Temporal Annotation in the Clinical Domain

William F. Styler IV[1], **Steven Bethard**[2], **Sean Finan**[3], **Martha Palmer**[1], **Sameer Pradhan**[3], **Piet C de Groen**[4], **Brad Erickson**[4], **Timothy Miller**[3], **Chen Lin**[3], **Guergana Savova**[3] and **James Pustejovsky**[5]

[1] Department of Linguistics, University of Colorado at Boulder
[2] Department of Computer and Information Sciences, University of Alabama at Birmingham
[3] Children's Hospital Boston Informatics Program and Harvard Medical School
[4] Mayo Clinic College of Medicine, Mayo Clinic, Rochester, MN
[5] Department of Computer Science, Brandeis University

Abstract

This article discusses the requirements of a formal specification for the annotation of temporal information in clinical narratives. We discuss the implementation and extension of ISO-TimeML for annotating a corpus of clinical notes, known as the THYME corpus. To reflect the information task and the heavily inference-based reasoning demands in the domain, a new annotation guideline has been developed, "the THYME Guidelines to ISO-TimeML (THYME-TimeML)". To clarify what relations merit annotation, we distinguish between linguistically-derived and inferentially-derived temporal orderings in the text. We also apply a top performing TempEval 2013 system against this new resource to measure the difficulty of adapting systems to the clinical domain. The corpus is available to the community and has been proposed for use in a SemEval 2015 task.

1 Introduction

There is a long-standing interest in temporal reasoning within the biomedical community (Savova et al., 2009; Hripcsak et al., 2009; Meystre et al., 2008; Bramsen et al., 2006; Combi et al., 1997; Keravnou, 1997; Dolin, 1995; Irvine et al., 2008; Sullivan et al., 2008). This interest extends to the automatic extraction and interpretation of temporal information from medical texts, such as electronic discharge summaries and patient case summaries. Making effective use of temporal information from such narratives is a crucial step in the intelligent analysis of informatics for medical researchers, while an awareness of temporal information (both implicit and explicit) in a text is also necessary for many data mining tasks.

It has also been demonstrated that the temporal information in clinical narratives can be usefully mined to provide information for some higher-level temporal reasoning (Zhao et al., 2005). Robust temporal understanding of such narratives, however, has been difficult to achieve, due to the complexity of determining temporal relations among events, the diversity of temporal expressions, and the interaction with broader computational linguistic issues.

Recent work on Electronic Health Records (EHRs) points to new ways to exploit and mine the information contained therein (Savova et al., 2009; Roberts et al., 2009; Zheng et al., 2011; Turchin et al., 2009). We target two main use cases for extracted data. First, we hope to enable interactive displays and summaries of the patient's records to the physician at the time of visit, making a comprehensive review of the patient's history both faster and less prone to oversights. Second, we hope to enable temporally-aware secondary research across large databases of medical records (e.g., "What percentage of patients who undergo procedure X develop side-effect Y within Z months?"). Both of these applications require the extraction of time and date associations for critical events and the relative ordering of events during the patient's period of care, all from the various records which make up a patient's EHR. Although we have these two specific applications in mind, the schema we have developed is generalizable and could potentially be embedded in a wide variety of biomedical use cases.

Narrative texts in EHRs are temporally rich documents that frequently contain assertions about the timing of medical events, such as visits, laboratory values, symptoms, signs, diagnoses, and procedures (Bramsen et al., 2006; Hripcsak et al., 2009; Zhou et al., 2008). Temporal representation and reasoning in the medical record are difficult due to: (1) the diversity of time expressions; (2) the complexity of determining temporal relations among events (which are often left to inference); (3) the difficulty of handling the temporal granularity of an event; and (4)

Transactions of the Association for Computational Linguistics, 2 (2014) 143–154. Action Editor: Ellen Riloff.
Submitted 9/2013; Revised 2/2014; Published 4/2014.

general issues in natural language processing (e.g., ambiguity, anaphora, ellipsis, conjunction). As a result, the signals used for reconstructing a timeline can be both domain-specific and complex, and are often left implicit, requiring significant domain knowledge to accurately detect and interpret.

In this paper, we discuss the demands on accurately annotating such temporal information in clinical notes. We describe an implementation and extension of ISO-TimeML (Pustejovsky et al., 2010), developed specifically for the clinical domain, which we refer to as the "THYME Guidelines to ISO-TimeML" ("THYME-TimeML"), where THYME stands for "Temporal Histories of Your Medical Events". A simplified version of these guidelines formed the basis for the 2012 i2b2 medical-domain temporal relation challenge (Sun et al., 2013a).

This is being developed in the context of the THYME project, whose goal is to both create robust gold standards for semantic information in clinical notes, as well as to develop state-of-the-art algorithms to train and test on this dataset.

Deriving timelines from news text requires the concrete realization of context-dependent assumptions about temporal intervals, orderings and organization, underlying the explicit signals marked in the text (Pustejovsky and Stubbs, 2011). Deriving patient history timelines from clinical notes also involves these types of assumptions, but there are special demands imposed by the characteristics of the clinical narrative. Due to both medical shorthand practices and general domain knowledge, many event-event relations are not signaled in the text at all, and rely on a shared understanding and common conceptual models of the progressions of medical procedures available only to readers familiar with language use in the medical community.

Identifying these implicit relations and temporal properties puts a heavy burden on the annotation process. As such, in the THYME-TimeML guideline, considerable effort has gone into both describing and proscribing the annotation of temporal orderings that are inferable only through domain-specific temporal knowledge.

