a wider range of domains to fully understand the nature of temporal ordering.

The ranking process helps generate temporally ordered medical event sequences corresponding to each clinical narrative. Given these sequences, we next address the problem of cross-narrative temporal ordering using a novel sequence alignment approach.

Cross-narrative temporal ordering This cross-narrative problem is a huge challenge due to the lack of discourse coherence and context across narratives. We model this problem as a sequence alignment task by leveraging coreference and temporal relation information to align medical events across narratives in Chapter 7.

The important contribution here is the novel framework for multiple sequence alignment using cascaded weighted finite state transducers (WFSTs) to derive the most likely temporal ordering of medical events for a patient. The alignment scores are computed based on

learned pairwise temporal and coreference relations between medical events across clinical narratives. We demonstrate that this method outperforms iterative pairwise dynamic programming and an integer linear programming-based method [Do et al., 2012] for the task of cross-narrative alignment of multiple medical event sequences. The result of the cross-narrative alignment is the most probable sequence of medical events (timeline) across the patient's history.

Information fusion. We then explore how structured data in the EHR, that is usually available in plenty, can help the process of timeline extraction from unstructured clinical narratives in Chapter 8, Section 8.4. The main contribution here is the proposed temporal model estimated from timestamped medical events that can be used compute the probability of certain medical events following certain other events based on event frequencies. We demonstrate through are experiments that using the probability from this temporal model as a temporal feature helps improve the accuracy of the timeline generated from the structure data.

Utility of generated timeline. Finally, we investigate the utility of the framework for timeline generation in resolving temporal clinical trial eligibility criteria. With the help of a set of criteria extracted from `http://clinicaltrials.gov`, we demonstrate through experiments that information extracted from unstructured clinical narratives, and information fusion across structured and unstructured data is required to resolve certain constraints in temporal eligibility criteria.

9.1.1 How does this research affect the state-of-the art in the community?

The proposed framework is the first of its kind for timeline generation from longitudinal clinical text. While there have been limited efforts at temporal relation learning from

clinical text by Zhou and Hripcsak [2007]; Harkema et al. [2005]; Sun et al. [2013], among others, there isn't any integrated framework for temporal relation learning at various levels of granularity both within and across unstructured clinical narratives. Moreover, this work tries to address a difficult problem of learning cross-narrative events relations. Previous work has mostly been restricted to extracting information from a single clinical note of a certain type (say discharge summaries). Even generally, cross-document relations in NLP have only been explored to a limited extent. In this work, we learn relations between events within and across unstructured clinical narratives, enabling information extraction from longitudinal clinical text. Moreover, we do this across all types of clinical notes. Through experiments with classification and ranking on Timebank and clinical narratives, we demonstrated the need for novel NLP methods for real-world data such as clinical text. Moreover, generating a timeline of events from temporally incoherent unstructured text has been a topic of significant interest in the clinical informatics community. The proposed framework, is the first effort at producing a comprehensive timeline, considering unstructured and structured data, across the patient's history.

9.2 Future Work

This is the first attempt at generating such a timeline from longitudinal clinical text. There is a lot of scope for improving the features and models used in this dissertation to generate a more accurate timeline. Moreover, the generated timeline can be used as a tool to enable many other problems in natural language processing. In this section, we enumerate some future directions for this work.

- **Event and Temporal Expression Extraction.** The learning models proposed in this dissertation are trained on gold-standard medical events and temporal expressions (although we use TimeText [Zhou et al., 2006] to generate additional temporal expressions) tagged by our annotators. The popular tool for automatic event extraction in the clinical informatics community is MetaMap [Aronson, 2001]. However, MetaMap has its own limitations including the fact that it doesn't always identify contextually relevant medical concepts and often doesn't identify the entire long phrase that constitutes the concept. Better mechanisms for automatic extraction of medical concepts and temporal expressions and anchoring of temporal expressions to medical events will help the process of completely automated timeline extraction to a large extent [Jindal and Roth, 2013].

- **Duration of Events.** The framework for timeline generation only considers the relative temporal ordering between medical events in longitudinal clinical text. However, in many clinical applications, it is important to know the duration of the medical event. Although we learn this to some extent as we order event starts and stops (and some explicit dates that the events are anchored to), we fall short of learning the exact time interval between the event start and stop. There has been prior work in learning to interpret temporal phrases given a corpus of utterances and the times they reference, using a compositional grammar of time expressions [Angeli et al., 2012]. This allows us to ground temporal expressions probabilistically using a loosely supervised EM-style bootstrapping method. Leveraging such methods may be an important part of predicting the duration of medical events.

- **Evaluation and Scalability.** We have evaluated our methods on a limited dataset of clinical data due to the tedious nature of the annotation process. Expanding the dataset and verifying the applicability and scalability of these methods is something that needs to be addressed. If direct evaluation is not possible, as that would require detailed annotations on a larger dataset, the methods can be indirectly evaluated. The indirect evaluation can be done by measuring the use of the timeline in resolving temporal constraints for various clinical applications. Cross-institutional validation of these methods to demonstrate generalizability across clinical datasets from different EHRs.

