

Towards Event Sequence Representation, Reasoning and Visualization for EHR Data

Cui Tao

Dept. of Health Science Research
Mayo Clinic
Rochester, MN

Krist Wongsuphasawat

Human-Computer Interaction Lab &
Dept. of Computer Science
University of Maryland
College Park, MD

Kim Clark

Dept. of Biomedical Engineering
University of Minnesota
Minneapolis, MN
Boston Scientific Corporation
Maple Grove, MN

Catherine Plaisant

Human-Computer Interaction Lab
University of Maryland
College Park, MD

Ben Shneiderman

Human-Computer Interaction Lab &
Dept. of Computer Science
University of Maryland
College Park, MD

Christopher G. Chute

Dept. of Health Science Research
Mayo Clinic
Rochester, MN

ABSTRACT

Efficient analysis of event sequences and the ability to answer time-related, clinically important questions can accelerate clinical research in several areas such as causality assessments, decision support systems, and retrospective studies. The Clinical Narrative Temporal Reasoning Ontology (CNTRO)-based system is designed for semantically representing, annotating, and inferring temporal relations and constraints for clinical events in Electronic Health Records (EHR) represented in both structured and unstructured ways. The LifeFlow system is designed to support an interactive exploration of event sequences using visualization techniques. The combination of the two systems will provide a comprehensive environment for users to visualize inferred temporal relationships from EHR data. This paper discusses our preliminary efforts on connecting the two systems and the benefits we envision from such an environment.

Categories and Subject Descriptors

H.4.m [Information Systems Applications]: Miscellaneous;
J.3 [Computing Applications]: Life and Medical
Sciences—Medical information systems

General Terms

Information Systems Applications

Keywords

EHR, Temporal Relation Reasoning, Time Trend Visualization, Semantic Web

1. INTRODUCTION

Efficient analysis of event sequences and the ability to answer time-related, clinically important questions can boost a series of clinical research in different areas such as causality assessments, decision support systems, as well as retrospective studies.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or to publish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

IHI'12, January 28–30, 2012, Miami, Florida, USA.

Copyright 2012 ACM 978-1-4503-0781-9/12/01...\$10.00.

Potential time patterns may exist within EHR data for patients with similar conditions, diagnoses, or similar treatments/procedures. These patterns may include a similar sequence of events, similar durations of or between events, or a similar time/date during which the related events occurred. These temporal properties and relationships, however, are often buried within the text of the narratives, requiring an astute observer to detect patterns while reading through EHR data. In addition, because temporal relations may require inference if they are not explicitly expressed within the narrative, temporal reasoning is also needed in order to analyze the trends in time. Manually assessing tens of thousands of clinical reports is time consuming, expensive, and the potential exists for a missed pattern observation or error in interpreting event sequencing.

The CNTRO time-relation reasoning framework [8-10] is designed as a system for semantically-representing and automatically inferring temporal relations and constraints within EHR data. The system is centered by an ontology, which provides a formal mechanism to represent temporal relations and constraints for clinical events. We have evaluated CNTRO using real-world clinical notes, and the evaluation results indicate that CNTRO can faithfully represent most of the time-related data without losing any information [8]. In addition, we have also developed a temporal relation reasoning framework prototype using semantic-web technologies [10]. This framework provides an application program interface (API) for users to ask different kind of time-related queries and tries to retrieve and infer the answers of these queries. We have evaluated the CNTRO framework using a medical device adverse event use case, and the results indicate that the CNTRO system is capable of answering the majority of the time-related questions which were generated during the evaluation (~90% accuracy) [2].

Although preliminary studies on the CNTRO system illustrate a promising approach for representing, annotating, and inferring temporal relations in clinical events, we believe that using a graphical user interface (GUI) and information visualization techniques for displaying the CNTRO outputted data to the end users (clinical researchers) could further increase the potential for clinical data analysis. To this end, we attempted to connect the output from CNTRO to LifeFlow [12], an interactive information visualization system for event sequences. The LifeFlow GUI provides an innovative visualization technique to help the user understand clinical event sequences in a short period of time, and

users can simultaneously ask questions and perform queries on the temporally organized data.

