Journal of Biomedical Informatics 46 (2013) S5–S12

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Methodological Review

Annotating temporal information in clinical narratives

Weiyi Sun [a,*], Anna Rumshisky [b], Ozlem Uzuner [c]

[a] Department of Informatics, University at Albany, SUNY, 1400 Washington Ave., Draper 114B, Albany, NY 12222, United States
[b] Department of Computer Science, University of Massachusetts, 198 Riverside St., Olsen Hall, Lowell, MA 01854, United States
[c] Department of Information Studies, University at Albany, SUNY, 1400 Washington Ave., Draper 114A, Albany, NY 12222, United States

ARTICLE INFO

Article history:
Received 21 May 2013
Accepted 10 July 2013
Available online 19 July 2013

Keywords:
Natural Language Processing
Temporal Reasoning
Medical Informatics
Corpus Building
Annotation

ABSTRACT

Temporal information in clinical narratives plays an important role in patients' diagnosis, treatment and prognosis. In order to represent narrative information accurately, medical natural language processing (MLP) systems need to correctly identify and interpret temporal information. To promote research in this area, the Informatics for Integrating Biology and the Bedside (i2b2) project developed a temporally annotated corpus of clinical narratives. This corpus contains 310 de-identified discharge summaries, with annotations of clinical events, temporal expressions and temporal relations. This paper describes the process followed for the development of this corpus and discusses annotation guideline development, annotation methodology, and corpus quality.

1. Introduction

Electronic Medical Records (EMRs) contain significant amounts of unstructured narrative text, which can be turned into structured data with help from automated medical language processing (MLP) systems. Some sub-areas of MLP, such as de-identification and clinical concept (e.g. disorder, medication) extraction are well-studied. Other areas, such as analysis of the temporal structures embedded in clinical texts, are less so [1]. Besides being a more complicated task, we believe that the lack of availability of manually annotated clinical corpora with temporal information also hindered the progress of MLP in this area [2].

Temporal information in clinical narratives plays an important role in medical decision-making and care assessment [3]. Some examples of clinical applications that utilize temporal information include: diagnosis, prognosis and treatment decision support [3,4], time specific clinical information extraction [5–7], and time-related question answering [8–10]. These applications rely on *temporal reasoning systems* which extract temporal information from natural language, and perform temporal inference over the extracted information. *Temporal information* in narrative texts includes the events and the temporal expressions that appear in the text, as well as the temporal relations among them.

In order to develop and evaluate temporal reasoning systems, we need clinical corpora annotated with temporal information. Given this need, the 2012 Informatics for Integrating Biology and the Bedside (i2b2) project provided the community with a corpus of temporally annotated clinical narratives [2] This corpus contains the clinical history and the hospital course sections of 310 de-identified discharge summaries from Partners Healthcare and the Beth Israel Deaconess Medical Center, for a total of approximately 178,000 tokens. The corpus was annotated for three layers of information: events, temporal expressions and their normalization, and temporal relations. Our annotation scheme was adapted from TimeML [11]. More specifically, we annotate three types of temporal information: (1) EVENTs which represent the semantic events mentioned in the text that affect the patient's clinical timeline; (2) TIMEX3s that represent temporal expressions of date, times, durations, and frequencies; and (3) TLINKs which represent the temporal relations between EVENTs and TIMEX3s. We refer to this corpus as the *i2b2 temporal relations corpus*.

In this paper, we present the process followed for the development of the i2b2 temporal relations corpus, including the creation of a temporal annotation scheme tailored to clinical narratives, the methodology for applying the scheme to the i2b2 temporal relations corpus, the evaluation of the resulting annotation quality, and a description of the resulting annotated corpus. We hope that this paper will: (1) inform the MLP researchers about the temporal data preparation process, (2) guide the development of future clinical temporal annotation guidelines, (3) caution against pitfalls and the issues often raised in the representation of temporal information, and (4) share our solutions to these problems with the community.

The remainder of this paper is organized as follows: Section 2 summarizes the related work in temporal representation and existing temporally annotated corpora in the general as well as the MLP domains. Section 3 summarizes our annotation guidelines. Sec-

[*] Corresponding author.
E-mail address: wsun2@albany.edu (W. Sun).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.004

tion 4 presents our annotation methodology and procedures. The annotation quality evaluation and corpus statistics are presented in Sections 5 and 6, respectively.