Although the THYME guidelines describe a number of departures from the ISO-TimeML standard for expediency and ease of annotation, this paper will focus on those differences specifically motivated by the needs of the clinical domain, and on the consequences for systems built to extract temporal data in both the clinical and general domain.

2 The Nature of Clinical Documents

In the THYME corpus, we have been examining 1,254 de-identified[1] notes from a large healthcare practice (the Mayo Clinic), representing two distinct fields within oncology: brain cancer, and colon cancer. To date, we have principally examined two different general types of clinical narrative in our EHRs: clinical notes and pathology reports.

Clinical notes are records of physician interactions with a patient, and often include multiple, clearly delineated sections detailing different aspects of the patient's care and present illness. These notes are fairly generic across institutions and specialities, and although some terms and inferences may be specific to a particular type of practice (such as oncology), they share a uniform structure and pattern. The 'History of Present Illness', for example, summarizes the course of the patient's chief complaint, as well as the interventions and diagnostics which have been thus far attempted. In other sections, the doctor may outline her current plan for the patient's treatment, then later describe the patient's specific medical history, allergies, care directives, and so forth.

Most critically for temporal reasoning, each clinical note reflects a single time in the patient's treatment history at which all of the doctor's statements are accurate (the DOCTIME), and each section tends to describe events of a particular timeframe. For example, 'History of Present illness' predominantly describes events occuring before DOCTIME, whereas 'Medications' provides a snapshot at DOCTIME and 'Ongoing Care Orders' discusses events which have not yet occurred.[2]

Clinical notes contain rich temporal information and background, moving fluidly from prior treatments and symptoms to present conditions to future interventions. They are also often rich with hypothetical statements ("if the tumor recurs, we can..."), each of which can form its own separate timeline.

By constrast, pathology notes are quite different. Such notes are generated by a medical pathologist

[1] Although most patient information was removed, dates and temporal information were not modified according to this project's specific data use agreement.

[2] One complication is the propensity of doctors and automated systems to later update sections in a note without changing the timestamp or metadata. We have added a SECTIONTIME to keep these updated sections from affecting our overall timeline.

upon receipt and analysis of specimens (ranging from tissue samples from biopsy to excised portions of tumor or organs). Pathology notes provide crucial information to the patient's doctor confirming the malignancy (cancer) in samples, describing surgical margins (which indicate whether a tumor was completely excised), and classifying and 'staging' a tumor, describing the severity and spread of the cancer. Because the information in such notes pertains to samples taken at a single moment in time, they are temporally sparse, seldom referring to events before or after the examination of the specimen. However, they contain critical information about the state of the patient's illness and about the cancer itself, and must be interpreted to understand the history of the patient's illness.

Most importantly, in all EHRs, we must contend with the results of a fundamental tension in modern medical records: hyper-detailed records provide a crucial defense against malpractice litigation, but including such detail takes enormous time, which doctors seldom have. Given that these notes are written by and for medical professionals (who form a relatively insular speech community), a great many non-standard expressions, abbreviations, and assumptions of shared knowledge are used, which are simultaneously concise and detail-rich for others who have similar backgrounds.

These time-saving devices can range from temporally loaded acronyms (e.g., 'qid', Latin for *quater in die*, 'four times daily'), to assumed orderings (a diagnostic test for a disorder is assumed to come before the procedure which treats it), and even to completely implicit events and temporal details. For example, consider the sentence in (1).

(1) **Colonoscopy** 3/12/10, nodule **biopsies** negative

We must understand that during the colonoscopy, the doctor obtained biopsies of nodules, which were packaged and sent to a pathologist, who reviewed them and determined them to be 'negative' (non-cancerous).

In such documents, we must recover as much temporal detail as possible, even though it may be expressed in a way which is not easily understood outside of the medical community, let alone by linguists or automated systems. We must also be aware of the legal relevance of some events (e.g., "We **discussed** the possible side effects"), even when they may not seem relevant to the patient's actual care.

Finally, each specialty and note type has separate conventions. Within colon cancer notes, the American Joint Committee on Cancer (AJCC) Staging Codes (e.g., *T4N1*, indicating the nature of the tumor, lymph node and metastasis involvement) are meticulously recorded, but are largely absent in the brain cancer notes which make up the second corpus in our project. So, although clinical notes share many similarities, annotators without sufficient domain expertise may require additional training to adapt to the inferences and nuances of a new clinical subdomain.

3 Interpreting 'Event' and Temporal Expressions in the Clinical Domain

Much prior work has been done on standardizing the annotation of events and temporal expressions in text. The most widely used approach is the ISO-TimeML specification (Pustejovsky et al., 2010), an ISO standard that provides a common framework for annotating and analyzing time, events, and event relations. As defined by ISO-TimeML, an EVENT refers to anything that can be said "to obtain or hold true, to happen or to occur". This is a broad notion of event, consistent with Bach's use of the term "eventuality" (Bach, 1986) as well as the notion of fluents in AI (McCarthy, 2002).

Because the goals of the THYME project involve automatically identifying the clinical timeline for a patient from clincal records, the scope of what should be admitted into the domain of events is interpreted more broadly than in ISO-TimeML[3]. Within the THYME-TimeML guideline, an EVENT is *anything* relevant to the clinical timeline, i.e., anything that would show up on a detailed timeline of the patient's care or life. The best single-word syntactic head for the EVENT is then used as its span. For example, a *diagnosis* would certainly appear on such a timeline, as would a *tumor*, *illness*, or *procedure*. On the other hand, entities that persist throughout the relevant temporal period of the clinical timeline (*endurants* in ontological circles) would not be considered as event-like. This includes the patient, other humans mentioned (the patient's mother-in-law or the doctor), organizations (the emergency room), non-anatomical objects (the patient's car), or individual parts of the patient's anatomy (an arm is not an EVENT unless missing or otherwise notable).