In this paper, we introduce our vision and preliminary efforts on connecting the CNTRO pipeline with the LifeFlow visualization tool. The paper starts with a brief introduction of the CNTRO system in Section 2, followed by an overview of the LifeFlow visualization environment in Section 3. Section 4 discusses our efforts on connecting the two systems and the comparison of the query functionalities. Finally, in Section 5 we provide a concluding summary.

2. CNTRO SYSTEM OVERVIEW

CNTRO is a semantic-web [1] based system that contains an OWL (Web Ontology Language) ontology for the time domain, an annotation component for extracting and semantic annotating information of interest in source documents with respect to domain ontologies, and a reasoning framework which can infer new knowledge from what is known.

CNTRO can be used to model temporal information and relations found both in structured databases and in native language-based clinical reports. CNTRO models clinical events, temporal expressions (such as time instants, time intervals, repeated time periods, and durations), granularity (such as minute, hour, day, month, year), temporal relationships, and time uncertainties, and has been evaluated using real world clinical events [8]. CNTRO has been integrated with existing ontologies [9] that cover time-related components.

Using CNTRO, temporal data in clinical narratives as well as data stored in databases can be annotated and represented in RDF [4]. A Protégé plug-in, called Semantator [6], is being implemented for annotating events and time-related information. This GUI is built upon the Knowtator [3], an annotation interface developed in Mayo Clinic for manual text annotation to XML files based on a pre-defined schema. The new interface allows CNTRO to function with the Semantic web with RDF triple stores [4] and SPARQL queries [7]. Additionally, the CNTRO system is being connected with the Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES) [5] to automate the annotation process.

The annotated RDF file is run through our CNTRO temporal relation-reasoning framework to infer temporal information that is not explicitly expressed in the original documentation. This framework combines DL-based reasoning, SWRL-based reasoning, and SWRL Built-Ins library. It provides basic functions for answering a list of time-related questions. The feasibility of CNTRO has been evaluated with real-world clinical data and medical device adverse event narratives. The evaluation results indicate that the system is capable of answering the majority of the time-related questions that were generated during the evaluation (~90% accuracy) [2].

3. LIFEFLOW SYSTEM OVERVIEW

The core of the LifeFlow system consists of two parts: a visual representation that can aggregate and compress any number of event sequences into one picture, and an interactive features that supports data exploration.

3.1 VISUAL REPRESENTATION

Figure 1 illustrates the conversion from four records of clinical event sequences to a LifeFlow visual representation. In the

beginning, raw data is displayed on a horizontal timeline with colored triangles representing the events (an approach used in LifeLines2 [11], a former project that inspired LifeFlow). Each row represents one patient. The transformation into LifeFlow takes two steps: aggregation and visualization.

