System Architecture for Temporal Information Extraction, Representation and Reasoning in Clinical Narrative Reports

Li Zhou, MS[1], Carol Friedman, PhD[1], Simon Parsons, PhD[2], George Hripcsak, MD, MS[1]
[1]Department of Biomedical Informatics, Columbia University, New York, NY
[2]Department of Computer and Information Science, Brooklyn College, Brooklyn, NY

Exploring temporal information in narrative Electronic Medical Records (EMRs) is essential and challenging. We propose an architecture for an integrated approach to process temporal information in clinical narrative reports. The goal is to initiate and build a foundation that supports applications which assist healthcare practice and research by including the ability to determine the time of clinical events (e.g., past vs. present). Key components include: (1) a temporal constraint structure for temporal expressions and the development of an associated tagger; (2) a Natural Language Processing (NLP) system for encoding and extracting medical events and associating them with formalized temporal data; (3) a post-processor, with a knowledge-based subsystem to help discover implicit information, that resolves temporal expressions and deals with issues such as granularity and vagueness; and (4) a reasoning mechanism which models clinical reports as Simple Temporal Problems (STPs).

1. Introduction

Time, which is used to elucidate the changes of the world and order the events in a description, is crucial in biomedical informatics and the Electronic Medical Record (EMR). For example, healthcare providers record the progress of a disease or a hospital course chronologically in text, and procedures and laboratory tests are stored in databases with time-stamps. The management of temporal information is essential in computer systems designed to assist with medical problem-solving and decision-making.

Temporal representation and reasoning theories draw from many fields, including philosophy, cognitive science, linguistics and computer science. Temporal logics and ontologies have been widely discussed and many systems have been proposed with different expressive power and computational complexity[1]. To exploit time-oriented clinical data, a variety of methods have been developed for medical information systems to address the associated storage, processing and retrieval requirements[2,3]. However, many approaches are either highly application dependent or concentrate on a specific issue of processing temporal data (e.g., temporal granularity). In recent years, researchers have been working on more comprehensive approaches[3].

Temporal information in narratives is rich, flexible and realistic. Though Combi and Shahar[2] divided the research efforts in designing and developing time-oriented medical systems into two main directions: temporal reasoning and temporal data maintenance, they mainly focused on clinical databases, where time is usually stored as timestamps, and temporal abstraction. Recently, Augusto[3] argued that in the medical domain "*Natural Language* turns into a very fertile area of research where temporal issues are very important".

Temporal representation and reasoning in Natural Language (NL) is a nontrivial task due to: (1) the diversity of time expressions; (2) the complexity of determining temporal relations among events; (3) the difficulty of handling temporal granularity; and (4) other major problems in computational NLP (e.g., ambiguity, anaphora, ellipsis, and conjunction). To date, minimal work has been done in medical informatics on temporal representation and reasoning problems, and the work described here is one of the few attempts to build a system for handling temporal information in clinical texts.

To understand how to automatically handle temporal information, it is first necessary to analyze how temporal information is conveyed in text, to examine which aspects of existing NLP systems need to be improved to process temporal data, and to investigate and evaluate suitable temporal ontologies and reasoning mechanisms. We have previously published results of our work in these areas[4].

Our objective is to build on this earlier work by developing a comprehensive treatment of temporal information in clinical narrative data, including extraction, representation, and reasoning. The goal is to initiate and build a foundation that supports further applications which assist healthcare practice and research (such as detection of medical errors). First, we present a high-level overview of the architecture. Second, we describe in detail each of the individual components. Lastly, we discuss the advantages and potential disadvantages of the described architecture, and the future work we see following from what we describe here.

2. System architecture

A high level overview of the proposed system architecture is shown below (Figure 1). The individual components include: (1) an annotation structure and tagger for temporal expressions; (2) an NLP system for encoding and extracting medical

events and formalized temporal data; (3) a post-processor including a knowledge-based subsystem; and (4) a reasoning mechanism which models clinical reports as Simple Temporal Problems (STPs).