To meet our explicit goals, the THYME-TimeML guideline introduces two additional levels of interpre-

[3]Our use of the term 'EVENT' corresponds with the less specific ISO-TimeML term 'Eventuality'

tation beyond that specified by ISO-TimeML: (i) a well-defined task; and (ii) a clearly identified domain. By focusing on the creation of a *clinical timeline* from *clinical narrative*, the guideline imposes constraints that cannot be assumed for a broadly defined and domain independent annotation schema.

Some EVENTs annotated under our guideline are considered meaningful and eventive mostly by virtue of a specific clinical or legal value. For example, AJCC Staging Codes (discussed in Section 2) are eventive only in the sense of the code being assigned to a tumor at a given moment in the patient's care. However, they are of such critical importance and informative value to doctors that we have chosen to annotate them *specifically so that they will show up on the patient's timeline in a clinical setting*.

Similarly, because of legal pressures to establish informed consent and patient knowledge of risk, entire paragraphs of clinical notes are dedicated to documenting the doctor's discussion of risks, plans, and alternative strategies. As such, we annotate verbs of discussion ("We **talked** about the risks of this drug"), consent ("She **agreed** with the current plan"), and comprehension ("Mrs. Larsen **repeated** the potential side effects back to me"), even though they are more relevant to legal defense than medical treatment.

It is also because of this grounding in clinical language that entities and other non-events are often interpreted in terms of their associated eventive properties. There are two major types for which this is a significant shift in semantic interpretation:

(2) a Medication as Event:
Orders: Lariam twice daily.
b Disorder as Event:
Tumor of the left lung.

In both these cases, entities which are not typically marked as events are identified as such, because they contribute significant information to the clinical timeline being constructed. In (2a), for example, the TIMEX3 "twice daily" is interpreted as scoping over the eventuality of the patient *taking* the medication, not the prescription event. In sentence (2b), the "tumor" is interpreted as a stative eventuality of the patient *having* a tumor located within an anatomical region, rather than an entity within an entity.

Within the medical domain, these eventive interpretations of medications, growths and status codes are unambiguous and consistent. Doctors in clinical notes (unlike in biomedical research texts) do not discuss medications without an associated (implicit) administering EVENT (though some mentions may be hypothetical, generic or negated). Similarly, mentions of symptoms or disorders reflect occurrences in a patient's life, rather than abstract entities. With these interpretations in mind, we can safely infer, for instance, that all UMLS (Unified Medical Language System, (Bodenreider, 2004)) entities of the types Disorder, Chemical/Drug, Procedure and Sign/Symptom will be EVENTs.

In general, in the medical domain, it is essential to read "between the lines" of the shorthand expressions used by the doctors, and recognize implicit events that are being referred to by specific anatomical sites or medications.

4 Modifications to ISO-TimeML for the Clinical Domain

Overall, we have found that the specification required for temporal annotation in the clinical domain does not require substantial modification from existing specifications for the general domain. The clinical domain includes no shortage of inferences, shorthands, and unusual use of language, but the structure of the underlying timeline is not unique.

As a result of this, we have been able to adopt most of the framework from ISO-TimeML, adapting the guidelines where needed, as well as reframing the focus of what gets annotated. This is reflected in a comprehensive guideline, incorporating the specific patterns and uses of events and temporal expressions as seen in clinical data. This approach allows the resulting annotations to be interoperable with existing solutions, while still accommodating the major differences in the nature of the texts. Our guidelines, as well as the annotated data, are available at `http://thyme.healthnlp.org`[4]

Our extensions of the ISO-TimeML specification to the clinical domain are intended to address specific constructions, meanings, and phenomena in medical texts. Our schema differs from ISO-TimeML in a few notable ways.

EVENT Properties We have both simplified the ISO-TimeML coding of EVENTs, and extended it to meet the needs of the clinical domain and the specific language goals of the clinical narrative.

[4]Access to the corpus will require a data use agreement. More information about this process is available from the corpus website.

Consider, for example, how modal subordination is handled in ISO-TimeML. This involves the semantic characterization of an event as "likely", "possible", or as presented by observation, evidence, or hearsay. All of these are accounted for compositionally in ISO-TimeML within the SLINK (Subordinating Link) relation (Pustejovsky et al., 2005). While accepting ISO-TimeML's definition of event modality, we have simplified the annotation task within the current guideline, so that EVENTs now carry attributes for "contextual modality", "contextual aspect" and "permanence".

Contextual modality allows the values ACTUAL, HYPOTHETICAL, HEDGED, and GENERIC. ACTUAL covers EVENTs which have actually happened, e.g., "We've noted a tumor". HYPOTHETICAL covers conditionals and possibilities, e.g., "If she develops a tumor". HEDGED is for situations where doctors proffer a diagnosis, but do so cautiously, to avoid legal liability for an incorrect diagnosis or for overlooking a correct one. For example:

(3) a. The signal in the MRI is not inconsistent with a **tumor** in the spleen.
b. The rash appears to be **measles**, awaiting antibody test to confirm.

These HEDGED EVENTs are more real than a hypothetical diagnosis, and likely merit inclusion on a timeline as part of the diagnostic history, but must not be conflated with confirmed fact. These (and other forms of uncertainty in the medical domain) are discussed extensively in (Vincze et al., 2008). In contrast, GENERIC EVENTs do not refer to the patient's illness or treatment, but instead discuss illness or treatment in general (often in the patient's specific demographic). For example:

(4) In other patients without significant **comorbidity** that can **tolerate** adjuvant **chemotherapy**, there is a **benefit** to systemic adjuvant **chemotherapy**.

