Inter-Annotator Reliability of Medical Events, Coreferences and Temporal Relations in Clinical Narratives by Annotators with Varying Levels of Clinical Expertise

Preethi Raghavan, MS, Eric Fosler-Lussier, PhD, Albert M. Lai, PhD
The Ohio State University, Columbus, OH

Abstract

The manual annotation of clinical narratives is an important step for training and validating the performance of automated systems that utilize these clinical narratives. We build an annotation specification to capture medical events, and coreferences and temporal relations between medical events in clinical text. Unfortunately, the process of clinical data annotation is both time consuming and costly. Many annotation efforts have used physicians to annotate the data. We investigate using annotators that are current students or graduates from diverse clinical backgrounds with varying levels of clinical experience. In spite of this diversity, the annotation agreement across our team of annotators is high; the average inter-annotator kappa statistic for medical events, coreferences, temporal relations, and medical event concept unique identifiers was 0.843, 0.859, 0.833, and 0.806, respectively. We describe methods towards leveraging the annotations to support temporal reasoning with medical events.

Introduction

The clinical community creates and uses a variety of semi-structured and unstructured clinical narratives including medical reports such as progress notes, radiology reports, social work assessments and hospital discharge summaries. Longitudinal patient narratives account for various tests, procedures, and diagnoses in a patient's medical history. Patient narratives are often written in a medical sub-language with semantic categorization of words, domain specific terminology, incomplete phrases and omission of information. They are also very temporal in nature with multiple implicit and explicit temporal expressions co-occurring with medical events. However, the medical events in clinical narratives do not occur in a temporally coherent manner. The ability to automatically extract and induce temporal order in medical events found in clinical text has many applications in tasks such as document summarization, temporal information retrieval, and automatically matching patients to temporal clinical trials eligibility criteria.

High quality annotated narratives are essential for establishing training sets and gold standards for research. We build a gold standard corpus to facilitate development of methods to automatically learn the Temporal Constraint Structure[1] and to facilitate automatic temporal ordering of medical events in clinical text. In order to do this task efficiently, we need to solve other related problems such as medical event coreference resolution and learning to anchor temporal expressions to medical events. The idea is to create a corpus with high quality annotations, which will serve as a gold standard that can be leveraged for a number of different applications. These annotations are intended to eventually lead to the development of a system that can perform automatic temporal ordering of medical events in clinical narratives.

However creating annotated clinical corpora with such detailed features is tedious, expensive and requires experts with domain knowledge. Many clinical narrative annotation efforts have used physicians, which can potentially be cost-prohibitive. Within this population, it can be difficult to find individuals willing to devote the time and effort to doing manual annotations. The main focus of this paper is to describe and evaluate an annotation effort that leverages annotators that are current students or graduates from diverse clinical backgrounds with varying levels of clinical experience. We demonstrate that in spite of this diversity, the annotation agreement across the team of annotators is reasonably high.

The main contributions of this work are as follows:

1) **Measurement of inter-annotator reliability** by determining inconsistencies across annotators by various annotators at various levels.
    a. We measure the consistency of identifying a word or phrase as a medical event and applying the same concept code to the event.
    b. We also measure inconsistencies in noting that medical events corefer and in noting temporal relations between events.

2) **Demonstration of high agreement across annotators with diverse clinical expertise**. In this study, the annotators were current students and recent graduates from diverse medical and nursing backgrounds with varying levels of clinical experience. In spite of this diversity, we demonstrate that the annotation consistency across the team of annotators is high. We also describe the patterns of agreement between annotators from these different backgrounds.

Motivation

Time is an important aspect of longitudinal clinical narratives. Frequently, temporal expressions co-occur with medical events giving the clinician some intuition about when the event occurred relative to other events in the patient's history. However, the nature of temporal expressions tends to be varied and complicated. Moreover, the clinical narrative is usually temporally incoherent. Thus, there have been very few past efforts at tackling the hard problem of time in longitudinal clinical narratives.