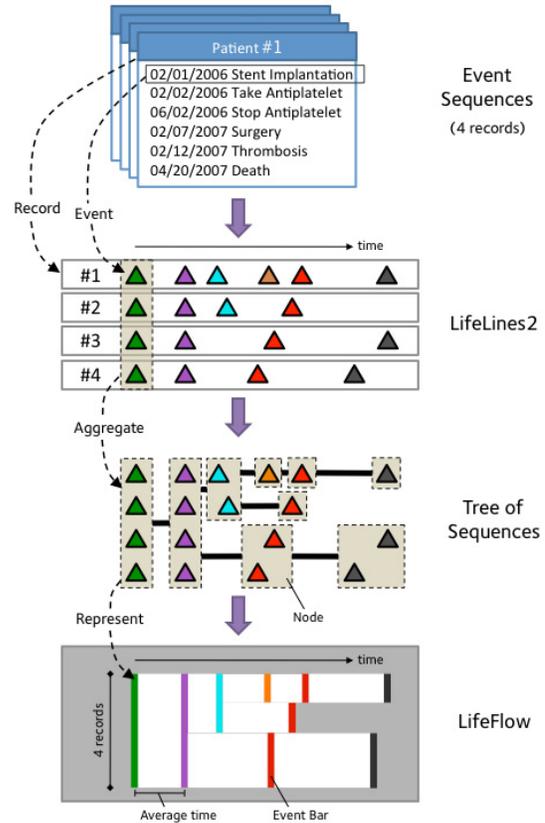


Figure 1: This diagram explains how to convert four records of event sequences into a LifeFlow visual representation.

1) *Aggregation* –All records are aggregated into a tree-based data structure called a *tree of sequences* based on the prefixes of their event sequences. For example, a record that contains event sequence “Stent implantation” → “Take antiplatelet” → “Stop taking antiplatelet” → “Thrombosis” and another record that contains event sequence “Stent implantation” → “Take antiplatelet” → “Thrombosis”, share the same prefix sequence “Stent implantation” → “Take antiplatelet”.

The records are grouped event-by-event from the beginning of the event sequences to the end. In Figure 1, all records start with the “Stent implantation” (green) event so they are grouped together (indicated by dashed rectangle) into a green tree node. Then, they all also have the “Take antiplatelet” (purple) event, so they are still grouped together into a purple node. In the next step, two of them have light blue events while the other two have red events so they are split into light blue and red nodes. The same steps follow for the rest of the event sequences.

2) *Visualization* –Once the tree of sequences is created, it can be plotted as a LifeFlow. Each node of the tree is represented with a color-coded *event bar*, matching the color of the event type. The height of a bar is determined by the number of records in that node proportionally to the total number of records. E.g., the light blue node contains two out of four records so the height of its corresponding event bar is 50% of the total height. The horizontal gap between bars

(e.g. the gap between the green bar and the purple bar) is proportional to the mean time between the two events (green → purple). By default, the representative time gap is the mean, but users can change to the median or other metrics. The LifeFlow display is scalable because it does not require additional space for a larger number of records. Users can display hundreds of records or millions of records using the same amount of screen space.

3.2 INTERACTIVE FEATURES

In addition to the visual representation, many user interactions were designed to support data exploration with LifeFlow. These interactions give users flexibility to manipulate the view to ask different questions or provide additional information when needed. Some of the main interactions are listed as follows:

11) *Zoom*: Zooming horizontally changes time granularity. Zooming vertically allows more rare sequences to be viewed in more detail.

12) *Tooltip*: Provides additional information about the sequence.

13) *Overlay a distribution of gap between events*: The horizontal gap between event bars shows the representative time (mean, by default) between events. Users can place a cursor over the gap to see a distribution of the actual time gaps. Users can also query by drawing a selection on the distribution.

14) *Drill down*: Clicking on any event bar will select all records that are included in that bar. Selected records are highlighted in the LifeLines2 view. In a symmetrical fashion, selecting a record in the LifeLines2 view highlights the pattern contained in that record in the LifeFlow view, allowing the users to find other records that contain the same sequence. (Figure 2 #1)

15) *Align*: Users can choose any event type to be the *alignment* point. By choosing “stent thrombosis” as the alignment point, users can answer the question: “what happened to the patients before and after stent thrombosis?” (Figure 3 (right).)

16) *Grouping by non-temporal attributes*: Records commonly contain non-temporal attributes, e.g., patient's gender. LifeFlow allows users to select a non-temporal attribute and groups records by that attribute before the sequences are aggregated.

17) *Include/Exclude event types*: Using the legend on the bottom left of the screen, users can check/uncheck event types to include or exclude them from the view. This simple functionality allows powerful transformations of the display to answer questions.