- **Discourse Relations.** There are different types of discourse relations between textual units in discourse. These include *comparison, expansion, contingency, causal, temporal.* We have focused on the temporal relation between medical events in clinical text. Events in clinical text can be related by other relations like causality which is of great significance to medicine. Learning different types of discourse relations, to represent more complex non-linear relationships with the events in the generated timeline, could be an interesting area of research.

- **Information Fusion.** There may be multiple ways to combine structured and unstructured in the EHR. We have explored information fusion in a limited setting using structured data to help the process of learning from unstructured data. However, training temporal models on very large disease datasets, will help generate more accurate predictions. It would be interesting to see how much this helps unstructured data learning.

- **Visualization of the timeline.** The output of our system is a partially ordered medical event timeline. Such a timeline may contain over 100 or more medical events. Designing a cognitively engineered display for visualizing the timeline in an intuitive manner, to help clinical understand and analyze the information available in the timeline, will be of immense use to the clinical community.

9.3 Conclusions

Electronic health records have brought an emerging challenge to health care: how do clinicians extract critical information, from rapidly growing quantities of health records, which can be perhaps used to influence physician behavior, and improve the quality of health care? Clinical data has distinct sub-language specific characteristics that present opportunities for natural language processing to enable unstructured data-analysis using clinical domain-heuristics in training machine learning models.

In this dissertation, we describe a novel framework for timeline generation from clinical narratives. Specifically, we propose methods for representing and reasoning with medical events and temporal information, both within and across narratives. We learn to resolve medical event coreferences and incorporate the learned information as an integral part of the temporal ordering process. We propose a novel ranking-based approach for learning intra-narrative temporal relations. Cross-narrative temporal ordering of medical events is a huge challenge due to the lack of discourse coherence across narratives. We model this problem as a multiple sequence alignment task, using a cascaded weighted finite state transducer (WFST)-based approach, and derive the most likely temporal ordering of medical events for a patient. The generated medical event timeline is of great utility in clinical

applications with temporal constraints including clinical trial recruitment, multi-document summarization and adverse drug reaction mining.

APPENDIX A: TYPES OF CLINICAL NARRATIVES

The clinical narratives corpus includes radiology notes, history and physical reports, social work assessments, progress notes, discharge instructions and discharge summaries. Some common characteristics across different types of patient narratives are as follows. Every clinical narrative is characterized by a structured header with information such as "medical record number" (MRN), "patient name," "physician name," "admission date," "discharge date," and "patient's date of birth". This is followed by multiple sections with unstructured content describing details of the patient's medical condition and the health care administered. The sections and section content vary based on the type of clinical narrative. The sections are usually delineated by a section header. However, there may be some exceptions where the entire note is a continuous paragraph without any section grouping. We examine the characteristics of different types of clinical narratives to better understand their characteristics.

A.1 Discharge Summaries

A discharge summary is a clinical report prepared by a physician or a health professional at the conclusion of a hospital stay or series of treatments. It outlines the patient's chief complaint, the diagnostic findings, the therapy administered and the patient's response to it, and recommendations on discharge, usually described under appropriately titled sections. Based on the discharge summaries that were studied, we noted the use of mostly

complete sentences. However, there is considerable use of medical abbreviations and terminology. The sections that are usually found in a discharge summary are as follows:

- Past Medical History: This section briefly states medical conditions from the patient's history that maybe relevant to his present illness. For instance, "Medical history - Significant for a mitral valve prolapse and a questionable history of hypertension." The medical conditions stated in this section may be detailed in the "History of Present Illness" section.

- Social History: It describes any social habits that may affect the patient's health. For instance, "The patient is a smoker 10-pack-year history of IV drug abuse 5 years ago history of incarceration 5 years ago."

- History of Present Illness: This section describes the patient's medical condition before admission to the hospital. It also includes some initial observations after admission to the hospital. It outlines diseases, complaints, medication routines and habits of the patient that maybe relevant to his present illness. For instance, "The patient noted no palpitations. Episodes lasted less than 5 minutes. She said that she would typically take an aspirin lie down and she would feel better ...." "At the time of admission to the ER the patient was chest pain free ...." The sentences within this section are usually in sequential order of time. This section may contain medical conditions that are coreferring with those found in "Past Medical History" and "Social History."

- Physical Examination: This section includes observations from the physical examination of the patient throughout his stay in the hospital. This could include observations on blood pressure, pulse, and respirations. It may also include cardiovascular

details, condition of the lungs, abdomen, etc., based on the patient's present illness. For instance, "Vital signs - On admission the patient's blood pressure was 141/82 pulse 98 respirations 18 afebrile."