In this paper, we describe annotation efforts in clinical narratives for marking medical events, related temporal expressions, temporal relations and coreference information in order to develop automatic methods to order medical events in clinical text by time. A step in this direction is to automatically induce the Temporal Constraint Structure (TCS) for clinical narratives. The TCS, proposed by Zhou et al.[1], models the time over which a medical event occurs as an interval. Each interval has "start" and "finish" time points, each of which may be constrained by temporal expressions. Based on the type of medical event and the context surrounding it, it may be possible to capture either the start time, finish time or both these times from the clinical narrative. Given medical events represented in this manner, automatically learning to order unique events on a timeline requires extensive linguistic analysis along with domain specific knowledge. An important step towards creating a timeline of unique medical events is coreference resolution. Multiple mentions of medical events may resolve to the same instance of a medical event. For instance, mentions of "*heart attack*," "*acute myocardial infarction*", and "*spontaneous myocardial infarction related to ischaemia*" may all refer to the same instance of "*myocardial infarction*" in the patient's history. Biomedical domain specific information such as the semantic group of the medical event, UMLS concept identifier, related concepts in the UMLS along with some information of when the medical event occurred may help the process of coreference resolution.

In this annotation effort, we intend to capture all such linguistic and domain-specific details that will allow development of methods to enable automatic induction of the TCS and facilitate temporal reasoning. Zhou et al. describes the TCS for medical events and temporal relations in the same discharge summary. However, we consider all clinical narratives such as admission notes, radiology and pathology reports, history and physical report, social work assessment report and discharge summaries in defining our annotation format. We demonstrate our efforts on a small subset of these clinical narratives. Importantly, we annotate the dataset of clinical narratives using annotators with diverse clinical backgrounds and demonstrate high agreement. This is valuable as getting expert physicians to annotate large datasets of clinical narrative is often infeasible and expensive.

Related Work

The TimeBank corpus[3] annotated using the TimeML specification is widely used for temporal relation learning in numerous natural language processing applications. A number of temporal relation annotation schemes have been developed for annotating clinical text.[2, 4, 5] In the medical domain, TimeText[6] defines annotations for temporal expressions found in discharge summaries. The annotated temporal expressions are extracted and represented using a Temporal Constraint Structure (TCS), which is then used for temporal reasoning.

There have been some efforts to annotate clinical text with coreference temporal relation information. Mowery et al.[7] annotated 24 clinical reports of different types and annotate temporal expressions using the format defined by Zhou et al.[2] They additionally annotate Trigger Terms that are explicit signals (words and phrases) in text other than temporal expressions whether a condition is recent or historical. However, they only examine temporal expressions annotation and the annotation format does not fully capture the requirements for temporal reasoning and coreference resolution within and across clinical narratives of a patient. We attempt to define an annotation format that supports these tasks and in turn enables creation of a longitudinal record of events over a patient's medical history.

Savova et al.[5] propose how they are going to work towards temporal relation discovery with the long-term goal of integrating temporal reasoning into cTAKES.[8] Their objective is to generate a timeline of events from clinical text. They observe how off the shelf parsers don't work well with medical data as most of the parsers are trained on The Wall Street Journal. They note for example that rash is typically an adjective in newswire, but is a noun in clinical

notes; erythema is not identified as a noun. They think a sufficient amount of domain training should help solve this. They also observe that in case of PropBank (used for semantic role labeling), the misidentification is due to inaccurate POS tagging. They propose using TimeML for tagging clinical narratives. They do not explain why or how these tags are the right choice.

Our annotations build on TimeML, but define new tags and tag attributes. Our contributions include the introduction of a new class of events called Medical. This is motivated by the observation that medical events do not have properties similar to events described in TimeML. Further, attributes like TENSE and ASPECT are not applicable to medical events. They may not be useful for other types of events either, as most clinical narratives are written in the PAST tense. Since our primary interest is in ordering medical events, we argue that there is a need to define special, domain-specific attributes for medical events that help identify coreferring medical events and help relate medical events to each other as well as other types of events, in order to generate a timeline of unique medical events.

