Towards Temporal Relation Discovery from the Clinical Narrative

Guergana Savova[1], PhD, Steven Bethard[2], PhD, Will Styler[2], James Martin[2], PhD, Martha Palmer[2], PhD, James Masanz[1], MS and Wayne Ward[2], PhD

[1]Mayo Clinic, Rochester, MN; [2]University of Colorado, Denver, CO

Abstract

Disease progression and understanding relies on temporal concepts. Discovery of automated temporal relations and timelines from the clinical narrative allows for mining large data sets of clinical text to uncover patterns at the disease and patient level. Our overall goal is the complex task of building a system for automated temporal relation discovery. As a first step, we evaluate enabling methods from the general natural language processing domain - deep parsing and semantic role labeling in predicate-argument structures - to explore their portability to the clinical domain. As a second step, we develop an annotation schema for temporal relations based on TimeML. In this paper we report results and findings from these first steps. Our next efforts will scale up the data collection to develop domain-specific modules for the enabling technologies within Mayo's open-source clinical Text Analysis and Knowledge Extraction System.

Background: Temporal Relation Discovery in the General Domain

Recent developments in natural language processing (NLP) have enabled a variety of fine-grained semantic components to be extracted automatically from text. Machine learning systems have shown good performance on a variety of tasks, including the detection of people, organizations and locations as well as the semantic roles these entities play. But other important semantic structures have not yet been addressed. Events are often tied together through a variety of temporal structures. The temporal relations are expressed both explicitly, through words like *after*, and implicitly through inference. Extracting these sorts of temporal structures is crucial for an understanding of the text. Machine reasoning requires an explicit representation of the temporal structure. Such an explicit representation can be formed by identifying specific words or phrases as the event anchors of the structure, and then drawing explicit temporal relation links between the various events. By breaking the text structure down to its event anchors and the temporal relations between them, this representation provides smaller atomic units which are more easily accessed by machines and which lend themselves better to computational reasoning.

In the general NLP community, researchers have developed a markup language, TimeML, and an annotation schema for expressing the structure of events and temporal relations in text [1] [2]. The main tags within the TimeML framework are Event, TLink, SLink and ALink. By their definition, Events are situations that *happen* or *occur*. A TLink or Temporal Link represents the temporal relation that holds between events, times or between an event and a time with the values being **simultaneous, before, after, immediately before, immediately after, including, being included, during, beginning, begun by, ending, identity, set/subset**. SLink or Subordination Link is used for contexts to introduce relations between events of type **modal, factive, counter-factive, evidential, negative evidential, conditional**. ALink or Aspectual link represents the relationship between an aspectual event and its argument event with the following types: **initiation, culmination, termination, continuation, reinitiation**.

TimeML has been used to encode parts of the temporal structure of the 186 document in the TimeBank corpus to serve as a testbed[1]. Recent work on temporal relations has mostly revolved around this corpus which contains a small set of newswire documents annotated for events, times and the temporal relations between them. Researchers trained models for extracting events and temporal structure typically combining machine learning techniques with low level features like word stems and parts of speech, and found that important events could be identified with precision and recall in the 70s and 80s [3] [4] [5]. The creators of the TimeBank organized the 2007 TempEval competition[6] where a stricter annotation interface and a simplified set of temporal relations was used. Systems performed well on its tense identification task, but poorly on the other tasks which often required multiple stages of implicit temporal logic [7] [8]. Building on the lessons of TimeBank and TempEval, we [9] annotated some verb-clause constructions in the TimeBank, and showed that with a small amount of data, support vector machine models could be trained to find these temporal relations with accuracies of nearly 90%.

PropBank [10] is a project that adds semantic role labels in predicate-argument structures to the annotation of syntactically parsed structures critical for the task of language understanding as they provide components necessary for deciphering the meaning of the sentence. High accuracy predicate-argument detection and semantic role labeling is a prerequisite for high-accuracy temporal resolution discovery. VerbNet [11] is the largest on-line verb lexicon for English and organizes verbs into classes to achieve syntactic and semantic coherence among members of a class. PropBank builds on the Penn TreeBank syntactic annotations and is mapped to the more informative VerbNet labels.

Background: Temporal Relation Discovery from the Clinical Narrative

As Zhou and colleagues [12] point out "To date, minimal work has been done in medical informatics on temporal representation and reasoning problems". An extensive and detailed review of temporal reasoning with medical data is in [13]. Most of the efforts in temporal reasoning are in processing structured data [14] [15]. As Zhou and Hripcsak [13] observe "Representing and reasoning about temporal information in text has drawn increasing attention in the area of computational linguistics. The text corpus used by research groups in this area was usually outside of medicine …, so the adaptability of these systems for processing temporal information in medical data has not been assessed. Nevertheless, the progress that has been made in this field is very valuable and worth delivering to MLP (medical language processing) researchers." (p. 188).