- Hospital Course: This section describes the treatment course that the patient underwent during his stay in the hospital. The temporal flow is not necessarily sequential. It generally details the various medical conditions that were diagnosed and treated, including details of tests done and drugs administered. For instance, "She did have an echo-cardiogram done which showed a normal ejection fraction of 55% mild diastolic dysfunction with mitral valve prolapse and mild aortic and mitral regurgitation moderate tricuspid regurgitation noted. The patient also had a stress test which was negative ...."

- Disposition: This section describes the patient's disposition with respect to his medical condition before discharge from the hospital. For instance, "The patient was not having chest pain was up and ambulating and was tolerating a regular diet ...."

- Diagnosis: This section describes the final diagnosis for the patient's present illness. This medical condition described in the final diagnosis is most likely coreferring with instances described in "Past Medical History," "History of Present Illness," and "Hospital Course." For instance, "Mitral valve prolapse."

There maybe other sections which are minor variations to the ones described above. These include "Review of Systems," "History," "Wound Clinic History," "Vital Signs at the Time of Discharge." These may vary based on the patient's illness and the physician writing the discharge summary. The patient's follow up and care could also form a separate section in the discharge summary. However, there may also be a different clinical narrative

outlining recommendations for the patient after discharge. This clinical narrative is usually called “Discharge Instructions.” There may be instances where a discharge summary only has one continuous paragraph which is a combination of the content under all sections.

A.2 History and Physical (H&P) Report

H&P reports contain an admitting diagnosis that answers the question, “Why is this patient being admitted?” or “Why does this patient need to be in the hospital?” They are structured similar to discharge summaries, but are usually more detailed. The H&P report describes the “Past Medical History (PMH)” and the History of Present Illness (HPI) of the patient. It is divided into sections such as “History of Present Illness,” “Past Medical History,” “Physical Examination.”

- History of Present Illness (HPI): The HPI tells the story of the patient from the time they are admitted to the moment the physician / nurse sees them. In case of a direct admission, it describes the whole outpatient story and what was tried, leading up to why they are being admitted. For a consult, the HPI tells the story of the patient up until the point of why the physician is being consulted.

- Past Medical History (PMH): It documents on-going medical problems, list of surgeries etc. If something is recent or pertinent to the current illness, it is usually described in more detail.

- Chief Complaint: It states the primary medical condition (disease, ailment) of the patient. For example,“Pelvic osteomyelitis and decubitus along with sepsis for IV antibiotics ....”

- Family History: This section describes any relevant medical conditions that run in the family.
- Laboratory Data: This section lists lab measurements and readings. For example, Shows sodium 128 potassium 4.1 chloride 97 bicarb 27.8.
- Assessment and Plan: This section describes the current assessment of the patient's condition outlines the treatment plan. For example, "Sepsis. Will continue IV Primaxin Tobramycin and Zyvox ...."

Generally, when the discharge summary lacks detail, the relevant details can instead be obtained from the "History and Physical" report under similar sections. It naturally follows that "History and Physical" report has multiple coreferring events with discharge summaries.

A.3 Radiology Report

Radiology reports contain observations from scans obtained through x-ray, ultrasound, MRI, CT, and other medical imaging technology. It is the primary means of communication between the radiologist and the referring physician. The measurements in a radiology report may be referenced in other clinical narratives. A sample snippet from a radiology report is as follows: "Left subclavian Groshong catheter tip in the SVC. No pneumothorax. The lungs are clear ...." They usually document the radiologist's observations of the scan reports.

A.4 Pathology Report

A pathology report is written by a physician who specializes in interpreting laboratory tests including evaluation of tissues, cells and organs to diagnose a disease. The report

may describe the lab specimen, its microscopic description and impression or diagnosis. A snippet from a pathology report is as follows: " Cytogenetic analysis of the sample showed a clone of cells with a deletion of 13q. Patients with deletions of 13q and CLL have an good prognosis ...." The description in a pathology report may have some information overlap with lab values in the structured data.

A.5 Social Work Assessment

Social Work Assessment reports are usually written by clinical social workers who assess patients with social, emotional, interpersonal, and socioeconomic issues. Areas of assessment[31] include, but are not limited to: Adjustment to chronic and catastrophic illness or procedures and treatment, ability to follow medical regimen, family functioning, social or financial concerns, abuse, neglect, violence, mental illness and emotional distress, substance abuse, adjustment to loss, cultural, religious, and language needs. These reports are mostly structured. There is usually a comments section that is unstructured. This section gives general details about where the patient currently lives, where he will go after being discharged from hospital, the transportation he will need, where his medical reports need to be sent to, etc. A sample snippet is as follows: "Pt is a non- smoker. Plan is for pt to return home with Interim home health care once medically stable. When pt discharged please complete coc's and fax copies with EDI and scripts ...."