Annotations

Medical Event

In the context of this work, a medical event or concept is any word or contiguous group of words found in a patient narrative that describes a medical condition affecting the patient's health. This includes all diseases and disorders, normal health situations like pregnancy that may affect the patient's health, as well as any treatments, procedures and drugs administered to the patient. A medical event describes the patient's state, at particular time point or time duration, as seen from a medical standpoint.

In order to make the annotation of medical events consistent across annotators, we ask them to consider words or phrases that have a meaningful and contextually relevant match in the Unified Medical Language System (UMLS version 2011AA).[9] The annotators have access to the UMLS Terminology Services Metathesaurus Browser.

The annotators were asked to mark the following as medical events:

1. Any disease or disorder. These include medical conditions, which are typically nouns. For example, *heart attack, chest pain, hypothermia, diastolic dysfunction* etc.
2. Any treatment, test or procedure. These are again generally nouns. For example, *echocardiogram, cholesterol profile* etc.
3. Any drugs administered. These are typically nouns which are names of drugs such as *beta blockers, nia*cin.
4. Any normal health condition that requires health care such as *pregnancy*.
5. Any observations related to the patient that may affect his health care. These could be nouns or verbs based on context (e.g., *smoking, drug abuse*).

Medical Event Coreference

Another important annotation label is the coreference label. Annotators are assigned the task of marking co-referential events within and across clinical narratives for each patient. This involves identifying all medical events across the narratives for a patient and determining which mentions of events across these clinical narratives are referring to the same instance of an event in time. This enables us to create a non-redundant list of medical events that have occurred in a patient's medical history.

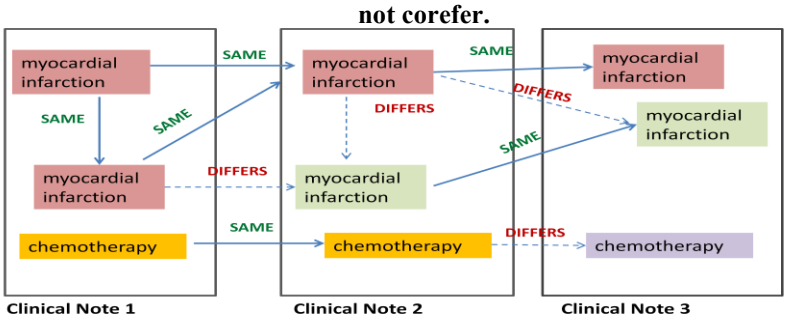
We can see examples of medical events that corefer within and across clinical notes in Figure 1. It also illustrates that multiple mentions of the same medical event may or may not resolve to the same instance of the medical event, i.e. corefer. The mentions of *myocardial infarction*/ *chemotherapy* indicated by the same color corefer whereas the ones that differ in color do not corefer.

We study the following coreference relations between medical event pairs, "same," "part of," and "type of." In case of "same" case, the medical event mentions are mostly synonymous and refer to the same instance. We also count abbreviations as part of this category of coreference relations. Examples of such coreference pairs include (*acute myocardial infarction, heart attack*), (B-*cell CLL, chronic lymphocytic leukemia*). In some cases, a medical concept may be "part of" or "type of" the other medical concept, as in the case of (*mass, specimen*) or (*B-cell CLL, evidence*).

Consider the following excerpt from a clinical narrative. "Endoscopic ultrasonographic examination of the upper gastrointestinal tract to the second part of the duodenum, performed 13 months before admission, revealed a large, submucosal, noncircumferential *mass*. The *mass* was well defined and did not appear to be locally invasive.

Examination of a *specimen* obtained by fine-needle aspiration showed *evidence* of an extranodal B-cell lymphoma of the MALT type."

Figure 1: Medical event coreferences within and across clinical notes of a patient. The medical events connected by arrow with label "SAME" corefer whereas the ones connected with the "DIFFERS" arrow do not corefer.



