Learning to Temporally Order Medical Events in Clinical Text

Preethi Raghavan*, Eric Fosler-Lussier*, and Albert M. Lai†
*Department of Computer Science and Engineering
†Department of Biomedical Informatics
The Ohio State University, Columbus, Ohio, USA
{raghavap, fosler}@cse.ohio-state.edu, albert.lai@osumc.edu

Abstract

We investigate the problem of ordering medical events in unstructured clinical narratives by learning to rank them based on their time of occurrence. We represent each medical event as a time duration, with a corresponding start and stop, and learn to rank the starts/stops based on their proximity to the admission date. Such a representation allows us to learn all of Allen's temporal relations between medical events. Interestingly, we observe that this methodology performs better than a classification-based approach for this domain, but worse on the relationships found in the Timebank corpus. This finding has important implications for styles of data representation and resources used for temporal relation learning: clinical narratives may have different language attributes corresponding to temporal ordering relative to Timebank, implying that the field may need to look at a wider range of domains to fully understand the nature of temporal ordering.

1 Introduction

There has been considerable research on learning temporal relations between events in natural language. Most learning problems try to classify event pairs as related by one of Allen's temporal relations (Allen, 1981) i.e., *before, simultaneous, includes/during, overlaps, begins/starts, ends/finishes* and their inverses (Mani et al., 2006). The Timebank corpus, widely used for temporal relation learning, consists of newswire text annotated for events, temporal expressions, and temporal relations between events using TimeML (Pustejovsky et al., 2003). In Timebank, the notion of an "event" primarily consists of verbs or phrases that denote change in state. However, there may be a need to rethink how we learn temporal relations between events in different domains. Timebank, its features, and established learning techniques like classification, may not work optimally in many real-world problems where temporal relation learning is of great importance.

We study the problem of learning temporal relations between medical events in clinical text. The idea of a medical "event" in clinical text is very different from events in Timebank. Medical events are temporally-associated concepts in clinical text that describe a medical condition affecting the patient's health, or procedures performed on a patient. Learning to temporally order events in clinical text is fundamental to understanding patient narratives and key to applications such as longitudinal studies, question answering, document summarization and information retrieval with temporal constraints. We propose learning temporal relations between medical events found in clinical narratives by learning to rank them. This is achieved by representing medical events as time durations with starts and stops and ranking them based on their proximity to the admission date.[1] This implicitly allows us to learn all of Allen's temporal relations between medical events.

In this paper, we establish the need to rethink the methods and resources used in temporal relation learning, as we demonstrate that the resources widely used for learning temporal relations in newswire text do not work on clinical text. When we model the temporal ordering problem in clinical text as a ranking problem, we empirically show that it outperforms classification; we perform similar experiments with Timebank and observe the opposite conclusion (classification outperforms ranking).

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

Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pages 70–74,
Jeju, Republic of Korea, 8-14 July 2012. ©2012 Association for Computational Linguistics

e1 *before* e2 e1.start e1.stop e2.start e2.stop	e1 *equals* e2 e1.start; e2.start e1.stop; e2.stop
e1 *overlaps with* e2 e1.start e2.start e1.stop e2.stop	e1 *starts* e2 e1.start; e2.start e1.stop e2.stop
e2 *during* e1 e1.start e2.start e2.stop e1.stop	e2 *finishes* e1 e1.start e2.start e1.stop; e2.stop

Table 1: Allen's temporal relations between medical events can be realized by ordering the starts and stops

2 Related Work

The Timebank corpus provides hand-tagged features, including tense, aspect, modality, polarity and event class. There have been significant efforts in machine learning of temporal relations between events using these features and a wide range of other features extracted from the Timebank corpus (Mani et al., 2006; Chambers et al., 2007; Lapata and Lascarides, 2011). The SemEval/TempEval (Verhagen et al., 2009) challenges have often focused on temporal relation learning between different types of events from Timebank. Zhou and Hripcsak (2007) provide a comprehensive survey of temporal reasoning with clinical data. There has also been some work in generating annotated corpora of clinical text for temporal relation learning (Roberts et al., 2008; Savova et al., 2009). However, none of these corpora are freely available. Zhou et al. (2006) propose a Temporal Constraint Structure (TCS) for medical events in discharge summaries. They use rule-based methods to induce this structure.

We demonstrate the need to rethink resources, features and methods of learning temporal relations between events in different domains with the help of experiments in learning temporal relations in clinical text. Specifically, we observe that we get better results in learning to rank chains of medical events to derive temporal relations (and their inverses) than learning a classifier for the same task.