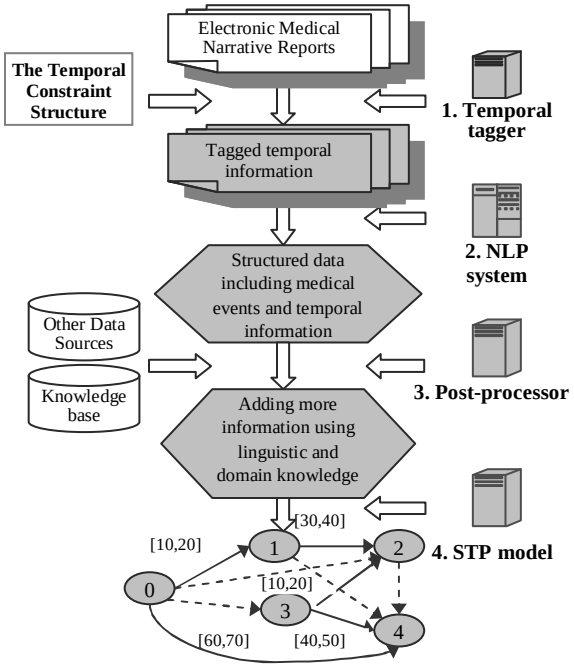

Figure 1: System Overview

The following example is used throughout this paper:

Example 1: *"Two years prior to admission the patient was diagnosed with hepatitis. The patient had an orthotopic liver transplant on June 7, 1992. He underwent a t-tube study, and now presents with a fever of 101°F lasting two days." (Note that admission date is September 18, 1992 obtained from an administrative data source)*

2.1. Annotation and tagging of temporal information

The first component in our system is the temporal tagger. This takes narrative records and represents the temporal information that they contain in a structured way. This facilitates later reasoning with the information. The challenge is to provide a model in which temporal expressions can be represented in an expressive, sound and unambiguous way.

We have developed a formal model framework, called the Temporal Constraint Structure (TCS)[5], based on the analysis of 200 discharge summaries from 1987 to 2004 drawn from the Columbia University Medical Center (CUMC) data repository. While record selection was random, we ensured that the samples of the text obtained represented the true complexity of medical language across the records and included a wide variety of expressions used to represent temporal information.

We model the time over which an event occurs as an interval. Each interval has start and finish time points, each of which may be constrained by temporal expressions. For example, from the statement "*the patient was a heavy smoker until 1984"*, we can infer that the patient stopped smoking in 1984. However, we do not know when he/she started smoking. We represent temporal expression by placing the endpoint(s) of an event in a relative time line and determining the relative or metric relationship between the endpoint and its anchor, e.g., the patient stopped smoking (the finish) in (equal) 1984 (the anchor). The TCS contains a set of fields that constrain the start and the finish of medical events. The definition of each field, possible values and examples are described in Table 1. We demonstrate two examples coded using the TCS in the following:

Example 2.1.1. "*two years prior to admission*" in "*two years prior to admission the patient was diagnosed with hepatitis*" constraining the event "*diagnosis of hepatitis*" is encoded as:

event_point = "*unspecified*"; anchor = "*admission*"; anchor_point = "*start*"; relation = "*equal*"; q*uantity* = "*2*"; time_unit = "*year*"*;* direction = "*minus*"; interval_operator = "*jump*"

Example 2.1.2. "*lasting more than two days*" in "*he presents with a fever of 101°F lasting more than two days.*" constraining the event "*fever of 101°F*" is encoded as:

event_point = "*start*"; anchor = "*event*" (*fever*); anchor_point = "*finish*"; relation = "*before*"; quantity = "*2*"; time_unit = "*day*"; direction = "*minus*"; interval_operator = "*jump*"

With this structure in hand, we developed a temporal expression tagging-program to recognize variant time expressions and normalize them into particular target forms determined by the TCS. The output of the tagger was integrated into an existing NLP system, Medical Language Extraction and Encoding System (MedLEE)[6], which we discuss in the next section. An example output for this step is:

```
<phr sem = "date" t =
"event_point~unspecified ^
anchor~admission^anchor_point~start^
relation_i~equal^ quantity~2 ^
time_unit~year ^ direction~minus ^
interval_op~jump" > two years prior to admission
</phr> the patient was diagnosed with hepatitis.
```

MedLEE recognizes the tagged XML expressions in the text for "sem" representing the semantic class of the phrase and "t" is its target form.