4. CONNECTING THE TWO SYSTEMS

To investigate the feasibility of adapting LifeFlow as the GUI of CNTRO output data, we used LifeFlow to access and query information inferred by the CNTRO system with an adverse event use case. Figure 2 and 3 show screenshots of the LifeFlow user interface with our use case data in display.

In order to connect the two systems, we have made the following assumptions: (1) a set of events of interest needs to be identified for each use case, and (2) the event names have been normalized in each data file, for LifeFlow to show the time trends across multiple files. In addition, since LifeFlow is designed for working with time instants only, we used the start time for time intervals. The LifeFlow environment also expects that each event has an explicit timestamp. This assumption cannot be always satisfied, especially for events extracted from narratives, in which events only associate with relative temporal relationships (i.e., event1 happened before event2) some times. In this situation, we had to assign timestamps for the events while still preserving the temporal relations. We also put a note on

the event indicating that the timestamp is assigned. For the above example, we have to assign timestamps for the two events to reflect the *before* relation. Note that the duration between the two events does not reflect the actual duration since it is unknown. With a large number of corpuses, the users can choose to exclude the data without explicit timestamps. Depending on the particular study, the assigned timestamps can still be helpful for identifying the time trends and reflecting the sequence of the events.

We have compared the query functionalities of the CNTRO API and the LifeFlow GUI. The CNTRO API includes 7 basic functions:

F1) `findEvent(searchText)`

- CNTRO API returns a list of events that match the search criteria.
- Since the event names are all normalized in the LifeFlow environment, to retrieve a particular kind of events is straightforward. LifeFlow has a legend panel that lists all event types. (See the bottom left corner in Figures 2 and 3. Users can check/uncheck the checkboxes to view the detailed information of selected event type (17).

F2) `getEventFeature(event, featureflag)`

- CNTRO API returns a specific time feature for a given event. The parameter `featureflag` indicates which time feature the user wants to retrieve: start time, end time, note taken time, or event time.
- In the LifeFlow environment, all events are plotted on a timeline according to their start time. In addition, users can view the timestamp of any given event by putting the cursor on top of that event in LifeLines2 view to bring up a tooltip (12). Note that for the event with time intervals, only the start time will be shown as the timestamp, any other information can be displayed as a note associated to the event.

F3) `getDurationBetweenEvents(event1, event2)`

- CNTRO returns the time interval between two events.
- LifeFlow currently does not support retrieving duration between two events in one record automatically. The user may have to retrieve the time for each event and calculate the difference. However, LifeFlow can show a summary of the duration between two event types from all records by unchecking all the checkboxes of other event types, leaving only two event types.

F4) `getDuration(event)`

- CNTRO API returns the duration of a given event. The duration can be either retrieved directly or calculated.
- This does not apply to the LifeFlow environment since the assumption is each event occurred at a point in time, not an interval.

F5) `getTemporalRelationType(event1, event2)`

- CNTRO API returns the temporal relations between two events if it can be retrieved differently or inferred.
- This can be checked intuitively in the LifeFlow environment since all the events are displayed on the timeline.

F6) `getTemporalRelationType(event1, time)`

- CNTRO API returns the temporal relations between an event and a specific time if it can be inferred or retrieved.
- This can be checked intuitively in the LifeFlow environment since all the events are displayed on the timeline.

F7) `sortEventsByTemporalRelationsOrTimeline(events)`

- CNTRO API returns the order (timeline) of a set of events.
- This can be checked intuitively in the LifeFlow environment since all the events are displayed on the timeline.