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

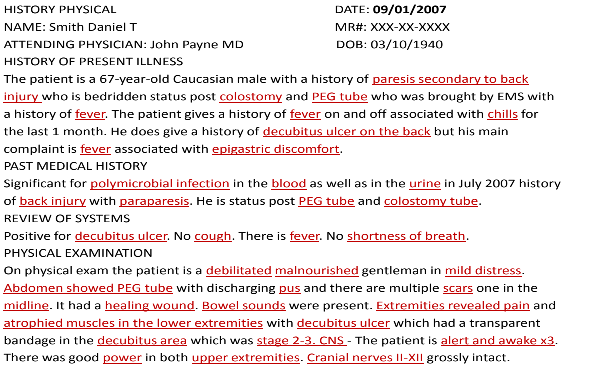
HISTORY PHYSICAL DATE: **09/01/2007**
NAME: Smith Daniel T MR#: XXX-XX-XXXX
ATTENDING PHYSICIAN: John Payne MD DOB: 03/10/1940
HISTORY OF PRESENT ILLNESS
The patient is a 67-year-old Caucasian male with a history of paresis secondary to back injury who is bedridden status post colostomy and PEG tube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

Figure 1: Excerpt from a sanitized clinical narrative (history & physical report) with medical events underlined.

3 Representation of Medical Events (MEs)

Clinical narratives contain unstructured text describing various MEs including conditions, diagnoses and tests in the history of a patient, along with some information on when they occurred. Much of the temporal information in clinical text is implicit and embedded in relative temporal relations between MEs. A sample excerpt from a note is shown in Figure 1. MEs are temporally related both qualitatively (e.g., *paresis before colostomy*) and quantitatively (e.g. *chills 1 month before admission*). Relative time may be more prevalent than absolute time (e.g., *last 1 month, post colostomy* rather than *on July 2007*). Temporal expressions may also be fuzzy where *history* may refer to an event *1 year ago* or *3 months ago*. The relationship between MEs and time is complicated. MEs could be recurring or continuous vs. discrete date or time, such as *fever* vs. *blood in urine*. Some are long lasting vs. short-lived, such as *cancer, leukemia* vs. *palpitations*.

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

4 Ranking Model and Experiments

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

The training data consists of medical event (ME) chains, where each chain consists of an instance of the start or stop of a ME belonging to the same clinical narrative along with a rank. The assumption is that the MEs in the same narrative are more or less semantically related by virtue of narrative discourse structure and are hence considered part of the same ME chain. The rank assigned to an instance indicates the temporal order of the event instance in the chain. Multiple MEs could occupy the same rank. Based on the rank of the starts and stops of event instances relative to other event instances, the temporal relations between them can be derived as indicated in Table 1. Our corpus for ranking consisted of 47 clinical narratives obtained from the medical center and annotated with MEs, temporal expressions, relations and event chains. The annotation agreement across our team of annotators is high; all annotators agreed on 89.5% of the events and our overall inter-annotator Cohen's kappa statistic (Conger, 1980) for MEs was 0.865. Thus, we extracted 47 ME chains across 4 patients. The distribution of MEs across event chains and chains across patients (p) is as as follows. p1 had 5 chains with 68 MEs, p2 had 9 chains with 90 MEs, p3 had 20 chains with 119 MEs and p4 had 13 chains with 82 MEs. The distribution of chains across different types of clinical narratives is shown in Figure 2. We construct a vector of features, from the manually annotated corpus, for each medical event instance. Although there is no real query in our set up, the admission date for each chain can be thought of as the query "date" and the MEs are ordered based on how close or far they are from each other and the admission date. The features extracted for each ME include the the type of clinical narrative, section information, ME polarity, position of the medical concept in the narrative and verb pattern. We extract temporal expressions linked to the ME like *history, before admission, past, during examination, on discharge, after discharge, on admission.* Temporal references to specific times like *next day, previously* are resolved and included in the feature set. We also extract features from each temporal expression indicating its closeness to the admission date. Differences between each explicit date in the narrative is also extracted. The UMLS(Bodenreider, 2004) semantic category of each medical concept is also included based on the intuition that MEs of a certain semantic group may occur closer to admission. We tried using features like the tense of ME or the verb preceding the ME (if any), POS tag in ranking. We found no improvement in accuracy upon their inclusion.

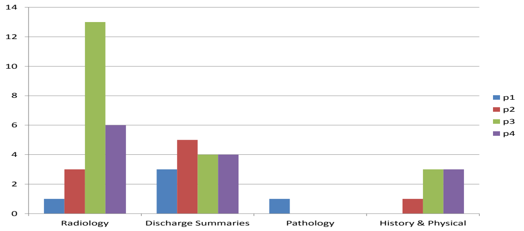

Figure 2: Distribution of the 47 medical event chains derived from discharge summaries, history and physical reports, pathology and radiology notes across the 4 patients.