2.2. A natural language processing system

Once the temporal data are tagged, the next step is to extract, structure, and encode the clinical information in the tagged patient reports. To do this we used MedLEE, which has been used at NewYork-Presbyterian Hospital since 1995. MedLEE uses a frame-based representation. Primary information (e.g., problem, lab test, and medication) and their

Table 1: Fields of the Temporal Constraint Structure

Fields	Definition	Values & Examples
event_ point[1]	endpoint (s) of the event which is constrained by the temporal expression values.	*start*, *finish*, *both* or *unspecified*
anchor[1]	constraining time point (e.g., in "operation on *10/20/2003"*, *anchor is 10/20/2003*)	a calendar date, a time of day, a relative date or time, an event, or a time reference
anchor_ point	If anchor is an event, the endpoint of the event is specified	*start*, *finish*, *both* or *unspecified*
anchor_ modifier	indicates the stage of a period of time, or the course of an event (e.g., anchor "*1990's*" in "*late 1990's*" has a modifier of "*late*")	*early*, *mid* and *late*
relation[1]	a temporal relation between an endpoint of an event and its anchor or an interval constructed by the constraint structure with respect to the anchor	*equal*, *before*, *after*, *equal_or_before*, or *equal_or_after*
time_unit	unit for measuring time periods	*year*, *month*, *day*, *hour*, etc.
quantity	specified or indefinite number or amount for measuring the length of a time period	a number or a vague quantifier (e.g., *many*)
direction	indicates the direction of an interval relative to its anchor	*plus* (future), *minus* (past), or *both* (e.g., *within three weeks*).
interval_ operator	characterizes an endpoint of the event; determines whether an endpoint of an event occurred a specified duration away from the anchor (*jump*), or any time between the anchor and a specified duration away from the anchor (*drag*)	*jump* (e.g., *three weeks ago*) or *drag* (e.g., *within the past two month*)
vagueness	indicates if a vagueness modifier is contained within the expression (e.g., about in "*about two weeks ago*", approximately)	*Yes*

[1] these fields are required for the temporal constraint structure.

associated values are in the top level frames, and modifiers are nested. In this step, tagged temporal information, medical events and other information are processed to form a structured output. Temporal information and events are linked with each other directly or indirectly. The output of this step for the same example in the previous section is as follows (shown in simplified XML format):

```
<problem v="hepatitis" umls="C0019158"
      idref="p29">
    <date v="event_point~unspecified ^
        anchor~admission^anchor_point~start^
        relation_i~equal^ quantity~2 ^
        time_unit~year ^ direction~minus ^
        interval_op~jump" idref="p7">
    </date>
    <sectname v="report history of present
      illness item"> </sectname>
    <sid idref = "s1.1.1"> </sid>
</problem>
```

MedLEE output shows that not only medical events (e.g., a specific "problem") and temporal information are obtained, but also other contextual information, including the section ("sectname"), paragraph and sentence in which the term appeared ("sid"), and sentence position ("idref").

2.3. Post-processor

The XML MedLEE output feeds into the post-processor where the structured temporal information is used as the basis of temporal reasoning. The post-processor implements methods from our previous work[4] that deal with issues including temporal granularity, implicit and explicit vagueness, and uncertainty.

One additional task that is also implemented in the post-processor is to resolve time expressions, especially indexical expressions (e.g., "*now*" and "*today*"), which designate times that are dependent on the temporal context of the report. Data from other sources can help to solve this problem. For example, lab data can be easily retrieved from clinical databases in the future. Also for temporal coreference (e.g., "*On 2/10/05, the day before discharge*"), the more absolute temporal reference will be chosen.

The majority of temporal information extracted in the first two steps was explicit, but the kind of NL information that can be extracted by the post-processor is usually more implicit. Thus, a challenge in using the processed data from NLP systems is that the temporal concepts that one has to deal with may not always be linked. In order to link time to an event or one event to another, we have been developing a knowledge-based subsystem utilizing sources like those described next.

2.3.1. Linguistic knowledge

Temporal discourse analysis helps to order events and link them to time. One default rule we can use is "narrative time progression", which means that except when explicitly contradicted in the text, each successive statement either has the same "event-time" as the event in the preceding statement, or is temporally after it, but is never before it.

Nevertheless, studies[7] show that rules can be built based on discourse structure, lexical knowledge and syntax to override the default "narrative time progression".