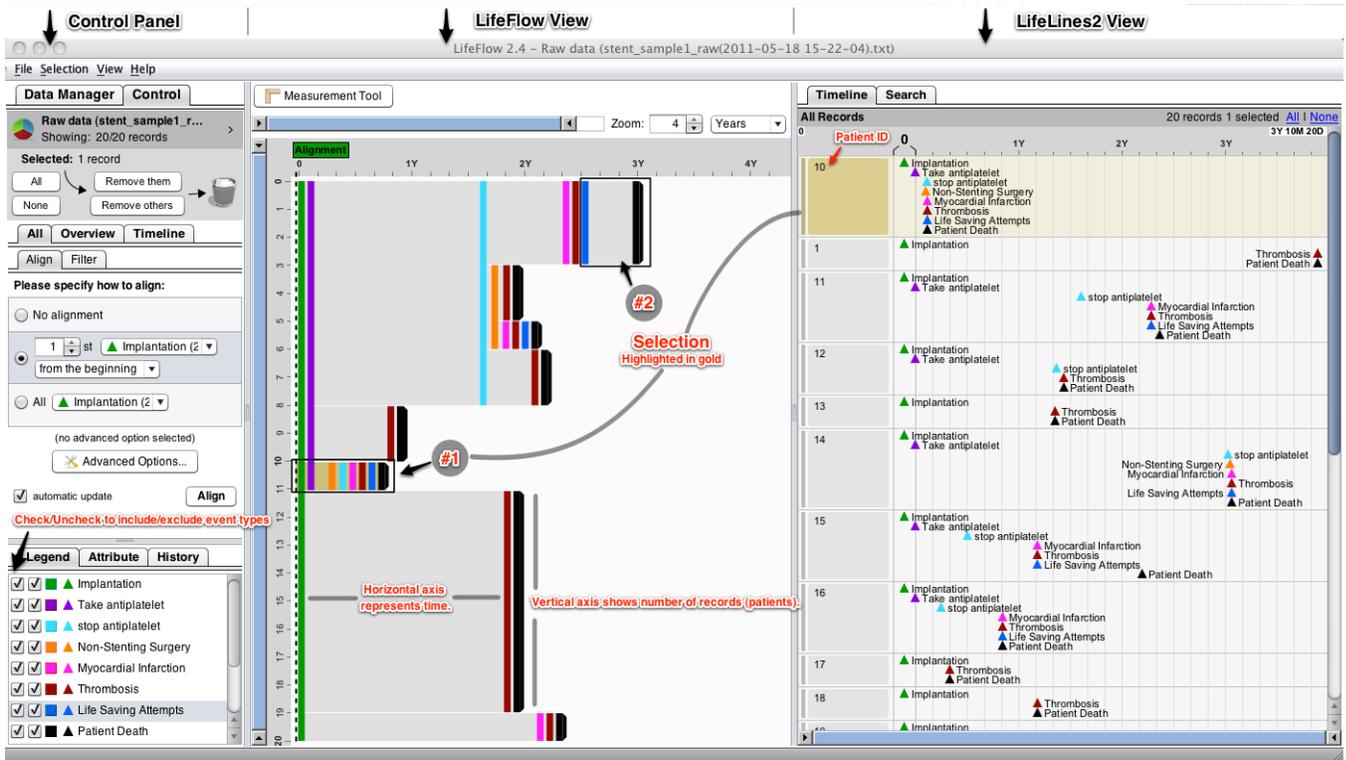


Figure 2: LifeFlow user interface can be divided into three parts: a control panel that allows the users to manipulate the display, a LifeFlow view that shows a summary of all event sequences, and a LifeLines2 view that shows details of each event sequence individually. This screenshot shows CNTRO output data of twenty patients with stent thrombosis.