University of Utah's system SPRUS [17] and its most recent version MPLUS [18] do not discover explicitly temporal information and events beyond some change of status attributes, e.g. improved, recurrence, etc. MedSyndikate [19] is a system that processes medical findings reports and discovers simple facts, complex propositions and evaluative assertions. For the conceptual representation of medical processes and events, modal and auxiliary verbs are analyzed, and the final conceptual representation uses an anaphora resolution component. The Medical Language Extraction and Encoding System, or MedLEE, [16] discovers clinical named entities and encodes them in a controlled terminology. Zhou and colleagues [12] propose an architecture for an integrated approach to processing temporal information in clinical narrative reports. They developed a temporal reasoning system with MedLEE as the NLP component followed by a reasoning mechanism [20]. A description and an evaluation of their system, TimeText, is provided in [25].

Research Questions, Data Set and Methods

Our long-term goal is to build a system for temporal relation discovery from the clinical narrative and for the creation of timelines of clinically-relevant events. We build on the work in the general NLP domain. In the process, Mayo's open-source clinical Text Analysis and Knowledge Extraction System (cTAKES) (www.ohnlp.org) [21] will be extended with a temporal relation discovery component and a reasoner to create timelines of clinically relevant concepts. cTAKES is an NLP tool for discovering clinically relevant concepts along with a set of attributes. Its main components are a sentence boundary detector, tokenizer, part-of-speech tagger, shallow parser, named entity recognizer and context discovery module. All components are trained for the clinical domain.

Towards our long-term goal, in the current study we aim at 1) assessing the accuracy of enabling technologies such as off-the-shelf deep parsers and semantic role labelers to determine if additional Treebanking and PropBanking would be required; and 2) creating an annotation schema for temporal relations in the clinical domain based on TimeML[2] and adding a temporal annotation module to Knowtator (http://knowtator.sourceforge.net/) to perform temporal annotations. To achieve these goals, we created a data set consisting of a 5K token clinical corpus (clinical and radiology notes) from the Mayo Clinic repository.

Treebanking and PropBanking Evaluation

The clinical notes are particularly problematic, since over a third of the sentences are sentence fragments which are not handled correctly by off-the-shelf parsers. Of note, cTAKES implements domain-specific part-of-speech tagging and phrasal chunking, but not deep parsing. In our current study, there were only 11% correct deep parses, although 42% had correct semantic role labels, surprisingly. The majority of the attachment problems appear to be associated with out-of-vocabulary (OOV) items that did not occur in the training data, or occurred there in different senses. For example, *rash* is typically an adjective in newswire but is a noun in clinical notes; *erythema* is not identified as a noun; quantifiers and negations are more likely to scope entire conjoined noun phrases as in *no edema, cyanosis and pallor*. A sufficient amount of domain specific training should alleviate most of these problems, especially if accurate named entity (NE) tags can be used as a back-off for OOV nouns. The field is only too familiar with the degradation in performance that occurs when parsers trained on the Wall Street Journal (WSJ) are tested on different corpora. This was showcased in the 2005 CoNLL shared task for PropBanking which included an evaluation on the Brown Corpus, to "test the robustness of the presented systems." [22]. The Charniak POStagger degrades by 5%, and the Charniak parser F score by 8%, from 88.25 to 80.84. There are two issues to take into consideration when trying to improve parser performance on a new corpus: lexical variance and syntactic construction variance. If the syntactic construction variance is minor the parser performance can be increased significantly by addressing solely the lexical issues, as illustrated in the gain achieved by Lease & Charniak on the biomedical domain [23]. Where the syntactic construction variance is major, additional treebanking is needed, although not necessarily in large quantities. Daelemans describes this as "the 'simple but effective' way of solving it [the portability problem]: retraining of modules using annotated data from the new domain." [24]. The preponderance of parsing errors due to sentence fragments is an example of syntactic variance which would benefit immensely from an additional 100K word Treebank of clinical notes which the Mayo Clinic NLP team has already annotated for sentence boundaries, tokens, part-of-speech tags and phrasal chunks following the Penn Treebank guidelines.

With respect to PropBanking, a major source of error for this data is mis-identification of predicating expressions, due primarily to inaccurate POS-tagging and parsing of fragments, abbreviations such as *b.i.d. , p.o., 100-mg, 15 cc*, and complex noun phrases. The Current Medications section is especially problematic. Even when the semantic roles have been correctly identified the ArgM labels are often wrong, once again a direct result of OOV items. Since the radiology notes have more accurate parses, the PropBanking is correspondingly more accurate as well, although still below WSJ levels.