[31] http://www.mghsocialwork.org/aboutus.html

BIBLIOGRAPHY

Allen, J. F. (1981). An interval-based representation of temporal knowledge. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 221–226. xvii, 24, 25, 26, 30, 71, 104, 108, 123

Angeli, G., Manning, C. D., and Jurafsky, D. (2012). Parsing time: Learning to interpret time expressions. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 446–455. Association for Computational Linguistics. 164

Aramaki, E., Imai, T., Miyo, K., and Ohe, K. (2006). Automatic deidentification by using sentence features and label consistency. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, pages 10–11. 10

Aronson, A. R. (2001). Effective mapping of biomedical text to the UMLS metathesaurus: the MetaMap program. *Proceedings of AMIA Symposium*, pages 17–21. 82, 84, 143, 150, 164

Bach, E. (1986). The algebra of events. *Linguistics and Philosophy*, 9(1):5–16. 19

Bagga, A. and Baldwin, B. (1998). Algorithms for scoring coreference chains. In *The First International Conference on Language Resources and Evaluation Workshop on Linguistics Coreference*, pages 563–566. 35

Barzilay, R., Elhadad, N., and McKeown, K. (2002). Inferring strategies for sentence ordering in multidocument summarization. *Journal of Artificial Intelligence Research (JAIR)*, 17:35–55. 119

Barzilay, R. and McKeown, K. R. (2005). Sentence fusion for multidocument news summarization. *Computational Linguistics*, 31(3):297–328. 120

Bean, D. L. and Riloff, E. (2004). Unsupervised learning of contextual role knowledge for coreference resolution. In *Proceedings of the North American Association for Computational Linguistics (HLT-NAACL)*, pages 297–304. 33

Bechhofer, S., Carr, L., Goble, C., Kampa, S., and Miles-Board, T. (2002). The semantics of semantic annotation. *On the Move to Meaningful Internet Systems 2002: CoopIS, DOA, and ODBASE*, pages 1152–1167. 17

Bellare, K., Druck, G., and McCallum, A. (2009). Alternating projections for learning with expectation constraints. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 43–50. 102

Björkelund, A., Farkas, R., Müller, T., and Seeker, W. (2013). (Re)ranking Meets Morphosyntax: State-of-the-art Results from the SPMRL 2013 Shared Task*. Citeseer. 34

Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT)*, pages 92–100. ACM. xix, 90, 95, 96, 97

Bodenreider, O. (2004). The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(suppl 1):D267–D270. 35, 111

Boguraev, B., Pustejovsky, J., Ando, R. K., and Verhagen, M. (2007). Timebank evolution as a community resource for timeml parsing. *Language Resources and Evaluation (LREC)*, 41(1):91–115. 30

Böhlen, M. H. (1995). Temporal database system implementations. *SIGMOD Record*, 24(4):53–60. 27

Boland, M. R., Tu, S. W., Carini, S., Sim, I., and Weng, C. (2012). Elixr-time: A temporal knowledge representation for clinical research eligibility criteria. *AMIA summits on translational science proceedings*, 2012:71. 144

Bollegala, D., Okazaki, N., and Ishizuka, M. (2010). A bottom-up approach to sentence ordering for multi-document summarization. *Information Processing & Management*, 46(1):89–109. 119

Bramsen, P., Deshpande, P., Lee, Y. K., and Barzilay, R. (2006). Inducing temporal graphs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 189–198. 73, 74, 118

Brigitte, B. (2003). Using Kullback-Leibler distance for text categorization. In *Proceedings of the 25th European conference on IR research*, ECIR'03, pages 305–319. 93

Buchanan, B. G. and Shortliffe, E. H. (1984). *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project*. The Addison-Wesley series in Artificial Intelligence. 17, 25

Bunt, H. (2007). The semantics of semantic annotation. In *Proceedings of the 21st Pacific Asia Conference on Language, Information and Computation (PACLIC21)*, pages 13–28. 23

Carlo, L., Chase, H. S., and Weng, C. (2010). Aligning structured and unstructured medical problems using UMLS. In *AMIA Annual Symposium Proceedings*, volume 2010, page 91. 144

Chambers, N. and Jurafsky, D. (2008). Jointly combining implicit constraints improves temporal ordering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 698–706. 31, 106

Chambers, N. and Jurafsky, D. (2009). Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Association for Computational Linguistics (ACL/AFNLP*, pages 602–610. 119

Chambers, N., Wang, S., and Jurafsky, D. (2007). Classifying temporal relations between events. In *Proceedings of the Association for Computational Linguistics (ACL)*. 73, 105

Charniak, E. (1972). Toward a model of children's story comprehension. Technical report, DTIC Document. 32

Chen, Z. and Ji, H. (2010). Graph-based clustering for computational linguistics: A survey. In *Proceedings of the 2010 Workshop on Graph-based Methods for Natural Language Processing*, TextGraphs-5, pages 1–9, Stroudsburg, PA, USA. Association for Computational Linguistics. 33

Chiang, J.-H., Lin, J.-W., and Yang, C.-W. (2010). Automated evaluation of electronic discharge notes to assess quality of care for cardiovascular diseases using Medical Language Extraction and Encoding System (MedLEE). *Journal of the American Medical Informatics Association (JAMIA)*, pages 245–252. 28, 90, 143, 144