In this example, the medical events *specimen* and *evidence* corefer with the medical event *mass*. Here, *specimen* and *mass* are linked by a "part of" relation, i.e. specimen "part of" *mass*; whereas *specimen* and *evidence* are linked by "type of" relation i.e. *specimen* "type of" *evidence*. We intend to capture all such examples of coreference described above in our gold standard that will help us build and evaluate automatic methods to identify pairs of medical events that corefer.

Temporal Relations

We annotate temporal relations between medical events using two different representations: 1) Pairwise relations between events and 2) Events as time durations.

We identify if medical events are temporally related using a subset of Allen's temporal relations[10]. The annotators mark the following temporal relations between medical events: before, after, simultaneous, ends, begins, and includes. The other relations can be learned by inversing these relations. This is to evaluate if machine learning methods can be applied to learn these temporal relations between medical events as established in the case of newswire text in the general natural language processing domain.

We also model the start and finish time of medical events whenever possible. This allows us to model clinical narratives using the Temporal Constraint Structure described in Zhou et al.[2] Additionally, for each medical event, the annotator also marks how it is temporally related with other medical events in terms of Allen's temporal relations[10], as well as the overall temporal order of medical events within a narrative. However, in this paper, we calculate agreement only for the temporal relations across annotators, leaving the agreement for overall temporal order as part of our ongoing work.

Medical Event Concept Unique Identifier (CUI)

In addition to identifying which sets of words indicates a medical event, for the events identified, we also asked our annotators to identify the CUI from the UMLS[9] that was most appropriate for describing the medical event, including the most appropriate semantic type. We use the semantic types of the identified CUI as defined in the UMLS Semantic Network.

Data

The corpus used in this study consists of three clinic notes from a chronic lymphocytic leukemia (CLL) patient's record. This patient was one of approximately 2060 CLL patient records we have collected over the last 10 years at The Ohio State University Wexner Medical Center' (OSUWMC) and are continuing to annotate. The notes consisted of a discharge summary, radiology report and a history and physical report with an average of 600 words per narrative.

Methods

A team of 5 annotators with diverse backgrounds, but with some experience in understanding medical terminology was hired to annotate the corpus. The background profiles of our annotators were as follows. Our team had one

medical student with a degree biomedical engineering, but no clinical experience (medstud); three recently graduated nurse practitioners with clinical experience gathered through the process of receiving their nurse practitioner degrees (np); and one graduate entry nurse practitioner student with some clinical experience and experience in working with biomedical documents (nstud). In order to achieve the level of detail we wanted in our annotations, each annotator required approximately one month's effort. This included clinical informatics IRB training, getting them familiar with the UMLS, explaining the motivation behind the task, having them read and understand the annotation guidelines and annotating a sample clinical narrative. The annotations efforts described in this paper were coded in Excel sheets.

An important aspect of annotating a large corpus is consistency. We measure consistency in terms of inter-annotator reliability. Inter-annotator agreement measures the consistency in annotating a particular concept across annotators. We measure inter-annotator reliability using Cohen's kappa statistic[11]. Kappa is interpreted as the proportion of agreement among raters after chance agreement has been removed. It can be expressed as follows:

$$\text{Kappa} = \frac{\text{Proportion of observed agreement - chance agreement}}{\text{1 - chance agreement}}$$

Chance agreement is estimated by the proportion of agreements that would be expected if the observer's ratings were completely random. Chance agreement increases as the variability of observed ratings decreases. The use of Kappa requires minimal assumptions about the underlying nature of the data. Three data collection conditions should be met: 1) The subjects to be rated are independent of each other, 2) the raters score the subjects in an independent fashion, and 3) the rating categories are mutually exclusive and exhaustive. The flexibility of different forms of kappa is also a major advantage. Kappa is appropriate for nominal and ordinal data, where there are two or more raters per subject. Kappa can be calculated for each scale point or averaged into a generalized Kappa across the entire set of ratings.

We use the methods proposed by Conger[12] to calculate agreement between multiple annotators. Conger suggests a multiple-rater agreement statistic obtained by averaging all pairwise overall and chance-corrected probabilities proposed by Cohen (1960)[11].