Temporal Relation Annotation Schema

Our dataset was hand annotated to develop an annotation schema based on TimeML. The schema was implemented in Knowtator. The following types of temporal information are annotated:

- **TIMEX3** TIMEX3 objects are definitive references to time that provide concrete temporal references, e.g., *today, 24 hours ago, early March, April 15.* Examples: *[His anterior chest rash has not reoccurred since the PCN VK was discontinued **24-hours ago**. The last cyclosporine level was 373 in **January**.]*

- **EVENTS** Following the TimeML definition, EVENTS are situations that happen or occur. *[The **rash** has not **reappeared** and we will **monitor** closely. There are again noted postoperative **changes** consistent with prior right frontoparietal **craniotomy** for resection of a right frontal brain **tumor**.]* The events of *rash* and *reappeared* are two separate events; the original rash event, and a future rash which never occurred. Of note, *postoperative* and *prior* are not EVENTs; rather they suggest TLINKs between the events of changes and craniotomy. *Resection* is considered the cause of the *craniotomy* EVENT. In our schema, states and conditions are labeled as EVENTs in addition to traditional events. They can be signaled by adjectives and predications in addition to verbs, (the adjective is marked instead of the copula (*the eye is **watery*** rather than ***is** watery*). Also, modifiers (such as adjectives or locations) are not marked as part of the EVENT, only the head noun is.

- **TENSE** is an attribute of EVENT and refers not to the morphological tense on the verb, but rather, to the temporal relation of the event to the time of the patient-physician encounter. Currently, our schema includes three basic tense labels: 1) **PAST** is used where the event occurred before the encounter. *[He is taking in adequate nutrition and adequate fluids; **consumed** 3500 calories and **drank** 2-3 liters of fluid];* 2) **PRESENT** is used for events which refer to the time of the examination, MRI, or pathology report: *[Moderate sized retention **cyst or polyp** in the right maxillary antrum again **noted]**;* 3) **FUTURE** is used where the event is scheduled or planned to occur following the time of the encounter: *[Levaquin 750 mg p.o. q. day (will **restart** today)].*

- **CLASS** is an attribute of EVENT. Five different types of event class are specified. OCCURRENCE is used for events which happened. These are usually verbs or symptoms, such as *rash, visit, MRI* or even *tumor.[Otherwise, he has not had any nausea, **vomiting**, **diarrhea**, **chest pain**, **shortness-of-breath**, or swelling**.]**.* In the case of a condition or state, the class STATE is used. Symptoms like *nausea* or *swelling* are STATES, as are most other sorts of descriptors and chronic conditions. *[Additional **nodularity** along the right lateral aspect of the resection bed is also unchanged.].* The vast majority of clinical events like medication events, signs/symptoms, diseases/disorders and procedures are annotated as either OCCURRENCE or STATE. Two other less common event classes are our two primary epistemic markers, PERCEPTION and REPORTING. The following is a PERCEPTION event, an event where the observer's perception is especially salient: [*The patient's confusion was **noted** during the exam.]*. This contrasts with a REPORTing event, which involves a perception through a third party: *[The patient has **noticed** some rectal itching and mild pain this morning.]* The fifth and final event class is ASPECTUAL, which is used to indicate an event whose function is to emphasize or code the aspect of a later event, like *continues* or *restart. [The rash has not **reappeared** and we will monitor closely]*

- **DEGREE** is an attribute of EVENT. In order to express both the polarity of an event and the degree of a discussed change, the DEGREE of an event is also specified. The most commonly used degree is ALL. This is, in effect, the marker of both complete degree and positive polarity. This is used for an event that did, in fact, occur, and/or occurred fully. Most events annotated are of this degree: *[The patient has **hepatosplenomegaly**. PO **changes** right pterional craniotomy.]*. The opposite is NONE, which is used to indicate when the event did not take place, or has an otherwise negative polarity: *[No evidence for new **suprasellar mass**. This is un**changed** and may be related to treatment changes.]*. The final two degrees are MOST and LITTLE. These are used when there has been a little change of an event, or a large (but not complete) change, e.g. the events *signal* and *disappeared* in *[There is a small amount of bright T1 **signal**. Abdominal tenderness has nearly **disappeared**.]*

- **MODALITY** is an attribute of EVENT. Our current schema only has a single modality, which is HYPOTHETICAL. This is useful when annotating diagnoses, theories, or other medically relevant but hypothetical events like the events *related, tumor* and *stroke* in *[This is unchanged and may be **related** to treatment changes**.** An approximately 3cm nodular region of intermediate T2 signal involving the body of the corpus callosum is suspicious for residual or recurrent **tumor** but appears unchanged from the patients prior examination. The patient may have undergone a mild **stroke]***.