Cohen, R., Elhadad, M., and Elhadad, N. (2013). Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. *BMC Bioinformatics*, 14(1):10. 118

Conger, A. (1980). Integration and generalization of kappas for multiple raters. In *Psychological Bulletin Vol 88(2)*, pages 322–328. 60, 64, 134

Dasarathy, B. V. (2001). Information fusion - what, where, why, when, and how? editorial. *Information Fusion*, pages 75–76. 141

Davidson, D. (2001). *Essays on actions and events*, volume 1. Oxford University Press, USA. 18, 35

Demner-Fushman, D., Chapman, W. W., and McDonald, C. J. (2009). What can natural language processing do for clinical decision support? *Journal of Biomedical Informatics*, 42(5):760–772. 118, 143, 149

Do, Q., Lu, W., and Roth, D. (2013). Joint inference for event timeline construction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning (EMNLP-CoNLL)*. 34

Do, Q. X., Lu, W., and Roth, D. (2012). Joint inference for event timeline construction. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, EMNLP-CoNLL '12, pages 677–687. Association for Computational Linguistics. 13, 118, 119, 125, 137, 162

Doddington, G. R., Mitchell, A., Przybocki, M. A., Ramshaw, L. A., Strassel, S., and Weischedel, R. M. (2004). The automatic content extraction (ace) program-tasks, data, and evaluation. In *Proceedings of LREC*. 32, 36

Dowty, D. (1986). The effects of aspectual class on the temporal structure of discourse: semantics or pragmatics? *Linguistics and Philosophy*, 9(1):37–61. 19

Durrett, G. and Klein, D. (2013). Easy victories and uphill battles in coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 34

Fries, J. F. (1972). Time-oriented patient records and a computer databank. *JAMA: The Journal of the American Medical Association*, 222(12):1536–1542. 25

Gaizauskas, R., Harkema, H., Hepple, M., and Setzer, A. (2006). Task-oriented extraction of temporal information: The case of clinical narratives. In *Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning*, TIME '06, pages 188–195. 73

Ganchev, K., Graa, J., Gillenwater, J., and Taskar, B. (2010). Posterior regularization for structured latent variable models. *Journal of Machine Learning Research (JMLR)*, pages 2001–2049. 90, 98

Grishman, R., Sager, N., Raze, C., and Bookchin, B. (1973). The linguistic string parser. In *Proceedings of the National Computer Conference and Exposition*, pages 427–434. ACM. 26

Guo, Y., Gaizauskas, R., Roberts, I., Demetriou, G., and Hepple, M. (2006). Identifying personal health information using support vector machines. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, pages 10–11. 16

Haghighi, A. and Klein, D. (2010). Coreference resolution in a modular, entity-centered model. In *Human Language Technologies: The 2010 Annual Conference of the North*

American Chapter of the Association for Computational Linguistics, pages 385–393. Association for Computational Linguistics. 33, 35

Harkema, H., Setzer, A., Gaizauskas, R., Hepple, M., Power, R., and Rogers, J. (2005). Mining and modelling temporal clinical data. In Cox, S., editor, *Proceedings of the 4th UK e-Science All Hands Meeting*, Nottingham, UK. Available at: http://www.all-hands.org.uk/2005/proceedings/. 163

He, T. Y. (2007). *Coreference Resolution on Entities and Events for Hospital Discharge Summaries*. EECS, Cambridge, MA, MIT. M.Eng. 35, 36, 90

Hobbs, J. R. (1977). Coherence and interpretation in english texts. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 110–116. 22

Hripcsak, G., Elhadad, N., Chen, Y.-H., Zhou, L., and Morrison, F. P. (2009). Research paper: Using empiric semantic correlation to interpret temporal assertions in clinical texts. *Journal of the American Medical Informatics Association (JAMIA)*, pages 220–227. 141

Humphreys, K., Gaizauskas, R., and Azzam, S. (1997). Event coreference for information extraction. In *Proceedings of a Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, ANARESOLUTION '97, pages 75–81, Stroudsburg, PA, USA. Association for Computational Linguistics. 35

Jensen, C. S. and Snodgrass, R. T. (1999). Temporal data management. *IEEE Transactions on Knowledge and Data Engineering*, 11(1):36–44. 29

Ji, H. and Grishman, R. (2008). Refining event extraction through cross-document inference. In *Association for Computational Linguistics*. 120

Jindal, P. and Roth, D. (2013). Extraction of events and temporal expressions from clinical narratives. *Journal of Biomedical Informatics*, 46:S13–S19. 164

Joachims, T. (1999). Making large-scale SVM learning practical. In Schölkopf, B., Burges, C. J. C., and Smola, A. J., editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press. 113

Joachims, T. (2006). Training linear SVMs in linear time. In *Proceedings of the Conference on Knowledge Discovery and Data Mining (KDD)*, pages 217–226. 112