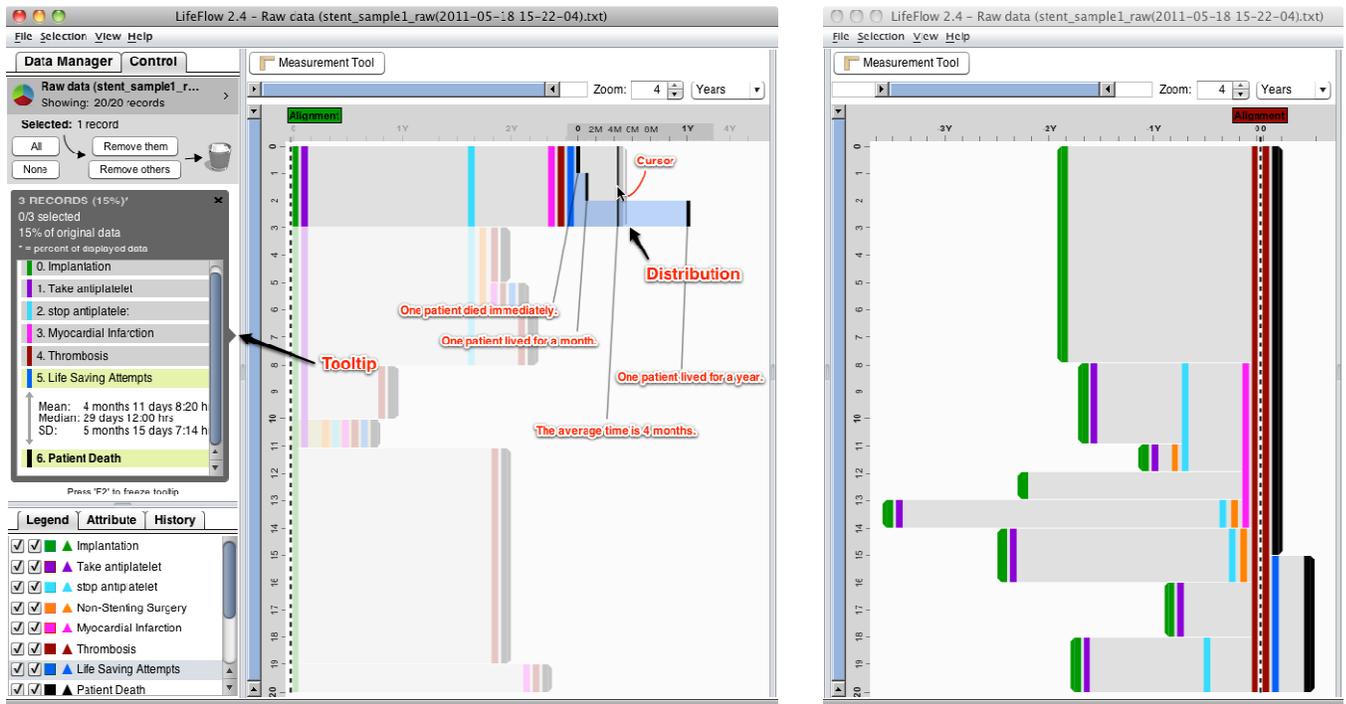


Figure 3: (Left) LifeFlow overlays distribution and shows a tooltip after the users place the mouse over the gap between life saving attempts (blue) and patient death (black). It indicates that one patient died immediately, another died after a month and another died after a year. / (Right) LifeFlow after aligned by "thrombosis" (red): The dashed line indicates the alignment point. The bars on the left are what happened before "thrombosis" while the bars on the right are what happened after.

The CNTRO API is designed for inferring the temporal relations as well as sorting events on the timeline for a particular patient or record. The LifeFlow environment, on the other hand, is an interactive visualization tool that supports an exploration of event sequences. Therefore, in addition to using the LifeFlow environment to retrieve information listed above, we can also leverage its visual power to gain more insight from the data.

To demonstrate this, we have described a sample use case scenario when clinical researchers would like to track patients after stent implantation. Users can use LifeFlow to quickly see the overview of all patients. Figure 2 shows the patient histories aligned (I5) by the implantation (green). Reading from left to right: After implantation, more than 50% of the patients were given antiplatelet medication, indicated by the purple bar that followed the green bar. Most of these patients stop taking antiplatelet about one and a half years after started, indicated by the light blue bar that followed the purple bar and horizontal gap between them. Another group of patients did not stop taking antiplatelet but had thrombosis directly (red after purple). The horizontal gap shows that the thrombosis occurred about one year after implantation. After thrombosis, all of them died (black after red). Another large group of patients did not take antiplatelet and had thrombosis (or the complaint file did not mention that antiplatelet therapy had been given) (red after green) approximately two years following stent implantation (horizontal gap). Similarly, all of them died after thrombosis (black after red).