2. Related work

2.1. Temporal annotation

A *temporal representation scheme* translates time-related information into a computer readable form to support temporal reasoning. Temporal annotation is a type of temporal representation that focuses on interpreting time-related natural language information. Defining a temporal representation scheme is non-trivial in that it requires the specification of many fundamental assumptions about time [12]. This task becomes even more challenging when the targeted temporal information is embedded in natural language because time-related concepts are usually vaguely and implicitly conveyed in free text [13,14]. For instance the verbal event 'know' describes a continuous state, and the event 'catch' is instantaneous [2].

The most prominent challenges in temporal annotation include: (1) large search space in the assignment of TLINKs. Given the EVENTs and TIMEX3s in one document, the theoretical search space for TLINKs is (N-1)N/2 (N: total number of entities). (2) Multiple ways to represent the same set of TLINKs. TLINKs can be transitive (e.g. before or after) or equivalent (e.g. concurrence). For example, let '<' represent the before relation, '>' for the after relation, and '=' for the concurrence relation, and for entities A, B, and C, we have 'A < B, B = C'. We can equivalently represent this relation with 'B > A, A < C, B = C' among many other sets of TLINKs, which give the annotators flexibility during annotation but whose equivalence is difficult to manually confirm. For this reason, we usually need to compute temporal closure when handling TLINKs. The *temporal closure* of a set of TLINKs is the set of minimal transitive relations that contains the original TLINKs. In the previous example, the temporal closure of 'A < B, B = C' contains the following relations: 'A < B, A < C, B > A, C > A, B = C, C = B'. (3) Conflict in TLINKs. The transitive and equivalent relations can give rise to conflicts in the annotations. For example, if 'A < C' is already established, then annotating 'A > C' would create a conflict. Such a conflict can be inferred from 'non-conflicting' relations (e.g. A = B, B > C) during temporal closure and can be difficult for even human annotators to spot. We will discuss our approaches to addressing these issues in Section 4.

2.2. Existing temporal annotation guidelines and corpora

With temporal reasoning attracting increasingly more research attention [15], the creation of temporally annotated datasets becomes a pressing task. As a result, several temporal annotation schemes and annotated datasets have become available [16–21]. In the general domain, these datasets include:

- TimeBank [16] and the AQUAINT corpora[1] contain newswire articles annotated under the TimeML guidelines [11]. The TimeBank corpus contains 183 news reports and the AQUAINT corpus contains 73 news reports. The TimeML guidelines specify three types of entities (EVENTs, TIMEX3s, and Signals) as well as three types of relations (TLINKs, ALINKs, and SLINKs). In addition to the EVENT, TIMEX3 and TLINK tags that we introduced in Section 1, signals are functional words or phrases that indicate the temporal relation between two entities; ALINKs describe the aspectual relation between entities, such as initiating, terminating and continuing; SLINKs indicate the subordinate relations between EVENTs (e.g. the conditional or evidential relations between two EVENTs) [22].
- The TempEval [17–19] 2007 corpus applied a simplified TimeML annotation, and restricted the TLINK assignment to those (1) between EVENTs and document creation times, that is, the time stamp of the document creation; (2) between EVENTs/TIMEX3s in the same sentence and (3) between main EVENTs (syntactically dominating verbs) in adjacent sentences. The TempEval 2010 extended the 2007 annotations to multiple languages. TempEval 2012 used subsets of the TimeBank and AQUAINT corpora, as well as an automatically annotated English Gigaword corpus [23–25].

In the clinical domain, Galescu and Blaylock [21] applied an adaptation of TimeML guidelines to 40 discharge summaries [26]. Savova et al. [27] also described an adaptation of TimeML to clinical narratives. The Clinical E-Science Framework (CLEF) project annotated a corpus of 167 clinical records for temporal relations [20]; however, they limited their annotations to intra-sentence temporal relations and to the temporal relations between events and document creation times. In addition to these full temporal relation annotation schema, there are also annotations that focus on some more specific temporal elements in the clinical narratives, such as conditions, temporal expressions [28,7,29]. These resources served as a good start at addressing the need for a temporally annotated MLP corpus, and highlighted the need for comprehensive temporal annotations that can support the extraction of the complete patient clinical timeline from narrative patient records. We aimed to fill this gap.

3. Design of i2b2 annotation guidelines

We built our annotation guidelines on the following principles:

(1) **Ease of Use**: the annotators should become proficient in the task after a few short training sessions, and the human annotation burden should be light.
(2) **Completeness**: the annotation should capture a broad range of key clinical concepts and it should support complete timeline extraction from medical records.
(3) **Definitiveness**: the guidelines should be unambiguous so as to ensure inter-annotator agreement.
(4) **Maximum utilization of existing annotations**: The guidelines should reuse and add value to existing corpora and annotations.