These sections would be true if pasted into any patient's note, and are often identical chunks of text repeatedly used to justify a course of action or treatment as well as to defend against liability.

Contextual Aspect (to distinguish from grammatical aspect), allows the clinically-necessary category, INTERMITTENT. This serves to distinguish intermittent EVENTs (such as vomiting or seizures) from constant, more stative EVENTs (such as fever or soreness). For example, the bolded EVENT in (5a) would be marked as INTERMITTENT, while that in (5b) would not:

(5) a She has been **vomiting** since June.
b She has had **swelling** since June.

In the first case, we assume that her vomiting has been intermittent, i.e., there were several points since June in which she was not actively vomiting. In the second case, unless made otherwise explicit ("she has had occasional swelling"), we assume that swelling was a constant state. This property is also used when a particular instance of an EVENT is intermittent, even though it generally would not be:

(6) Since starting her new regime, she has had occasional bouts of **fever**, but is feeling much better.

The permanence attribute has two values, FINITE and PERMANENT. Permanence is a property of diseases themselves, roughly corresponding to the medical concept of "chronic" vs. "acute" disease, which marks whether a disease is persistent following diagnosis. For example, a (currently) uncurable disease like Multiple Sclerosis would be classed as PERMANENT, and thus, once mentioned in a patient's note, will be assumed to persist through the end of the patient's timeline. This is compared with FINITE disorders like "Influenza" or "fever", which, if not mentioned in subsequent notes, should be considered cured and no longer belongs on the patient's timeline. Because it requires domain-specific knowledge, although present in the specification, Permanence is not currently annotated. However, annotators are trained on the basic idea and told about subsequent axiomatic assignment. The addition of this property to our schema is designed to relieve annotators of any feeling of obligation to express this inferred information in some other way.

TIMEX3 Types Temporal expressions (TIMEX3s) in the clinical domain function the same as in the general linguistic community, with two notable exceptions. ISO-TimeML SETs (statements of frequency) occur quite frequently in the medical domain, particularly with regard to medications and treatments. Medication sections within notes often contain long lists of medications, each with a particular associated set ("Claritin 30mg *twice daily*"), and further temporal specification is not uncommon (e.g., "three times per day at meals", "once a week at bedtime").

The second major change for the medical domain is a new type of TIMEX3 which we call PREPOSTEXP. This covers temporally complex terms like

"preoperative", "postoperative", and "intraoperative". These temporal expressions designate a span of time bordered, usually only on one side, by the incorporated event (an operation, in the previous EVENTs). In many cases, the referent is clear:

(7) She underwent **hemicolectomy** *last week*, and had some postoperative **bleeding**.

Here we understand that "postoperative" refers to "the period of time following the hemicolectomy". In these cases, the PREPOSTEXP makes explicit a temporal link between the bleeding and the hemicolectomy. In other cases, no clear referent is present:

(8) Patient shows some **post-procedure** scarring.

In these situations, where no procedure is mentioned (or the reference is never explicitly resolved), we treat the PREPOSTEXP as a narrative container (see Section 5), covering the span of time following the unnamed procedure.

Finally, it is worth noting that the process of normalizing those TIMEX3s is significantly more complex relative to the general domain, because many temporal expressions are anchored not to dates or times, but to other EVENTs (whose dates are often not mentioned or not known by the physician). As we move towards a complete system, we are working to expand the ISO-TimeML system for TIMEX3 normalization to allow some value to be assigned to a phrase like "in the months after her hemicolectomy" when no referent date is present. ISO-TimeML, in discussion with ISO TC 37SC 4, plans to reference to such TIMEX3s in a future release of the standard.

5 Temporal Ordering and Narrative Containers

The semantic content and informational impact of a timeline is encoded in the ordering relations that are identified between the temporal and event expressions present in clinical notes. ISO-TimeML specifies the standard thirteen "Allen relations" from the interval calculus (Allen, 1983), which it refers to as TLINK values. For unguided, general-purpose annotation, the number of relations that could be annotated grows quadratically with the number of events and times, and the task quickly becomes unmanageable. There are, however, strategies that we can adopt to make this labeling task more tractable. Temporal ordering relations in text are of three kinds:

1. Relations between two events
2. Relations between two times
3. Relations between a time and an event.

ISO-TimeML, as a formal specification of the temporal information conveyed in language, makes no distinction between these ordering types. Humans, however, do make distinctions, based on local temporal markers and the discourse relations established in a narrative (Miltsakaki et al., 2004; Poesio, 2004).

Because of the difficulty of humans capturing every relationship present in the note (and the disagreement which arises when annotators attempt to do so), it is vital that the annotation guidelines describe an approach that reduces the number of relations that must be considered, but still results in maximally informative temporal links. We have found that many of the weaknesses in prior annotation approaches stem from interaction between two competing goals:

- The guideline should specify certain types of annotations that *should* be performed;
- The guideline should not force annotations to be performed when they need not be.

Failing in the first goal will result in under-annotation and the neglect of relations which provide necessary information for inference and analysis. Failure in the second goal results in over-annotation, creating complex webs of temporal relations which yield mostly inferable information, but which complicate annotation and adjudication considerably.