This view can also identify the rare sequences, which could be anomalies or interesting spots in the dataset. For example, users noticed one small sequence (Figure 2 #1) that had non-stenting surgery (orange) immediately after taking antiplatelet (purple), so they clicked on it (I4) to query for the patients with this pattern. Both the pattern and selected patient were highlighted in gold. They found that there was only one patient for this sequence and used the patient ID to lookup more details about this patient.

In Figure 2 #2, users spotted that a few patients lived for a few months after life saving attempts (blue), so they placed the mouse over the gap to see more details if it is common that life saving attempts are successful and make the patients live for a few more months. Figure 3 (left) shows a screenshot after moving the mouse over the gap. The tooltip on the left (I2) lists all events from the beginning to the place where the mouse was placed and statistical information between life saving attempts and patient death. The overlay distribution (I3) showed that only one of the patients lived for a year after life saving attempt, which made the average duration between events become unexpectedly large.

5. CONCLUDING REMARKS

In summary, this preliminary study on connecting the two systems indicates that it is feasible to use LifeFlow on interactively visualizing the data represented in and inferred by the CNTRO system. It also shows the potential of enhancing data analysis to gain better understanding and more insight from the CNTRO outputted data.

Our future direction includes (1) detecting intuitive ways to display intervals and repeated events on the LifeFlow system; (2) investigating how to handle uncertainties; and (3) connecting the system on the API level so that the data inferred from CNTRO can be automatically viewed in LifeFlow.

6. ACKNOWLEDGMENTS

This research is partially supported by the National Science Foundation under Grant #0937060 to the Computing Research Association for the CIFellows Project, and the ONC Strategic Health IT Advanced Research (SHARP) award under Grant #90TR0002-01.

7. REFERENCE

- [1] Berners-Lee, T., *et al.* 2001. The Semantic Web. *Scientific American*. (May 2011).
- [2] Clark, K., *et al.* 2011. Application of a Temporal Reasoning Framework Tool in Analysis of Medical Device Adverse Events. In *Proc. AMIA Annual Symp.*
- [3] Ogren, P. V. 2006. Knowtator: a protege plug-in for annotated corpus construction. In *Proc. Conf. of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*. 273-275.
- [4] Resource Description Framework (RDF). <http://www.3.org/RDF>
- [5] Savova, G. K., *et al.* 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc.* 17, 507-513.
- [6] Semantator. <http://informatics.mayo.edu/CNTRO/index.php/Semantator>
- [7] SPARQL Query Language for RDF. <http://www.w3.org/TR/rdf-sparql-query/>
- [8] Tao, C., *et al.* 2010. CNTRO: A Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives. In *Proc. AMIA Annual Symp.* 787-791.
- [9] Tao, C., *et al.* 2011. CNTRO 2.0: A Harmonized Semantic Web Ontology for Temporal Relation Inferencing in Clinical Narratives. In *Proc. AMIA Summits on Translational Science*. 64-68.
- [10] Tao, C., *et al.* 2010. Time-oriented question answering from clinical narratives using semantic-web techniques. In *Proc. 9th International Semantic Web Conf.* 241-256.
- [11] Wang, T. D., *et al.* 2008. Aligning temporal data by sentinel events: discovering patterns in electronic health records. In *Proc. Conf. on Human Factors in Computing Systems (CHI)*. 457-466.
- [12] Wongsuphasawat, K., *et al.* 2011. LifeFlow: Visualizing an Overview of Event Sequences. In *Proc. Conf. on Human Factors in Computing Systems (CHI)*. 1747-1756.