For annotation of medical events, if the annotators marked a partially overlapping section of text as a medical event, we considered it to be an agreement. For medical event coreference, we considered agreement in a pairwise fashion. For example, for events A, B, and C, if annotator #1 identified events A and B corefer, and events B and C corefer, but annotator #2 only identified events B and C corefer, we count that as having 1 annotation in agreement. We also did not consider transitive closure. In the example given, if annotator #1 also identified events A and C coreferring and annotator #2 also identified events A and B as coreferring, which would count as having 2 annotations in agreement, and not 3. For temporal relations, we considered whether or not the same pairwise temporal relationship between events was identified. With medical events, we further analyzed whether or not the coders identified the medical events with the same UMLS CUIs.

Results

Inter-annotator agreement metrics

The total number of words in each clinical narrative (CN) is as follows: CN1= 454, CN2 = 612, CN3=386. Given the text in each narrative, the main unit of annotation is a medical event. Some examples of medical events in these clinical narratives include B-cell lymphoma, mass, physical examination, and beta blockers. We present statistics on the number of medical events, coreference pairs and temporal relations annotated by each annotator in the clinical narratives in Table 1.

We also present precision and recall metrics for each annotator measured against a reference annotator. In Tables 2, 3, and 4 respectively, we present precision and recall values for medical event mentions, coreference pairs, and temporal relation pairs across the three narratives with the medical student (medstud) as the reference annotator, respectively. The other annotators are a nursing student (nstud) and three nurse practitioners (np1, np2, np3).

Table 1. The number of medical events, coreference pairs and temporal relations noted by each annotator in three different clinical narratives.

	No. of Medical Events			Coreference-Pairs			Temporal Relations		
Annotator	CN1	CN2	CN3	CN1	CN2	CN3	CN1	CN2	CN3
medstud	65	81	53	15	19	12	15	19	12

nstud	58	70	58	8	10	7	8	10	7
np1	67	95	69	13	15	10	13	15	10
np2	52	87	70	12	16	10	12	16	10
np3	59	76	55	12	15	10	12	15	10

Table 2. Precision and recall values for medical event mentions across the three narratives with (medstud) as the reference annotator.

Annotator	CN1		CN2		CN3	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
nstud	94.82	84.61	89.3	87.5	88.2	80.4
np1	86.56	89.23	85.1	98.76	90.23	96.6
np2	96.1	76.92	90.8	97.53	88.57	95.2
np3	94.91	81.15	97.37	91.36	91.4	89.8

Table 3. Precision and recall values for coreference pairs across the three narratives with (medstud) as the reference annotator

Annotator	CN1		CN2		CN3	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
nstud	87.69	98.27	86.4	85.2	84.1	79.6
np1	92.53	89.65	72.6	98.57	94.56	94.3
np2	92.3	82.75	75.28	95.71	85.27	92.4
np3	84.7	86.2	92	100	93.4	90.7

Table 4. Precision and recall values for temporal relation pairs across the three narratives with (medstud) as the reference annotator.

Annotator	CN1		CN2		CN3	
	Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
nstud	96.92	94.02	98.76	85.26	92.76	82.46
np1	89.65	77.6	97.1	71.5	87.4	75.7
np2	92.3	71.64	96.5	88.4	86.8	85.8
np3	91.52	80.5	96.1	78.94	83.6	74.34

In Table 5, we present the average pairwise Cohen's kappa for medical events, coreferences, temporal relations, and medical event concept unique identifiers across narratives CN1, CN2, and CN3. To further illustrate and clarify our results, we plot these values in Figure 1.

Table 5. The average pair wise Cohen's kappa for medical events, coreferences, temporal relations, and medical event concept unique identifiers across CN1, CN2 and CN3.