Joachims, T., Li, H., Liu, T.-Y., and Zhai, C. (2007). Learning to rank for information retrieval (lr4ir 2007). *SIGIR Forum*, 41(2):58–62. 107

Jung, H., Allen, J., Blaylock, N., de Beaumont, W., Galescu, L., and Swift, M. (2011). Building timelines from narrative clinical records: initial results based-on deep natural language understanding. In *Proceedings of BioNLP 2011 Workshop*, BioNLP '11, pages 146–154. 7, 71, 73

Kahn, K. and Gorry, G. A. (1977). Mechanizing temporal knowledge. *Artificial Intelligence*, 9(1):87–108. 24

Kalyanpur, A., Boguraev, B., Patwardhan, S., Murdock, J. W., Lally, A., Welty, C., Prager, J. M., Coppola, B., Fokoue-Nkoutche, A., Zhang, L., et al. (2012). Structured data and inference in DeepQA. *IBM Journal of Research and Development*, 56(3.4):10–1. 37, 149

Kamp, H. and Reyle, U. (1993). *From discourse to logic: Introduction to modeltheoretic semantics of natural language, formal logic and discourse representation theory*, volume 42. Kluwer Academic. 20

Kho, A. N., Rasmussen, L. V., Connolly, J. J., Peissig, P. L., Starren, J., Hakonarson, H., and Hayes, M. G. (2013). Practical challenges in integrating genomic data into the electronic health record. *Genetics in Medicine*, 15(10):772–778. 37

Kolya, A. K., Ekbal, A., and Bandyopadhyay, S. (2010). Ju_cse_temp: A first step towards evaluating events, time expressions and temporal relations. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 345–350. Association for Computational Linguistics. 31

Köpcke, F., Trinczek, B., Majeed, R. W., Schreiweis, B., Wenk, J., Leusch, T., Ganslandt, T., Ohmann, C., Bergh, B., Röhrig, R., et al. (2013). Evaluation of data completeness in the electronic health record for the purpose of patient recruitment into clinical trials: a retrospective analysis of element presence. *BMC Medical Informatics & Decision Making*, 13:37. 145, 149

Krstev, C., Vitas, D., Obradović, I., and Utvić, M. (2011). E-dictionaries and Finite-state automata for the recognition of named entities. In *Proceedings of the 9th International Workshop on Finite State Methods and Natural Language Processing*, pages 48–56. 121

Kumar, S. and Byrne, W. (2003). A weighted finite state transducer implementation of the alignment template model for statistical machine translation. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, pages 63–70. 120, 121

Lacatusu, V. F., Maiorano, S. J., and Harabagiu, S. M. (2004). Multi-document summarization using multiple-sequence alignment. In *Proceedings of LREC*. 121

Lapata, M. (2003). Probabilistic text structuring: Experiments with sentence ordering. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*, pages 545–552. Association for Computational Linguistics. 119

Lapata, M. and Lascarides, A. (2006). Learning sentence-internal temporal relations. *Journal of Artificial Intelligence Research (JAIR)*, 27:85–117. 31, 106, 119

Lee, C. M. and Katz, G. (2009). Error analysis of the tempeval temporal relation identification task. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 138–145. Association for Computational Linguistics. 31

Luo, Z., Johnson, S. B., Lai, A. M., and Weng, C. (2011). Extracting temporal constraints from clinical research eligibility criteria using conditional random fields. In *Proceedings of AMIA Symposium*. 7, 118

Malpas, J. E. (1992). *Donald Davidson and the mirror of meaning: holism, truth, interpretation*. CUP Archive. 35

Mani, I. (2005). *Chronoscopes* : A theory of underspecified temporal representations. In *Annotating, Extracting and Reasoning about Time and Events*, pages 127–139. 1

Mani, I., Verhagen, M., Wellner, B., Lee, C. M., and Pustejovsky, J. (2006). Machine learning of temporal relations. In *Proceedings of the Association for Computational Linguistics (ACL)*. 10, 31, 73, 104, 105, 106, 119

Marrocco, C., Duin, R. P., and Tortorella, F. (2008). Maximizing the area under the roc curve by pairwise feature combination. *Pattern Recognition*, 41(6):1961–1974. 112

Martin, J. and Jurafsky, D. (2000). Speech and Language Processing. 15

McDermott, D. (1982). A temporal logic for reasoning about processes and plans*. *Cognitive Science*, 6(2):101–155. 24

Mohri, M., Pereira, F., and Riley, M. (2002). Weighted finite-state transducers in speech recognition. *Computer Speech & Language*, 16(1):69–88. 127, 131, 132

Mohri, M., Pereira, F. C. N., and Riley, M. (2000). The design principles of a weighted finite-state transducer library. *Theoretical Computer Science*, 231(1):17–32. 121

Myers, C. and Habiner, L. (1981). A comparative study of several dynamic time-warping algorithms for connected-word. *Bell System Technical Journal*. 132

Needleman, S., Wunsch, C., et al. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, 48(3):443–453. 118, 125, 132