With these design principles in mind, we separated the annotation task into: clinical event annotation (EVENT), temporal expression annotation (TIMEX3), and temporal relation annotation (TLINK). Two annotators with clinical background assisted in the development of the annotation guidelines. After each round of pilot training (see Fig. 1), the annotators were asked to independently annotate 5 clinical records (pilot annotations). We analyzed the errors and the disagreements in the pilot annotations after each round, and modified the guidelines accordingly. We repeated this process until the annotations stabilized. The guideline development process lasted two months. The finalized annotation guidelines can be found in Appendix.

3.1. Annotation scope

We annotated a corpus consisting of de-identified discharge summaries from Partners Healthcare and the Beth Israel Deaconess Medical Center [26,30,31]. After analyzing a set of stratified sam-

[1] http://timeml.org/site/timebank/timebank.html.

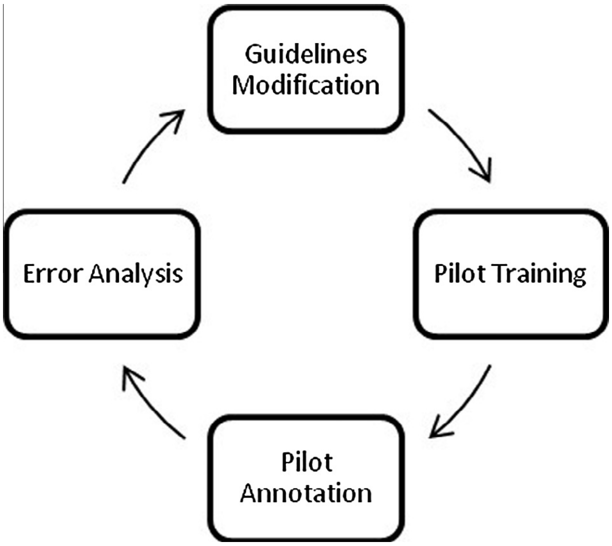

Fig. 1. Annotation guidelines development process.

ples from these sources, we found that the clinical history and the hospital course sections of discharge summaries contained abundant temporal information expressed in narrative text. We therefore focused our efforts on these sections.

Our temporal annotation guidelines are adapted from TimeML. In addition to TimeML, we consulted the annotation guidelines of the THYME project [32]. As an effort to simplify the annotation task, we removed TimeML's SIGNALs, as well as the ALINKs and the SLINKs. Our guidelines included EVENTs, TIMEX3s and TLINKs, with modified attributes (see Sections 3.2, 3.3 and 3.4). We also introduced a SECTIME (section time) tag, which keeps track of the 'section creation date' of each section in the discharge summary. The SECTIME for the clinical history section is defined as the date of admission, and the SECTIME for the hospital course section is the date of discharge.

Our i2b2 temporal relations corpus included previously generated layers of gold standard annotations, in the form of clinical concepts (problems, tests, treatments) [30] and coreference relations [31] which can support temporal reasoning. Locating a patient's disease, treatment and test results on a timeline is important for care providers. Coreference, linking two mentions that refer to the same incidence of the same event, is a prerequisite for temporal reasoning. We used clinical concepts as pre-annotated EVENTs (see Section 3.2), and the coreference relations as SIMULTANEOUS type TLINKs (see Section 3.4).

3.2. EVENT annotation

EVENTs include: clinical concepts (i.e. PROBLEMs, TESTs and TREATMENTs [30]), clinical departments (the mentions of the clinical departments or services where the patient was, is or will be admitted to), EVIDENTIALs (words or phrases that indicate the source of information such as the word 'complained' in 'The patient complained about a week-long headache') and OCCURRENCEs (other events such as 'admit', 'transfer' or 'discharge', ... that affect the patient's clinical timeline).

EVENTs have three attributes, TYPE, MODALITY and POLARITY. The TYPE attribute specifies the EVENT as a PROBLEM, TEST, TREATMENT, CLINICAL_DEPT, EVIDENTIAL or OCCURRENCE. MODALITY specifies if an EVENT is factual, hypothetical, hedged or conditional. POLARITY specifies whether an EVENT has positive (POS) or negative (NEG) polarity. Fig. 2 shows a snippet of a sample discharge summary; the EVENTs in this record are shown in Table 1.