2.3.2. Biomedical terminology

Biomedical terminology helps NLP systems not only to reduce redundancy and ambiguity, but also to improve temporal reasoning. The hierarchical structure of terminology can help identify whether two different statements can refer to the same event (e.g., *ampicillin* and *antibiotics*). Using its well-defined semantic relations (e.g., causal relationship), terminologies can help order events. There is also a semi-intervals problem, which means either the beginning or the ending of an interval is not directly known (e.g., *patient took medication until 10 days ago* or *hypertension beginning in 1999*). In most cases, an end point is never mentioned. Looking at other data sources or using a domain knowledge base might provide ways to handle them. For example, a chronic disease might be alleviated but still exists even after discharge, but an acute problem may be resolved before the discharge. MedLEE automatically maps its structured output to UMLS codes. Therefore, we can use the UMLS or other terminologies such as SNOMED as knowledge bases.

2.3.3. Domain specific knowledge

Clinical narrative reports may follow specific formats, which we can use to unearth implicit temporal references. In Example 1 (from the section "history of present illness"), the indexical time expression "*now*" means that the event, specifically "*fever*", happens as the same time as or before admission. We also identified several general reference events like admission, discharge, operation, or transfer that are usually located at an absolute time point and can be used as anchors. Some linguists have described the medical summary as "*a sequence of episodes that correspond to phrases, sentences, or groups of sentences dealing with a single topic*". In specific medical domains, a set of key events can be defined[8] and used to order and group events around those key events. A textbook or decision support tool can facilitate this work.

2.4. Modeling clinical narrative reports as an STP

After formalizing the notion of time, representing, and extracting the temporal information and medical events from real data, an appropriate computational mechanism is needed for automatically and efficiently reasoning about temporal relationships. Systems applying constraint propagation techniques usually use a graph-based representation where vertices represent times and arcs represent the possible temporal relationships. An STP is a subset of the Temporal Constraint Satisfaction Problems (TCSPs) proposed by Dechter et al.[9]. They can be solved by a polynomial time algorithm and are sufficient to represent primitive Allen relations, simple metric constraints and points anchored in absolute time.

In previous work[4], we modeled electronic discharge summaries as STPs. Each event was modeled as an interval with a start and finish. All assertions about events were encoded using temporal information in the report, and mapped to the model as constraints.

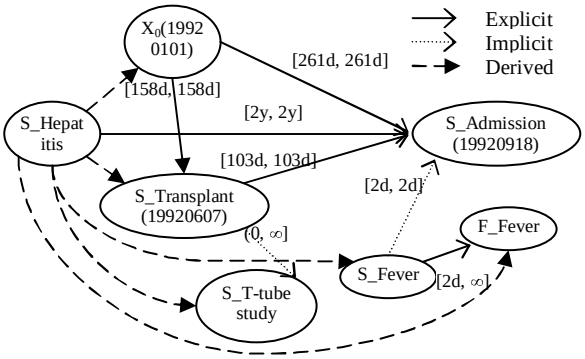

Figure 2: Constraint graph representing Example 1.

Given all encoded explicit information (e.g., "*the start of hepatitis (S_Hepatitis) is 2 years before the start of this admission (S_Admission)"* and "*the start of fever (S_Fever) is equal or more than 2 days before the end of the fever (F_Fever)"*) and implicit information obtained after applying the linguistic or domain knowledge (e.g., "*the start of T-tube study (S_T-tube study) was conducted after the start of liver transplant (S_Transplant)*"), a modified all-paired-shortest-paths algorithm[4] was used to derive more information and the domain was restricted further. Figure 2 shows the constraint graph representing Example 1 and the derived constraints for *S_Hepatitis* (note that X_0 is a predefined time origin and all times are relative to X_0). After the above steps, the node *S_hepatitis* may be connected to every other node in the network, and, as a result, we can figure out how long after hepatitis was diagnosed that a transplant was performed, and when it was that the fever occurred. The nodes may include a timestamp (e.g., "*19920918*" for *S_Admission* represents September 18, 1992) and the numbers in brackets represent the limits of duration between pairs of nodes (e.g., "*[103d, 103d]*" between *S_Transplant* and *S_Admission* conveys that the liver transplant was performed 103 days before this admission).

3. Discussion

In this paper, we have described the four components of our system. The temporal tagger, NLP system, and STP model have been implemented and tested. Currently, we are enhancing the post-processor.