Our method of addressing both goals in temporal relations annotation is that of the narrative container, discussed in Pustejovsky and Stubbs (2011). A narrative container can be thought of as a temporal bucket into which an EVENT or series of EVENTs may fall, or a natural cluster of EVENTs around a given time or situation. These narrative containers are often represented (or "anchored") by dates or other temporal expressions (within which a variety of different EVENTs occur), although they can also be anchored to more abstract concepts ("recovery" which might involve a variety of EVENTs) or even durative EVENTs (many other EVENTs can occur during a surgery). Rather than marking every possible TLINK between each EVENT, we instead try to link all EVENTs to their narrative containers, and then link those containers so that the contained EVENTs can be linked by inference.

First, annotators assign each event to one of four broad narrative containers: before the DOCTIME, before and overlapping the DOCTIME, just overlapping the DOCTIME or after the DOCTIME. This narrative

container is identified by the EVENT attribute DocTimeRel. After the assignment of DocTimeRel, the remainder of the narrative container relations must be specified using temporal links (TLINKs). There are five different temporal relations used for such TLINKs: BEFORE, OVERLAP, BEGINS-ON, ENDS-ON and CONTAINS[5]. Due to our narrative container approach, CONTAINS is the most frequent relation by a large margin.

EVENTs serving as narrative container anchors are not tagged as containers per-se. Instead, annotators use the narrative container idea to help them visualize the temporal relations within a document, and then make a series of CONTAINS TLINK annotations which establish EVENTs and TIMEX3s as anchors, and specify their contents. If the annotators do their jobs correctly, properly implementing DocTimeRel and creating accurate TLINKs, a good understanding of the narrative containers present in a document will naturally emerge from the annotated text.

The major advantage introduced with narrative containers is this: a narrative event is placed *within* a bounding temporal interval which is explicitly mentioned in the text. This allows EVENTs within separate containers to be linked by post-hoc inference, temporal reasoning, and domain knowledge, rather than by explicit (and time-consuming) one-by-one temporal relations annotation.

A secondary advantage is that this approach works nicely with the general structure of story-telling in both the general and clinical domains, and provides a compelling and useful metaphor for interpreting timelines. Often, especially in clinical histories, doctors will cluster discussions of symptoms, interventions and diagnoses around a given date (e.g. a whole paragraph starting "June 2009:"), a specific hospitalization ("During her January stay at Mercy"), or a given illness or treatment ("While she underwent Chemo"). Even when specific EVENTs are not explicitly ordered within a cluster (often because the order can be easily inferred with domain knowledge), it is often quite easy to place the EVENTs into containers, and just a few TLINKs can order the containers relative to one another with enough detail to create a clinically useful understanding of the overall timeline.

Narrative containers also allow the inference of relations between sub-events within nested containers:

(9) December 19th: The patient underwent an **MRI** and **EKG** as well as emergency **surgery**. During the **surgery**, the patient experienced mild **tachycardia**, and she also **bled** significantly during the initial **incision**.

1. December 19th CONTAINS **MRI**
2. December 19th CONTAINS **EKG**
3. December 19th CONTAINS **surgery**

a. **surgery** CONTAINS **tachycardia**
b. **surgery** CONTAINS **incision**
c. **incision** CONTAINS **bled**

Through our container nesting, we can automatically infer that 'bled' occurred on December 19th (because '19th' CONTAINS 'surgery' which CONTAINS 'incision' which CONTAINS 'bled'). This also allows the capture of EVENT/sub-event relations, and the rapid expression of complex temporal interactions.

6 Explicit vs. Inferable Annotation

Given a specification language, there are essentially two ways of introducing the elements into the document (data source) being annotated:[6]

- Manual annotation: Elements are introduced into the document directly by the human annotator following the guideline.
- Automatic (inferred) annotation: Elements are created by applying an automated procedure that introduces new elements that are derivable from the human annotations.

As such, there is a complex interaction between specification and guideline, and we focus on how the clinical annotation task has helped shape and refine the annotation guidelines. It is important to note that an annotation guideline does not necessarily force the markup of certain elements in a text, even though the specification language (and the eventual goal of the project) might require those annotations to exist.

In some cases, these added annotations are derived logically from human annotations. Explicitly marked temporal relations can be used to infer others that are not marked but exist implicitly through closure. For instance, given EVENTs A, B and C and TLINKs 'A BEFORE B' and 'B BEFORE C', the TLINK 'A BEFORE C' can be automatically inferred. Repeatedly applying such inference rules allows all inferable

[5]This is a subset of the ISO-TimeML TLINK types, excluding those seldom occurring in medical records, like 'simultaneous' as well as inverse relations like 'during' or 'after'.

[6]We ignore the application of automatic techniques, such as classifiers trained on external datasets, as our focus here is on the preparation of the gold standard used for such classifiers.

TLINKs to be generated (Verhagen, 2005). We can use this idea of closure to show our annotators which annotations need not be marked explicitly, saving time and effort. We have also incorporated these closure rules into our inter-annotator agreement (IAA) calculation for temporal relations, described further in Section 7.2.

The automatic application of rules following the annotation of the text is not limited to the marking of logically inferable relations or EVENTs. In the clinical domain, the combination of within-group shared knowledge and pressure towards concise writing leads to a number of common, inferred relations. Take, for example, the sentence:

(10) Jan 2013: **Colonoscopy**, **biopsies**. **Pathology** showed **adenocarcinoma**, **resected** at Mercy. **Diagnosis T3N1 Adenocarcinoma**.