3.3. Temporal expression annotation

Temporal expressions in the clinical records are marked as TIMEX3s. Our guidelines include four types of TIMEX3s: dates, times, durations and frequencies. Each TIMEX3 needs to be normalized to the ISO8601 standard in its value (VAL). ISO8601 requires date/time TIMEX3s to be normalized to [YYYY-MM-DD]T[HH:MM] format, and duration/frequency TIMEX3s to be normalized to R[#1 times]P[#2][Units] (repeat for #1 times during #2 units of time). For example, 'twice every three weeks' is normalized as R2P3W. Like the TimeML TIMEX3s, the i2b2 TIMEX3s also have a modifier attribute (MOD), which represents a subset of the TimeML TIMEX3 modifier values: MORE, LESS, APPROX, START, END MIDDLE and the default NA. Table 2 shows the sample annotations of TIMEX3s in the snippet displayed in Fig. 2. TimeML uses temporal function, a mechanism that allows TIMEX3s to anchor to each other, to handle durations and relative time annotations. To simplify the annotation procedure, we omitted temporal functions, and used TLINKs between two TIMEX3s to handle the anchoring of durations and relative times (see guidelines for details).

3.4. Temporal relation annotation

TLINKs mark the temporal relation between EVENTs and TIMEX3s. Our TLINK TYPEs include a subset of the TimeML TLINK TYPEs. These TYPEs are: BEFORE, AFTER, BEGUN_BY, ENDED_BY, DURING, SIMULTANEOUS, OVERLAP and BEFORE_OVERLAP. Table 3 shows the TLINKs of the snippet in Fig. 2.

In order to support the extraction of a complete timeline from discharge summaries, our guidelines allow the annotators to assign TLINKs to any pair of EVENTs/TIMEX3s in a record. Nonetheless, as pointed out in Section 2.1, there are multiple ways to represent the same set of TLINKs (e.g. any relations in the set 'A < B, A < C, B > A, C > A, B = C, C = B' are correct for representing 'A < B, B = C'). Requiring the annotator to mark every relation in the temporal closure is time-consuming and unnecessary. Instead, we informed our annotators that we would compute temporal closure on the TLINKs that they marked, and hence they only needed to mark a minimal set of TLINKs. We provided them the following instructions to help them select candidate entity pairs when facing multiple possibilities to assign TLINKs:

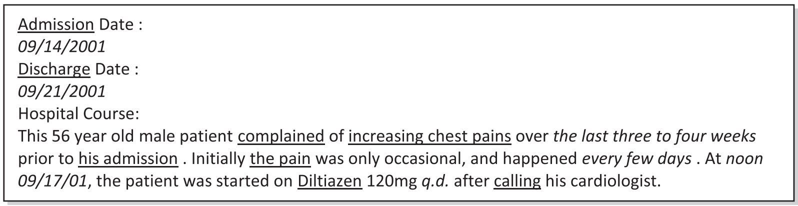
Admission Date :
09/14/2001
Discharge Date :
09/21/2001
Hospital Course:
This 56 year old male patient complained of increasing chest pains over *the last three to four weeks* prior to his admission . Initially the pain was only occasional, and happened *every few days* . At *noon 09/17/01,* the patient was started on Diltiazen 120mg *q.d.* after calling his cardiologist.

Fig. 2. Sample clinical record snippet (Underscore: EVENTs, Italics: TIMEX3s).

Table 1
EVENT annotation examples.

Event	Type	Modality	Polarity
[Admission]	OCCURRENCE	FACTUAL	POS
[Discharge]	OCCURRENCE	FACTUAL	POS
[complained]	EVIDENTIAL	FACTUAL	POS
[increasing chest pains]	PROBLEM	FACTUAL	POS
[his admission]	OCCURRENCE	FACTUAL	POS
[the pain]	PROBLEM	FACTUAL	POS
[Diltiazen]	TREATMENT	FACTUAL	POS
[calling]	OCCURRENCE	FACTUAL	POS

Table 2
TIMEX3 annotation examples.

TIMEX3	Type	Val	Mod
[09/14/2001]	DATE	2001-09-14	NA
[09/21/2001]	DATE	2001-09-21	NA
[the last three to four weeks]	DURATION	P3.5W	APPROX
[every few days]	FREQUENCY	RP2D	APPROX
[noon 09/17/01]	TIME	2001-08-17T12:00	NA
[q.d.]	FREQUENCY	RP1D	NA

Table 3
TLINK annotation examples.