Annotator Pairs	ME	Coref	TempRel	ME CUI
medstud, nstud	0.84	0.82	0.81	0.82
medstud, np1	0.86	0.90	0.85	0.78
medstud, np2	0.83	0.92	0.86	0.78
medstud, np3	0.85	0.91	0.88	0.80
nstud, np1	0.78	0.81	0.79	0.79
nstud,np2	0.85	0.82	0.85	0.81
nstud,np3	0.83	0.85	0.80	0.82
np1,np2	0.80	0.81	0.83	0.83
np1, np3	0.83	0.86	0.82	0.82
np2, np3	0.96	0.89	0.84	0.81
Average kappa	0.843	0.859	0.833	0.806

In looking at our results, we note that for medical events, the average kappa agreement between annotators from different backgrounds mostly varies between 0.80–0.85. The highest agreement is between np2 and np3 of 0.96. The lowest agreement was between nstud and np1 (kappa=0.78). For medical event coreference, the agreement is high

between the medical student and the nurse practitioners, with kappa ranging from 0.81–0.92. The highest agreement of 0.92 is between medstud and np2. With temporal relations, the kappa agreement varies between 0.79-0.88 with the highest Kappa of 0.88 between medstud and np3. When looking at medical event CUIs, the agreement varies between 0.78–0.83.

The overall average inter-annotator kappa statistic for medical events, coreferences, temporal relations, and medical event concept unique identifiers was 0.843, 0.859, 0.833, and 0.806 (Table 5), respectively, all of which show excellent agreement.[13] The average pairwise Cohen's kappa is highest between when medstud is paired with other annotators (kappa=0.86). The average of the pairwise agreement among the nurse practitioners is 0.846, whereas the average of the pairwise Cohen's kappa between each nstud/np from the nursing group and medstud is 0.82. This is across all three categories medical events, coreferences, and temporal relations.

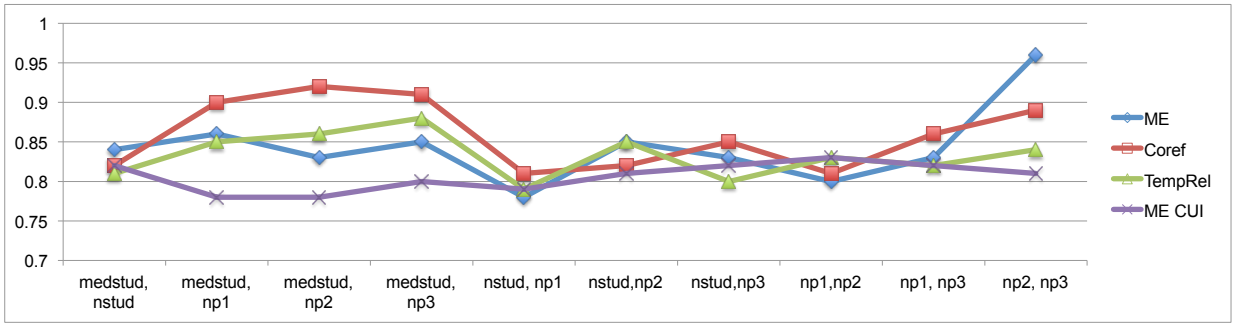


Figure 1. Pairwise Kappa agreement for medical events, coreference, temporal relations, and medical event CUIs. The pattern of agreement across the categories for different annotator pairs is more or less the same.

While overall agreement for coreferences is already high (0.859), it may be an underestimate. We did not consider transitive closure in our calculations. For example, if annotator #1 marked events A and B corefer, B and C corefer, and A and C corefer, but annotator #2 marked events A and B corefer and B and C corefer, we still consider there to be a missing annotation (A and C corefer).

Error Analysis

While the inter-annotator agreement for medical event CUIs was lower than for medical events, coreference, and temporal relations, agreement was still very high. Figure 1 indicates that the pattern of agreement across various annotations by different annotators with varying clinical expertise is more or less uniform.