In this sentence, only the CONTAINS relations between "Jan 2013" and the EVENTs (in bold) are explicitly stated. However, based on the known progression-of-care for colon cancer, we can infer that the colonoscopy occurs first, biopsies occur during the colonoscopy, pathology happens afterwards, a diagnosis (here, adenocarcinoma) is returned after pathology, and resection of the tumor occurs after diagnosis. The presence of the AJCC staging information in the final sentence (along with the confirmation of the adenocarcinoma diagnosis) implies a post-surgical pathology exam of the resected specimen, as the AJCC staging information cannot be determined without this additional examination.

These inferences come naturally to domain experts but are largely inaccessible to people outside the medical community without considerable annotator training. Making explicit our understanding of these "understood orderings" is crucial; although they are not marked by human annotators in our schema, the annotators often found it initially frustrating to leave these (purely inferential) relations unstated. Although many of our (primarily linguistically trained) annotators learned to see these patterns, we chose to exclude them from the manual task since newer annotators with varying degrees of domain knowledge may struggle if asked to manually annotate them.

Similar unspoken-but-understood orderings are found throughout the clinical domain. As mentioned in Section 3, both Permanence and Contextual Aspect:Intermittent are properties of symptoms and diseases themselves, rather than of the patient's particular situation. As such, these properties could easily be identified and marked across a medical ontology, and then be automatically assigned to EVENTs recognized as specific medical named entities.

Finally, due to the peculiarities of EHR systems, some annotations *must* be done programatically. Exact dates of patient visit (or of pathology/radiology consult) are often recorded as metadata on the EHR itself, rather than within the text, making the canonical DOCTIME (or time of automatic section modifications) difficult to access in de-identified plaintext data, but easy to find automatically.

Annotation Type	Raw Count
EVENT	15,769
TIMEX3	1,426
LINK	7935
Total	25,130

Table 1: Raw Frequency of Annotation Types

TLINK Type	Raw Count	% of TLINKs
CONTAINS	5,112	64.42%
OVERLAP	1,205	15.19%
BEFORE	1,004	12.65%
BEGINS-ON	488	6.15%
ENDS-ON	126	1.59%
Total	7,935	100.00%

Table 2: Relative Frequency of TLINK types

7 Results

We report results on the annotations from the here-released subset of the THYME colon cancer corpus, which includes clinical notes and pathology reports for 35 patients diagnosed with colon cancer for a total of 107 documents. Each note was annotated by a pair of graduate or undergraduate students in Linguistics at the University of Colorado, then adjudicated by a domain expert. These clinical narratives were sampled from the EHRs of a major healthcare center (the Mayo Clinic). They were deidentified for all patient-sensitive information; however, original dates were retained.

7.1 Descriptive Statistics

Table 1 presents the raw counts for events, temporal expressions and links in the adjudicated gold annotations. Table 2 presents the number and percentage of TLINKs by type in the adjudicated relations gold annotations.

Annotation Type	F1-Score	Alpha
EVENT	0.8038	0.7899
TIMEX3	0.8047	0.6705
LINK: Participants only	0.5012	0.4999
LINK: Participants+type	0.4506	0.4503
LINK: CONTAINS	0.5630	0.5626

Table 3: IAA (F1-Score and Alpha) by annotation type

EVENT Property	F1-Score	Alpha
DocTimeRel	0.7189	0.6889
Cont.Aspect	0.9947	0.9930
Cont.Modality	0.9547	0.9420

Table 4: IAA (F1-Score and Alpha) for EVENT properties

7.2 Inter-annotator Agreement

We report inter-annotator agreement (IAA) results on the THYME corpus. Each note was annotated by two independent annotators. The final gold standard was produced after disagreement adjudication by a third annotator was performed.

We computed the IAA as F1-score and Krippendorff's Alpha (Krippendorff, 2012) by applying closure, using explicitly marked temporal relations to identify others that are not marked but exist implicitly. In the computation of the IAA, inferred-only TLINKs do not contribute to the score, matched or unmatched. For instance, if both annotators mark A BEFORE B and B BEFORE C, to prevent artificially inflating the agreement score, the inferred A BEFORE C is ignored. Likewise, if one annotator marked A BEFORE B and B BEFORE C and the other annotator did not, the inferred A BEFORE C is not counted. However, if one annotator did explicitly mark A BEFORE C, then an equivalent inferred TLINK would be used to match it. EVENT and TIMEX3 IAA was generated based on exact and overlapping spans, respectively. These results are reported in Table 3.

The THYME corpus also differs from ISO-TimeML in terms of EVENT properties, with the addition of DocTimeRel, ContextualModality and ContextualAspect. IAA for these properties is in Table 4.

7.3 Baseline Systems

To get an idea of how much work will be necessary to adapt existing temporal information extraction systems to the clinical domain, we took the freely available ClearTK-TimeML system (Bethard, 2013),

	TempEval 2013			THYME Corpus		
	P	R	F_1	P	R	F_1
TIMEX3	83.2	71.7	77.0	59.3	42.8	49.7
EVENT	81.4	76.4	78.8	78.9	23.9	36.6
DocTimeRel	-	-	-	47.4	47.4	47.4
LINK[7]	28.6	30.9	26.6	22.7	18.6	20.4
EVENT-TIMEX3	-	-	-	32.3	60.7	42.1
EVENT-EVENT	-	-	-	7.0	3.0	4.2

Table 5: Performance of ClearTK-TimeML models, as reported in the TempEval 2013 competition, and as applied to the THYME Corpus development set.

which was among the top performing systems in TempEval 2013 (UzZaman et al., 2013), and evaluated its performance on the THYME corpus.