In addition to the above features, we also anchor each ME to a coarse time-bin and use that as a feature in ranking. We define the following sequence of time-bins centered around admission, {*way before admission, before admission, on admission, after admission, after discharge*}. The time-bins are learned using a linear-chain CRF,[2] where the observation sequence is MEs in the order in which they appear in a clinical narrative, and the state sequence is the corresponding label sequence of time-bins.

We ran ranking experiments using SVM-rank (Joachims, 2006), and based on the ranking score assigned to each start/stop instance, we derive the relative temporal order of MEs in a chain.[3] This in turn allows us to infer temporal relations between

[2] http://mallet.cs.umass.edu/sequences.php

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

Relation	Clinical Text		Timebank	
	Ranking	Classifier	Ranking	Classifier
begins	81.21	73.34	52.63	58.82
ends	76.33	69.85	61.32	82.87
simulatenous	85.45	71.31	50.23	56.58
includes	83.67	74.20	59.56	60.65
before	88.3	77.14	61.34	70.38

Table 2: Per-class accuracy (%) for ranking, classification on clinical text and Timebank. We merge class ibefore into before.

all MEs in a chain. The ranking error on the test set is 28.2%. On introducing the time-bin feature, the ranking error drops to 16.8%. The overall accuracy of ranking MEs on including the time-bin feature is 82.16%. Each learned relation is now compared with the pairwise classification of temporal relations between MEs. We train a SVM classifier (Joachims, 1999) with an RBF kernel for pairwise classification of temporal relations. The average classification accuracy for clinical text using the same feature set is 71.33%. We used Timebank (v1.1) for evaluation, 186 newswire documents with 3345 event pairs. We traverse transitive relations between events in Timebank, increasing the number of event-event links to 6750 and create chains of related events to be ranked. Classification works better on Timebank, resulting in an overall accuracy of 63.88%, but ranking gives only 55.41% accuracy. All classification and ranking results from 10-fold cross validation are presented in Table 2.

5 Discussion

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

6 Conclusions

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

Acknowledgments

The project was supported by the NCRR, Grant UL1RR025755, KL2RR025754, and TL1RR025753, is now at the NCATS, Grant 8KL2TR000112-05, 8UL1TR000090-05, 8TL1TR000091-05. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

James F. Allen. 1981. An interval-based representation of temporal knowledge. In *IJCAI*, pages 221–226.

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

Nathanael Chambers, Shan Wang, and Daniel Jurafsky. 2007. Classifying temporal relations between events. In *ACL*.

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

Thorsten Joachims, Hang Li, Tie-Yan Liu, and ChengXiang Zhai. 2007. Learning to rank for information retrieval (lr4ir 2007). *SIGIR Forum*, 41(2):58–62.

Thorsten Joachims. 1999. Making large-scale SVM learning practical. In Bernhard Schölkopf, Christopher John C. Burges, and Alexander J. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184. MIT Press.

Thorsten Joachims. 2006. Training linear SVMs in linear time. In *KDD*, pages 217–226.

Mirella Lapata and Alex Lascarides. 2011. Learning sentence-internal temporal relations. *CoRR*, abs/1110.1394.

Inderjeet Mani, Marc Verhagen, Ben Wellner, Chong Min Lee, and James Pustejovsky. 2006. Machine learning of temporal relations. In *ACL*.

James Pustejovsky, Jos M. Castao, Robert Ingria, Roser Sauri, Robert J. Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R. Radev. 2003. TimeML: Robust specification of event and temporal expressions in text. In *New Directions in Question Answering'03*, pages 28–34.

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

Guergana K. Savova, Steven Bethard, Will Styler, James Martin, Martha Palmer, James Masanz, and Wayne Ward. 2009. Towards temporal relation discovery from the clinical narrative. *AMIA*.

Marc Verhagen, Robert J. Gaizauskas, Frank Schilder, Mark Hepple, Jessica Moszkowicz, and James Pustejovsky. 2009. The tempeval challenge: identifying temporal relations in text. *Language Resources and Evaluation*, 43(2):161–179.

Li Zhou and George Hripcsak. 2007. Temporal reasoning with medical data - a review with emphasis on medical natural language processing. *Journal of Biomedical Informatics*, pages 183–202.

Li Zhou, Genevieve B. Melton, Simon Parsons, and George Hripcsak. 2006. A temporal constraint structure for extracting temporal information from clinical narrative. *Journal of Biomedical Informatics*, pages 424–439.