In an analysis of the reason behind the discrepancies we discovered that in many cases there was either a discrepancy in the granularity to which the medical events were coded or whether or not clinical judgment was used in selecting the CUI. For example, all of our annotators marked "B-Cell CLL" as an event. The three NPs coded this term as "C0023434: Chronic Lymphocytic Leukemia." Both medstud and nstud coded this event as "C0475774: B-cell chronic lymphocytic leukemia variant." While both could be considered correct annotations for "B-Cell CLL," C0475774 is the more specific term. In another example, all of the annotators marked the phrase "white blood cell count of 10,000." For this situation, medstud selected "C0750426: white blood cell count increased," while nstud selected "C0023508: White Blood Cell count procedure." In contrast, all three NPs selected different CUIs, applying clinical judgment to the medical events. Np2 selected "C0860797: differential white blood cell count normal." Overall we found the medical student's (who did not have any real life clinic experience) annotations remained true to what was observed and could be inferred based on the data. However, the nursing student and the nurse practitioners often used clinical judgment to infer certain annotations that were not directly observed in the data. For instance, classifying something as an acute condition based on certain readings or values in the text.

Discussion

The temporal constraint structure (TCS) proposed by Zhou et al.[1] was restricted to individual discharge summaries. Our annotation scheme is intended to capture medical event attributes in longitudinal patient data found in the form of clinical text in admission and progress notes, radiology and pathology reports, discharge summaries and other types of clinical notes. Given these attributes, we can build systems to automatically induce the TCS. The TCS represents medical events related temporally and constrained by time. These can be learned with the help of the

medical event annotations and the temporal relation annotations respectively. We can further improve this structure by resolving coreferences between medical events that are the same. This leads to a more accurate and effective TCS representation for temporal reasoning between medical events. Moreover, since our annotations capture attributes in longitudinal patient data, the induced TCS can be used to temporally reason across medical events found in various clinical narratives of a patient. As a step in this direction, our initial efforts towards learning a coarse temporal order for medical events within a clinical narrative is explained in Raghavan et.al.[14] Further, our efforts towards automatically learning pairs of medical events that corefer is illustrated in Raghavan et al.[15, 16].

Consider the following excerpt from a patient case report in the New England Journal of Medicine:

"**Twenty-two months** after the *implantation procedure*, the level of *PSA* was 0.6 ng per milliliter; the *PSA* was less than 0.1 ng per milliliter **5 years after** the procedure. **Approximately 6 months before** this evaluation, routine follow-up *urinalysis* revealed *microscopic hematuria*. The level of *PSA* was reportedly less than 0.2 ng per milliliter."

There are 3 mentions of PSA in the excerpt. A PSA is a prostate specific antigen test that measures a protein produced by the cells of the prostate. As these were 3 separate tests involving 3 separate blood draws at different times (and having 3 different values), they do not corefer. But there may be other mentions of PSA in the narrative that do corefer. We can learn to automatically identify which mentions of PSA corefer with the help of the annotated temporal relations. The temporal relations clearly indicate the temporal order of the 3 PSA tests. This in turn helps us differentiate that the mentions of the PSA tests are indeed 3 different instances of a PSA test.

However, we do not annotate for pronoun anaphora resolution. Savova et al.[17] focused on the annotation of anaphoric coreference.

One limitation of our study is the small number of narratives. The main reason for this limitation is due to the large amount of effort required to annotate the narratives to the detail that we desired. Given that the 3 narratives used in the study required a month of effort for each annotator, we needed to begin having the annotators annotate non-overlapping narratives in order to increase the overall size of our gold standard. Another limitation of our study is the lack of a physician annotator to compare their annotations. Given the amount of time required for our existing annotators to complete the annotations, having an additional physician annotator was not feasible. One reason for the time required to generate these annotations may be the lack of sophisticated annotation tools. In future, we plan to create the annotation schema using Knowtator[18] and train our annotators to use this tool.

Conclusion

We have demonstrated that across our diverse group of annotators, ranging from current students to recent graduates from diverse medical and nursing backgrounds with varying levels of clinical experience, inter-annotator agreement is high for medical event identification, medical event coreference, temporal relation annotation, and medical event concept coding, inter-annotator reliability is high. The average inter-annotator kappa statistic for medical events, coreferences, temporal relations, and medical event concept unique identifiers was 0.843, 0.859, 0.833, and 0.806, respectively. The agreement was calculated on a small subset of chronic lymphocytic leukemia narratives.