The temporal constraint structure we describe preserves the original meaning of temporal information as completely and concisely as possible. Though the TimeML[10] group has developed a temporal annotation guideline, it mainly focuses on the news article domain. An important theoretical foundation for NLP in the medical domain is a sublanguage theory which shows that a language in a restricted domain is more well-defined than the general domain and can be characterized by a specified vocabulary, semantic relations and, in some cases, syntax. Time in clinical narratives has common characteristics with other domains, but also possesses unique aspects. For example, healthcare providers use jargon, like "*on postop day # 2"* which means "*2 days after operation*". Our structure addresses the specific characteristics of temporal information exhibited in the medical domain. Instead of using "document creation time" as a reference to normalize relative time, we identified a set of domain specific reference-events, such as admission or discharge times. Some of them can be easily located at an absolute time point. More significantly, they can be used to relate other medical events to one another.

The system we describe expands the functionality of an existing medical language processing system. Though there exists previous research focused on reasoning with the representations for temporal expressions[11], limited detail is given on how the representation can actually be obtained automatically by a tagging system. Depending on the grammar of the NLP system, time and event may be linked together. However, most implicit temporal relationships need to be inferred from domain knowledge since primary concepts are usually not linked. For example, the concepts "*hepatitis*" and "*t-tube study*" can be identified by an NLP system, but the NLP system in itself usually does not contain the knowledge to relate these concepts in a temporal relationship. In addition, ordering events in a complicated case or across multiple reports of the same patient is challenging. Our solution to this problem uses discourse analysis, biomedical terminologies and domain-specific knowledge. Currently, we have developed some rudimentary rules, and are working to enhance the set.

In the news articles that are widely studied in NLP research, events are defined as situations that happen or occur, generally expressed by means of tensed or untensed verbs, nominalizations. For example, John *presented* on Monday. Medical events can refer to any medical-related phenomena, and most of them are expressed by nouns in clinical reports. In addition, medical narratives are usually written in a semi-structured manner. It is our feeling that in historical medical reports, tense and aspect information is not as useful as in other domains (e.g., the phrase, "the patient *presented* with" does not necessarily mean that the event occurred in the past). However, further studies are needed, and form part of our future work.

In a previous paper[4] we have shown that an STP appears to be sufficient to represent most temporal assertions in discharge summaries, and that computationally tractable algorithms can be used to draw conclusions. Other important issues, including intermittence, periodicity, granularity, vagueness, ambiguity, uncertainty and plans, were also discussed in that paper.

4. Conclusion

In this paper, we have proposed a modular architecture for comprehensively processing the time-oriented information in clinical narrative reports. We have constructed implementations of the various modules and have linked them to form a prototype system. This system integrates NLP techniques, multiple knowledge-bases, and a temporal reasoning formalism. By providing a way to determine and discover temporal relationships among medical events, our prototype assists medical decision support. Further system enhancements and evaluation are required.

Acknowledgements

Supported by National Library of Medicine grants R01 LM06910; R01 LM07659, and R01 LM07268.

References

1. Pani AK, Bhattacharjee GP. Temporal representation and reasoning in artificial intelligence: A review. Mathematical and Computer Modelling. 2001; 34(1-2): 55-80.
2. Combi C, Shahar Y. Temporal reasoning and temporal data maintenance in medicine: Issues and challenges. Computers in Biology and Medicine. 1997; 27(5): 353-368.
3. Augusto J.C. Temporal Reasoning for Decision Support in Medicine. AI in Med. 2005; 33(1): 1-24.
4. Hripcsak G., Zhou L., Parson S., Das AK., Johnson SB. Modeling Electronic Discharge Summaries as a Simple Temporal Constraint Satisfaction Problem. JAMIA. 2005; 12(1):55-63.
5. Zhou L, Melton G, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. JBI 2005;Accepted.
6. Friedman C, Hripcsak G, Shagina L, and Liu H. Representing information in patient reports using natural language processing and the extensible markup language. JAMIA. 1999; 6: 76-87.
7. Jordan PW. Determining the Temporal Ordering of Events in Discourse. Masters Thesis for CMU Computational Linguistics Program, 1994.
8. Obermeier K, Temporal inference in medical texts. Proceedings of 23 Annual Meeting of the Association for Computational Linguistics. Chicago. 1985 July; 9-17.
9. Dechter R, Neiri I, Pearl J. Temporal constraint networks. Artificial Intelligence. 1991; 49:61–95.
10. Mani I, Pustejovsky J, Sundheim B, Introduction to the special issues on temporal information processing. ACM Trans. Asian Lang. Inf. Process. 2004; 3(1): 1-10.
11. Han B, Lavie A. A framework for resolution of time in natural language. ACM Trans. Asian Lang. Inf. Process. 2004; 3(1): 11-32.