- **TLINKs** TLINK (or temporal link) annotations define the relation between two events, or between an EVENT and a TIMEX3. Although these are annotated in a more interactive fashion using Knowtator, they have the basic format: *EVENT1 [temporal relation] EVENT2/TIMEX3.* These links are generally designed to be short span, and in order to simplify annotation, annotators are discouraged from making TLINK annotations which span more than one or two sentences. TLINK annotations should only be made where the temporal relation between two events is definitively known, and where they provide more information than PAST, PRESENT and FUTURE. There are three different temporal relations possible in our schema, BEFORE, AFTER and OVERLAP. BEFORE and AFTER simply order two events in time. When possible, BEFORE should be used for consistency, but in some cases when TIMEX3 annotations are used, AFTER is necessary. OVERLAP is a single temporal relation that encompasses all the different notions of two things happening at the same time. This can refer to two simultaneous events, an EVENT that occurs during another, larger EVENT or time reference, or any other sense in which two events are occurring in the same timeframe (Table 1).

- **ALINK Annotation** Aspectual links (ALINKs) are created between two events, one of which gives aspectual information about the second. Any event previously marked with the class ASPECTUAL will be ALINKed to another, non-aspectual event. As with TLINKs, they are created interactively within Knowtator with the basic form: *EVENT1 [aspectual relation] EVENT2***.** There are four different aspectual relations used in the schema, CONTINUES, INITIATES, REINITIATES, and TERMINATES. CONTINUES is used when an aspectual event shows the continuation of another event. INTIATES is used when an aspectual event indicates the start or initiation of another event. REINTIATES is used when an aspectual event indicates that another event will be restarted or reinitiated. TERMINATES is used when an aspectual event indicates the ending of another event (Table 2).

We consider the annotation of temporal and causal relations two separate tasks. An example of a discourse causal relation is the relation between *craniotomy* and *tumor* in [*There are again noted postoperative changes consistent with prior right frontoparietal **craniotomy** for resection of a right frontal brain **tumor**.]* Under the schema described in this paper, the relation between these two events is annotated as a temporal relation. The development of an annotation schema and guidelines for causal relations is a future task.

The corpus presented in this paper was manually annotated by a linguist and a domain expert working in collaboration. It was used to expand the TimeML annotation schema and guidelines to the clinical domain; hence no inter-annotator agreement (IAA) was computed which is a limitation of the current work. Future annotations will be created following the closed annotation methodology with IAA tracking. The annotations generated in this process will be used to build rich semantic representations of entities, events and temporal and causal relations. These structures will be the component blocks within an inferential mechanism for intra- and inter-sentential reasoning to constitute a comprehensive, open-source NLP system for clinical text

Conclusion

In this paper, we describe our work towards the larger task of temporal relation and timeline discovery from the clinical narrative. We evaluated the enabling technologies for such a system. In parallel, we implemented an annotation schema based on the general domain TimeML standards. These two steps are the foundation for our future efforts to build a comprehensive, open-source NLP system for clinical text that includes temporal reasoning.

References

[1] Pustejovsky J, Hanks P, Saur R, See A, Gaizauskas R, Setzer A, Radev D, Sundheim B, Day D, Ferro L, Lazo M. 2003. The timebank corpus. In Corpus Linguistics, pp. 647–656.

[2] Sauri R, Littman J, Knippen B, Gaizauskas R, Setzer A, Pustejovky J. 2006. TimeML annotation guidelines.http://www.timeml.org/site/publications/timeMLdocs/annguide_1.2.1.pdf

[3] Bethard S and Martin JH. 2006 Identification of event mentions and their semantic class. EMNLP.

[4] Boguraev B and Ando RK. 2005. Timebank-driven timeml analysis. In Graham Katz, James Pustejovsky, and Frank Schilder, editors, Annotating, Extracting and Reasoning about Time and Events, Dagstuhl Seminars. German Research Foundation.

[5] Saurı R, Knippen R, Verhagen M, Pustejovsky J. 2005. Evita: A robust event recognizer for qa systems. HLT-EMNLP, 2005.

[6] Verhagen M, Gaizauskas R, Schilder F, Hepple M, Pustejovsky J. 2007. Semeval-2007 task 15: Tempeval temporal relation identification. In SemEval- 2007.