Ng, V. (2010). Supervised noun phrase coreference research: The first fifteen years. In *Proceedings of the Association for Computational Linguistics (ACL)*, pages 1396–1411. 11, 33, 89

Nigam, K. and Ghani, R. (2000). Analyzing the effectiveness and applicability of co-training. In *Proceedings of CIKM*, pages 86–93. 92

Notredame, C. (2002). Recent progress in multiple sequence alignment: a survey. *Pharmacogenomics*, 3(1):131–144. 120, 124, 132

Obermeier, K. K. (1986). Groka knowledge-based text processing system. In *Proceedings of the 14th Annual Conference on Computer Science*, pages 331–339. ACM. 27

Ogren, P. V. (2006). Knowtator: a protégé plug-in for annotated corpus construction. In *Proceedings of the 2006 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 273–275, Morristown, NJ, USA. Association for Computational Linguistics. 66

Parsons, T. (1990). *Events in the Semantics of English*. MIT press Cambridge, MA:. 18, 21

Pedersen, T. B. and Jensen, C. S. (1998). Research issues in clinical data warehousing. In *Proceedings of the 10th International Conference on Scientific and Statistical Database Management*, SSDBM '98, pages 43–52. IEEE Computer Society. 27

Pustejovsky, J., Castao, J. M., Ingria, R., Sauri, R., Gaizauskas, R. J., Setzer, A., Katz, G., and Radev, D. R. (2003a). TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering'03*, pages 28–34. 29, 38, 39, 56, 73, 74, 104

Pustejovsky, J., Hanks, P., Sauri, R., See, A., Gaizauskas, R., Setzer, A., Radev, D., Sundheim, B., Day, D., Ferro, L., et al. (2003b). The Timebank Corpus. In *Corpus Linguistics*, volume 2003, page 40. 29

Radev, D. R. (2000). A common theory of information fusion from multiple text sources step one: cross-document structure. In *Proceedings of the 1st SIGdial workshop on Discourse and Dialogue - Volume 10*, SIGDIAL '00, pages 74–83. Association for Computational Linguistics. 117

Raghavan, P., Fosler-Lussier, E., and Lai, A. M. (2012). Learning to temporally order medical events in clinical text. In *The 50th Annual Meeting of the Association for Computational Linguistics-Short Papers (ACL Short Papers 2012)*. Association for Computational Linguistics. 121, 135

Raghavan, P. and Lai, A. M. (2010). Leveraging natural language processing of clinical narratives for phenotype modeling. In *3rd Ph.D. Workshop on Information and Knowledge Management (PIKM)*, pages 57–66. 6

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010a). A multi-pass sieve for coreference resolution. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP, pages 492–501, Stroudsburg, PA, USA. Association for Computational Linguistics. 33, 89

Raghunathan, K., Lee, H., Rangarajan, S., Chambers, N., Surdeanu, M., Jurafsky, D., and Manning, C. (2010b). A multi-pass sieve for coreference resolution. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 492–501. Association for Computational Linguistics. 34

Reichenbach, H. and Reichenbach, M. (1956). *The direction of time*, volume 65. Berkeley: University of California Press. 18

Reichert, D., Kaufman, D., Bloxham, B., Chase, H., and Elhadad, N. (2010). Cognitive analysis of the summarization of longitudinal patient records. In *AMIA Annual Symposium Proceedings*, volume 2010, page 667. American Medical Informatics Association. 118

Roberts, A., Gaizauskas, R., Hepple, M., Demetriou, G., Guo, Y., and Setzer, A. (2008). Semantic Annotation of Clinical Text: The CLEF Corpus. In *Proceedings of the LREC 2008 Workshop on Building and Evaluating Resources for Biomedical Text Mining*, pages 19–26. 10, 106, 120

Rosenbloom, S. T., Denny, J. C., Xu, H., Lorenzi, N., Stead, W. W., and Johnson, K. B. (2011). Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association*, 18(2):181–186. 149

Ross, J., Tu, S., Carini, S., and Sim, I. (2010). Analysis of eligibility criteria complexity in clinical trials. *AMIA Summits on Translational Science Proceedings*, 2010:46. 144

Savova, G. K., Bethard, S., Styler, W., Martin, J., Palmer, M., Masanz, J., and Ward, W. (2009). Towards temporal relation discovery from the clinical narrative. *American Medical Informatics Association (AMIA)*. 106

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010a). Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513. 28, 38, 144

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Schuler, K. K., and Chute, C. G. (2010b). Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association (JAMIA)*, pages 507–513. 11, 90, 143

Shahar, Y. (1999). Timing is everything: Temporal reasoning and temporal data maintenance in medicine. In *Proceedings of the Joint European Conference on Artificial Intelligence in Medicine and Medical Decision Making*, AIMDM '99, pages 30–46, London, UK, UK. Springer-Verlag. 27

Shortliffe, E. H., Scott, A. C., Bischoff, M. B., Campbell, A. B., Melle, W. V., and Jacobs, C. D. (1981). ONCOCIN: An expert system for oncology protocol management. pages 876–881. 25