ClearTK-TimeML uses support vector machine classifiers trained on the TempEval 2013 training data, employing a small set of features including character patterns, tokens, stems, part-of-speech tags, nearby nodes in the constituency tree, and a small time word gazetteer. For EVENTs and TIMEX3s, the ClearTK-TimeML system could be applied directly to the THYME corpus. For DocTimeRels, the relation for an EVENT was taken from the TLINK between that EVENT and the document creation time, after mapping INCLUDES to OVERLAP. EVENTs with no such TLINK were assumed to have a DocTimeRel of OVERLAP. For other temporal relations, INCLUDES was mapped to CONTAINS.

Results of this system on TempEval 2013 and the THYME corpus are shown in Table 5. For time expressions, performance when moving to the clinical data degrades about 25%, from F_1 of 77.0 to 49.7. For events, the degradation is much larger, about 40%, from 78.8 to 36.6, most likely because of the large number of clinical symptoms, diseases, disorders, etc. which have never been observed by the system during training. Temporal relations are a bit more difficult to compare because TempEval lumped DocTimeRel and other temporal relations together and had several differences in their evaluation metric[7]. However, we at least can see that performance of the ClearTK-TimeML system on temporal relations is low on clinical text, achieving only F_1 of 20.4.

These results suggest that clinical narratives do

[7]The TempEval 2013 evaluation metric penalized systems for parts of the text that were not examined by annotators, and used different variants of closure-based precision and recall.

indeed present new challenges for temporal information extraction systems, and that having access to domain specific training data will be crucial for accurate extraction in the clinical domain. At the same time, it is encouraging that we were able to apply existing ISO-TimeML-based systems to our corpus, despite the several extensions to ISO-TimeML that were necessary for clinical narratives.

8 Discussion

CONTAINS plays a large role in the THYME corpus, representing 66% of TLINK annotations made, compared with only 14.6% for OVERLAP, the second most frequent type. We also see that BEFORE links are relatively less common than OVERLAP and CONTAINS, illustrating that much of the temporal ordering on the timeline is accomplished by using many vertical links (CONTAINS, OVERLAP) to build containers, and few horizontal links (BEFORE, BEGINS-ON, ENDS-ON) to order them.

IAA on EVENTs and Temporal Expressions is strong, although differentiating implicit EVENTs (which should not be marked) from explicit, markable EVENTs remains one of the biggest sources of disagreement. When compared to the data from the 2012 i2b2 challenge (Sun et al., 2013b), our IAA figures are quite similar. Even with our more complex schema, we achieved an F1-score of 0.8038 for EVENTs (compared to the i2b2 score of 0.87 for partial match). For TIMEX3s, our F1-score was 0.8047, compared to an F1-score of 0.89 for i2b2.

TLINKing medical EVENTs remains a very difficult task. By using our narrative container approach to constrain the number of necessary annotations and by eliminating often-confusing inverse relations (like 'after' and 'during') (neither of which were done for the i2b2 data), we were able to significantly improve on the i2b2 TLINK span agreement F1-score of 0.39, achieving an agreement score of 0.5012 for all LINKs across our corpus. The majority of remaining annotator disagreement comes from different opinions about whether any two EVENTs require an explicit TLINK between them or an inferred one, rather than what type of TLINK it would be (e.g. BEFORE vs. CONTAINS). Although our results are still significantly higher than the results reported for i2b2, and in line with previously reported general news figures, we are not satisfied. Improving IAA is an important goal for future work, and with further training, specification, experience, and standardization, we hope to clarify contexts for explicit TLINKS.

News-trained temporal information extraction systems see a significant drop in performance when applied to the clinical texts of the THYME corpus. But as the corpus is an extension of ISO-TimeML, future work will be able to train ISO-TimeML compliant systems on the annotations of the THYME corpus to reduce or eliminate this performance gap.

Some applications that our work may enable include (1) better understanding of event semantics, such as whether a disease is chronic or acute and its usual natural history, (2) typical event duration for these events, (3) the interaction of general and domain-specific events and their importance in the final timeline, and, more generally, (4) the importance of rough temporality and narrative containers as a step towards finer-grained timelines.

We have several avenues of ongoing and future work. First, we are working to demonstrate the utility of the THYME corpus for training machine learning models. We have designed support vector machine models with constituency tree kernels that were able to reach an F1-score of 0.737 on an EVENT-TIMEX3 narrative container identification task (Miller et al., 2013), and we are working on training models to identify events, times and the remaining types of temporal relations. Second, as per our motivating use cases, we are working to integrate this annotation data with timeline visualization tools and to use these annotations in quality-of-care research. For example, we are using temporal reasoning built on this work to investigate the liver toxicity of methotrexate across a large corpus of EHRs (Lin et al., under review)]. Finally, we plan to explore the application of our notion of an event (anything that should be visible on a domain-appropriate timeline) to other domains. It should transfer naturally to clinical notes about other (non-cancer) conditions, and even to other types of clinical notes, as certain basic events should always be included in a patient's timeline. Applying our notion of event to more distant domains, such as legal opinions, would require first identifying a consensus within the domain about which events must appear on a timeline.

9 Conclusion

Much of the information in clinical notes critical to the construction of a detailed timeline is left implicit by the concise shorthand used by doctors. Many events are referred to only by a term such as "tu-

mor", while properties of the event itself, such as "intermittent", may not be specified. In addition, the ordering of events on a timeline is often left to the reader to infer, based on domain-specific knowledge. It is incumbent upon the annotation guideline to indicate that only informative event orderings should be annotated, while leaving domain-specific orderings to post-annotation inference. This document has detailed our approach to adapting the existing ISO-TimeML standard to this recovery of implicit information, and defining guidelines that support annotation within this complex domain. Our guidelines, as well as the annotated data, are available at `http://thyme.healthnlp.org`, and the full corpus has been proposed for use in a SemEval 2015 shared task.