Many other clinical narrative annotation efforts have used physicians, which can potentially be cost-prohibitive. In addition, identifying physicians who are willing to devote the time and effort to performing annotations is difficult. By using current students and recent graduates of clinical degree programs, we are able to capitalize on a readily available and valuable resource for annotation intensive research Considering the high inter-annotator agreement amongst this diverse cohort of annotators, we believe that it is unnecessary to utilize physicians for performing the annotations described in this study. However, we intend to further evaluate the quality of the annotations that we have generated by having a group of physicians review the annotations from the notes in this study. We believe that this study provides the foundations upon which to build upon for doing a larger study regarding the level of clinical expertise needed for developing a gold standard.

Acknowledgements

The project described was partially supported by the National Center for Research Resources, Grant UL1RR025755, KL2RR025754, and TL1RR025753, and is now at the National Center for Advancing Translational Sciences, 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

1. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform. 2006 Aug;39(4):424-39.
2. Pustejovsky J, Castaño J, Ingria R, et al. TimeML: Robust Specification of Event and Temporal Expressions in Text. Fifth International Workshop on Computational Semantics (IWCS-5); 2003; Tilburg, The Netherlands; 2003.
3. Pustejovsky J, Verhagen M, Sauri R, et al. TimeBank 1.2. Philadelphia: Linguistic Data Consortium; 2006.
4. Galescu L, Blaylock N. A corpus of clinical narratives annotated with temporal information. Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. Miami, Florida, USA: ACM; 2012. p. 715-20.
5. Savova GK, Bethard S, Styler W, et al. Towards temporal relation discovery from the clinical narrative. AMIA Annu Symp Proc. 2009;2009:568-72.
6. Zhou L, Parsons S, Hripcsak G. The evaluation of a temporal reasoning system in processing clinical discharge summaries. J Am Med Inform Assoc. 2008 Jan-Feb;15(1):99-106.
7. Mowery DL, Harkema H, Chapman WW. Temporal Annotation of Clinical Text. BioNLP 2008: Current Trends in Biomedical Natural Language Processing; 2008; 2008. p. 106-7.
8. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010 September 1, 2010;17(5):507-13.
9. Allen JF. Towards a general theory of action and time. Artif Intell. 1984;23(2):123-54.
10. Bodenreider O. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. Nucleic Acids Research. 2004; 2, D267-D270.
11. Cohen J. A Coefficient of Agreement for Nominal Scales. Educational and Psychological Measurement. 1960;20:37-46.
12. Conger AJ. Integration and Generalization of Kappas for Multiple Raters. Psychological Bulletin. 1980;88(2):322-8.
13. Fleiss JL. Statistical methods for rates and proportions. 2nd ed. New York: John Wiley; 1981.
14. Raghavan P, Fosler-Lussier E, Lai AM. Temporal Classification of Medical Events, BioNLP 2012, Workshop of North American Association for Computational Linguistics - Human Language Technologies Conference, 2012.
15. Raghavan P, Fosler-Lussier E, Brew C, Lai AM. Medical event coreference resolution using the UMLS metathesaurus and temporal reasoning. 2nd ACM SIGHIT Symposium on International Health Informatics. 2012;2012:465-472.
16. Raghavan P, Fosler-Lussier E, Lai AM. Exploring Semi-Supervised Coreference Resolution of Medical Concepts using Semantic and Temporal Features, North American Association for Computational Linguistics - Human Language Technologies Conference (NAACL HLT 2012), 2012.
17. Savova GK, Chapman WW, Zheng J, Crowley RJ. Anaphoric relations in the clinical narrative: corpus creation. JAMIA 2011; 18(4): 459-465.Ogren PV. Knowtator: A Protégé plug-in for annotated corpus construction. Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology: Companion Volume: Demonstrations (NAACL-Demonstrations 2006). Association for Computational Linguistics, Stroudsburg, PA, USA, 2006; 273-275.
18. Zhou L, Hripcsak G. Temporal reasoning with medical data – A review with emphasis on medical natural language processing. J Biomed Inform. 2007; 40(2):183-202.