Shu, H. (2006). *Multi-tape finite-state transducer for asynchronous multi-stream pattern recognition with application to speech*. PhD thesis, Massachusetts Institute of Technology. 130

Smith, T. and Waterman, M. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1). 118, 125, 132

Soon, W. M., Ng, H. T., and Lim, C. Y. (2001). A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, pages 521–544. 11, 34, 89

Souza, J. d., Espl-Gomis, M., Turchi, M., and Negri, M. (2013). Exploiting qualitative information from automatic word alignment for cross-lingual nlp tasks. In *The 51st Annual Meeting of the Association for Computational Linguistics-Short Papers (ACL Short Papers 2013)*. 37

Sproat, R. (2006). *A Computational Theory of Writing Systems (Studies in Natural Language Processing)*. Cambridge University Press. 121

Story, G. and Hirschman, L. (1982). Data base design for natural language medical data. *Journal of Medical Systems*, 6(1):77–88. 27

Strötgen, J. and Gertz, M. (2010). Heideltime: High quality rule-based extraction and normalization of temporal expressions. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 321–324. Association for Computational Linguistics. 31

Strube, M. and Ponzetto, S. P. (2006). Wikirelate! computing semantic relatedness using Wikipedia. In *Proceedings of AAAI*, volume 6, pages 1419–1424. 33

Sun, W., Rumshisky, A., and Uzuner, O. (2013). Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813. 29, 144, 163

Tange, H. J., Schouten, H. C., Kester, A. D., and Hasman, A. (1998). The granularity of medical narratives and its effect on the speed and completeness of information retrieval. *Journal of the American Medical Informatics Association*, 5(6):571–582. 149

Tenny, C. and Pustejovsky, J. (2000). A history of events in linguistic theory. *Events as grammatical objects*, 32. 19

Thadani, S. R., Weng, C., Bigger, J. T., Ennever, J. F., and Wajngurt, D. (2009). Electronic screening improves efficiency in clinical trial recruitment. *Journal of the American Medical Informatics Association*, 16(6):869–873. 6

UzZaman, N. and Allen, J. F. (2010). 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. 31

Van Deemter, K. and Kibble, R. (2000). On coreferring: Coreference in MUC and related annotation schemes. *Computational linguistics*, 26(4):629–637. 34, 35, 36

Vendler, Z. (1967). *Linguistics in philosophy*. Cornell University Press Ithaca. xvii, 19, 20

Verhagen, M., Gaizauskas, R. J., Schilder, F., Hepple, M., Moszkowicz, J., and Pustejovsky, J. (2009). The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation (LREC)*, 43(2):161–179. 30, 73, 106, 119

Wang, L. and Jiang, T. (1994). On the complexity of multiple sequence alignment. *Journal of Computational Biology*, 1(4):337–348. 124, 134

Wang, X., Hripcsak, G., Markatou, M., and Friedman, C. (2009). Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *Journal of the American Medical Informatics Association*, 16(3):328–337. 143

Weng, C., Bigger, J. T., Busacca, L., Wilcox, A., and Getaneh, A. (2010). Comparing the effectiveness of a clinical registry and a clinical data warehouse for supporting clinical trial recruitment: a case study. In *AMIA Annual Symposium Proceedings*, volume 2010, page 867. American Medical Informatics Association. 6

Whelan, C., Roark, B., and Sonmez, K. (2010). Designing antimicrobial peptides with weighted finite-state transducers. In *Proceedings of IEEE Engineering in Medical Biology Society*, page 764. 121

Wrenn, J. O., Stein, D. M., Bakken, S., and Stetson, P. D. (2010). Research paper: Quantifying clinical narrative redundancy in an electronic health record. *Journal of the American Medical Informatics Association (JAMIA)*, pages 49–53. 140

Xiang, Y., Lu, K., James, S. L., Borlawsky, T. B., Huang, K., and Payne, P. R. O. (2011). k-neighborhood decentralization: A comprehensive solution to index the UMLS for scale knowledge discovery. In *Journal of Biomedical Informatics*. 91, 92, 93, 99

Yarowsky, D. (1995). Unsupervised word sense disambiguation rivaling supervised methods. In *Association for Computational Linguistics*, pages 189–196. 120

Zheng, J., Chapman, W. W., Miller, T. A., Lin, C., Crowley, R. S., and Savova, G. K. (2012). A system for coreference resolution for the clinical narrative. *Journal of the American Medical Informatics Association*. 11, 90

Zhou, L. and Hripcsak, G. (2007). Temporal reasoning with medical data - a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, pages 183–202. 3, 27, 28, 55, 71, 73, 83, 94, 106, 144, 163

Zhou, L., Melton, G. B., Parsons, S., and Hripcsak, G. (2006). A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, pages 424–439. 10, 28, 38, 53, 73, 84, 106, 108, 113, 120, 144, 164