Acknowledgments

The project described is supported by Grant Number R01LM010090 and U54LM008748 from the National Library Of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library Of Medicine or the National Institutes of Health.

We would also like to thank Dr. Piet C. de Groen and Dr. Brad Erickson at the Mayo Clinic, as well as Dr. William F. Styler III, for their contributions to the schema and to our understanding of the intricacies of clinical language.

References

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

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy*, 9(1):5–16.

Steven Bethard. 2013. Cleartk-timeml: A minimalist approach to tempeval 2013. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 10–14, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(Database issue):D267–D270, January.

Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Finding temporal order in discharge summaries. In *AMIA Annual Symposium Proceedings*, volume 2006, page 81. American Medical Informatics Association.

Carlo Combi, Yuval Shahar, et al. 1997. Temporal reasoning and temporal data maintenance in medicine: issues and challenges. *Computers in biology and medicine*, 27(5):353–368.

Robert H Dolin. 1995. Modeling the temporal complexities of symptoms. *Journal of the American Medical Informatics Association*, 2(5):323–331.

George Hripcsak, Nicholas D Soulakis, Li Li, Frances P Morrison, Albert M Lai, Carol Friedman, Neil S Calman, and Farzad Mostashari. 2009. Syndromic surveillance using ambulatory electronic health records. *Journal of the American Medical Informatics Association*, 16(3):354–361.

Ann K Irvine, Stephanie W Haas, and Tessa Sullivan. 2008. Tn-ties: A system for extracting temporal information from emergency department triage notes. In *AMIA Annual Symposium proceedings*, volume 2008, page 328. American Medical Informatics Association.

Elpida T Keravnou. 1997. Temporal abstraction of medical data: Deriving periodicity. In *Intelligent Data Analysis in Medicine and Pharmacology*, pages 61–79. Springer.

Klaus H. Krippendorff. 2012. *Content Analysis: An Introduction to Its Methodology*. SAGE Publications, Inc, third edition edition, April.

Chen Lin, Elizabeth Karlson, Dmitriy Dligach, Monica Ramirez, Timothy Miller, Huan Mo, Natalie Braggs, Andrew Cagan, Joshua Denny, and Guergana. Savova. under review. Automatic identification of methotrexade-induced liver toxicity in rheumatoid arthritis patients from the electronic medical records. *Journal of the Medical Informatics Association*.

John McCarthy. 2002. Actions and other events in situation calculus. In *Proceedings of the International conference on Principles of Knowledge Representation and Reasoning*, pages 615–628. Morgan Kaufmann Publishers; 1998.

Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, John F Hurdle, et al. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform*, 35:128–44.

Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova. 2013. Discovering temporal narrative containers in clinical text. In *Proceedings of the 2013 Workshop on Biomedical Natural Langua ge Processing*, pages 18–26, Sofia, Bulgaria, August. Association for Computational Linguistics.

Eleni Miltsakaki, Rashmi Prasad, Aravind Joshi, and Bonnie Webber. 2004. The penn discourse treebank. In *In Proceedings of LREC 2004*.

Massimo Poesio. 2004. Discourse annotation and semantic annotation in the gnome corpus. In *In Proceedings of the ACL Workshop on Discourse Annotation*.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Sauri. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation*, 39(2-3):123–164.

James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. Iso-timeml: An international standard for semantic annotation. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC 2010)*, Valletta, Malta.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, and Ian Roberts. 2009. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966.

Guergana Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. 2009. Towards temporal relation discovery from the clinical narrative. In *AMIA Annual Symposium Proceedings*, volume 2009, page 568. American Medical Informatics Association.

Tessa Sullivan, Ann Irvine, and Stephanie W Haas. 2008. It's all relative: usage of relative temporal expressions in triage notes. *Proceedings of the American Society for Information Science and Technology*, 45(1):1–8.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013a. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*.

Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013b. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association*, 20(5):806–813.

Alexander Turchin, Maria Shubina, Eugene Breydo, Merri L Pendergrass, and Jonathan S Einbinder. 2009. Comparison of information content of structured and narrative text data sources on the example of medication intensification. *Journal of the American Medical Informatics Association*, 16(3):362–370.

Naushad UzZaman, Hector Llorens, Leon Derczynski, James Allen, Marc Verhagen, and James Pustejovsky. 2013. Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 1–9, Atlanta, Georgia, USA, June. Association for Computational Linguistics.

Marc Verhagen. 2005. Temporal Closure in an Annotation Environment. *Language Resources and Evaluation*, 39(2):211–241.

Veronika Vincze, Gyrgy Szarvas, Richrd Farkas, Gyrgy Mra, and Jnos Csirik. 2008. The bioscope corpus: biomedical texts annotated for uncertainty, negation and their scopes. *BMC Bioinformatics*, 9(Suppl 11):1–9.

Ying Zhao, George Karypis, and Usama M. Fayyad. 2005. Hierarchical clustering algorithms for document datasets. *Data Mining and Knowledge Discovery*, 10:141–168.

Jiaping Zheng, Wendy W Chapman, Rebecca S Crowley, and Guergana K Savova. 2011. Coreference resolution: A review of general methodologies and applications in the clinical domain. *Journal of biomedical informatics*, 44(6):1113–1122.

Li Zhou, Simon Parsons, and George Hripcsak. 2008. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *Journal of the American Medical Informatics Association*, 15(1):99–106.