[7] Puscasu G. 2007. Wvali: Temporal relation identification by syntactico-semantic analysis. In SemEval- 2007.

[8] Bethard S and Martin JH. 2007. CU-TMP: Temporal relation classification using syntactic and semantic features. In SemEval- 2007.

[9] Bethard S, Martin J, Klingenstein S. 2007. Finding Temporal Structure in Text: Machine Learning of Syntactic Temporal Relations. Proc IJSC, 1(4), Dec 2007.

[10] Palmer M, Gildea D, Kingsbury P. 2005. The Proposition Bank: A Corpus Annotated with Semantic Roles, Comp Linguistics J, 31:1, 2005.

[11] Kipper K, Korhonen A, Ryant N, Palmer M. 2006. Extensive Classifications of English verbs. Proc 12th EURALEX Int Congr. Turin, Italy. 2006.

[12] Zhou L, Friedman C, Parsons S, Hripcsak G. 2005. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports. Proc AMIA Symp. 2005:869-73.

[13] Zhou L and Hripcsak G. 2007. Temporal reasoning with medical data – a review with emphasis on medical natural language processing. JBI 40 (2007) 183-202.

[14] Augusto JC. 2005. Temporal reasoning for decision support in medicine. Artif Intell Med 2005; 33(1):1-24.

[15] Zhou L, Melton GB, Parsons S, Hripcsak G. 2006. A temporal constraint structure for extracting temporal information from clinical narrative. JBI 2006;39(4):424-39.

[16] Friedman C. 2000. A broad coverage natural language processing system. Proc AMIA Symp; 2000:270-4.

[17] Haug PJ, Ranum DL, Frederick PR. 1990. Computerized extraction of coded findings from free-text radiologic reports. Radiology 1990;174(2):543-8.

[18] Christensen LM, Haug PJ, Fiszman M. 2002. MPLUS: a probabilistic medical language understanding system. Proc Workshop on NLP in the biomedical domain; 2002. pp. 29-36.

[19] Hahn U, Romacker M, Schulz S. 2002. MedSyndikate – a natural language system for the extraction of medical information from findings reports. Int J Med Inform 2002;67(1-3):63-74.

[20] Hripcsak G, Zhou L, Parsons S, Das AK, Johnson SB. 2005. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. J Am Med Inform Assoc 2005;12(1):55-63.

[21] Savova GK, Kipper-Schuler K, Buntrock JD, Chute CG. UIMA-based clinical information extraction system. LREC 2008: Towards enhanced interoperability for large HLT systems: UIMA for NLP; 2008; Marrakech, Morocco; 2008

[22] Carreras C and Màrquez L. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. Proc CoNLL-2005, held in conjunction with ACL-2005, Ann Arbor, Michigan, June, 2005.

[23] Lease M and Charniak E. 2005. Parsing biomedical literature. In R. Dale, K.-F. Wong, J. Su, and O. Y. Kwong, eds, Proc 2nd IJCNLP, pp. 58–69, Jeju Island, Korea, Oct 11-13, 2005. Springer-Verlag.

[24] Daelemans W. 2008. Domain Adaptation in Memory-Based Shallow Parsing, Workshop on Parsing and Semantic Role Labeling, Tilburg, Netherlands, Feb, 2008.

[25] Zhou L, Parsons S, Hripcsak G. 2008. The evaluation of a temporal reasoning system in processing clinical discharge summaries. JAMIA, 15/1, pp. 99-106.

TLINK TYPE	Example sentence	TLINK
BEFORE	*He does have a history of peri-rectal **abscess** with his last round of **chemotherapy**.*	*(chemotherapy BEFORE abscess)*
AFTER	*The overall extent of contrast enhancement is **unchanged** since **February 29, 2005**.*	*(unchanged AFTER February 29th, 2005)*
OVERLAP	*Gengraf 300-mg p.o. b.i.d. (**decreased** in **early June**)*	*(decreased OVERLAP early June)*

Table 1: TLink examples

ALINK TYPE	Example sentence	ALINK
CONTINUE	*The patient will **remain** on **dialysis** until her condition changes.*	*(remain CONTINUES dialysis)*
INITIATE	*Patient will **begin** a high-fiber **diet** upon release.*	*(begin INITIATES diet)*
REINITIATE	*His anterior chest **rash** has not **reoccurred** since the PCN VK was discontinued 24-hours ago.*	*(reoccurred (not) REINITIATES rash)*
TERMINATE	*Because of this reaction, **Allegra** will be **discontinued***	*(discontinued TERMINATES Allegra)*

Table 2: ALink examples