From extent	Type	To extent
[Admission]	SIMULTANEOUS	[09/14/2001]
[Discharge]	SIMULTANEOUS	[09/21/2001]
[complained]	BEFORE	SECTIME: 09/21/2001
[increasing chest pains]	BEFORE	SECTIME: 09/21/2001
[increasing chest pains]	OVERLAP	[the last three to four weeks]
[increasing chest pains]	BEFORE_OVERLAP	[complained]
[his admission]	BEFORE	SECTIME: 09/21/2001
[the last three to four weeks]	ENDED_BY	[his admission]
[the pain]	BEFORE	SECTIME: 09/21/2001
[the pain]	SIMULTANEOUS	[increasing chest pain]
[the pain]	OVERLAP	[every few days]
[Diltiazen]	BEFORE_OVERLAP	SECTIME: 09/21/2001
[Diltiazen]	OVERLAP	[q.d.]
[Diltiazen]	BEGUN_BY	[noon 09/17/01]
[Diltiazen]	AFTER	[calling]
[calling]	BEFORE	SECTIME: 09/21/2001

- A TLINK can only be assigned to a pair of TIMEX3s, if:
  - it anchors a relative TIMEX3 (e.g. last Friday, three days before discharge) to an absolute TIMEX3 (e.g. a calendar date),
  - it marks the start point or the end point of a duration.
- A TLINK involving at least one EVENT can be marked, if:
  - there is a TIMEX3 in the same sentence or in adjacent sentences,
  - an explicit relation between EVENTs is signaled by words such as 'before' or 'after',
  - there is an implicit relation, such as a causal or concurrent relation, between EVENTs, and if such a relation cannot be inferred from existing TLINKs.
- A TLINK must be assigned to every EVENT and its SECTIME.

4. Annotation procedure

The i2b2 temporal relations corpus was annotated in a single-pass, dual annotation with adjudication. As Fig. 2 illustrates, the annotation process started with data selection and pre-annotation on EVENTs and coreference relations (i.e., SIMULTANEOUS TLINKs). Each record, with its pre-annotations, was assigned to two independent annotators. These annotators made a single pass over each record, completing all three layers of annotation in a single pass. Our pilot showed that single pass annotation was more efficient for our project. As opposed to a multi-pass annotation which requires the annotators to complete one of the layers, submit for adjudication, and continue to annotate the next layer on adjudicated records, single pass annotation had each annotator complete all three layers of annotation in a clinical record and then submit for adjudication, reducing total reading time as well as the overhead in submitting and receiving assignments. (see Fig. 3)

4.1. Annotator expertise

The team consists of eight annotators, four of whom have medical background. Roberts et al. [20] showed that annotators with medical background are more likely to find relations between clinical events. Our pilot study also showed that annotators with medical background were more successful in interpreting ambiguous or uncommon abbreviations, as well as finding TLINKs that were based on causal relations between clinical concepts. Therefore, in the dual annotation, the annotators with medical background were each paired with an annotator without medical background.

4.2. Annotation tool

We chose to use the Multi-purpose Annotation Environment (MAE) toolkit for annotation and the Multi-document Adjudication Interface (MAI) toolkit for adjudication [33]. Due to the fact that we allow TLINKs to span sentences, and even sections, the annotation tool needs to display each clinical record in its entirety during annotation. The MAE/MAI tools also enable fast look-up of all relations of an entity as well as the look-up of all entities involved in a relation, which makes the tool an ideal choice for our task.

4.3. Training

The annotator training process is shown in Fig. 4. We started with a 2-h group tutorial meeting. Afterwards, each annotator received 5 training discharge summaries for practice. The trainer then reviewed and conducted error analysis on these practice annotations. Afterwards, the trainer held individual meetings with the annotators, as necessary, to better understand the sources of errors. The entire training process was repeated twice before annotations stabilized. The average time that an annotator spent in training (including full annotation of 10 training records) was 15.25 h. During the practice annotation, the annotators were encouraged to utilize an online discussion board to raise questions and help each other understand the guidelines. A total of 37 threads, containing 128 messages, were posted in the discussion board. We found that the discussion board was helpful for annotators to quickly find answers to their questions when the trainer was not available; it also helped the trainer clarify the guidelines and prepare more targeted training sessions.

4.4. Dual annotation

The dual annotation ensured that each record was annotated by at least one medical background annotator. The average time for one annotator to complete a full annotation of one clinical record was about 55 min. The overall annotator-hours spent in annotation are 568 h.

4.5. Adjudication

The disagreements between the annotators were presented to an adjudicator, different from the original annotators for tie-break-

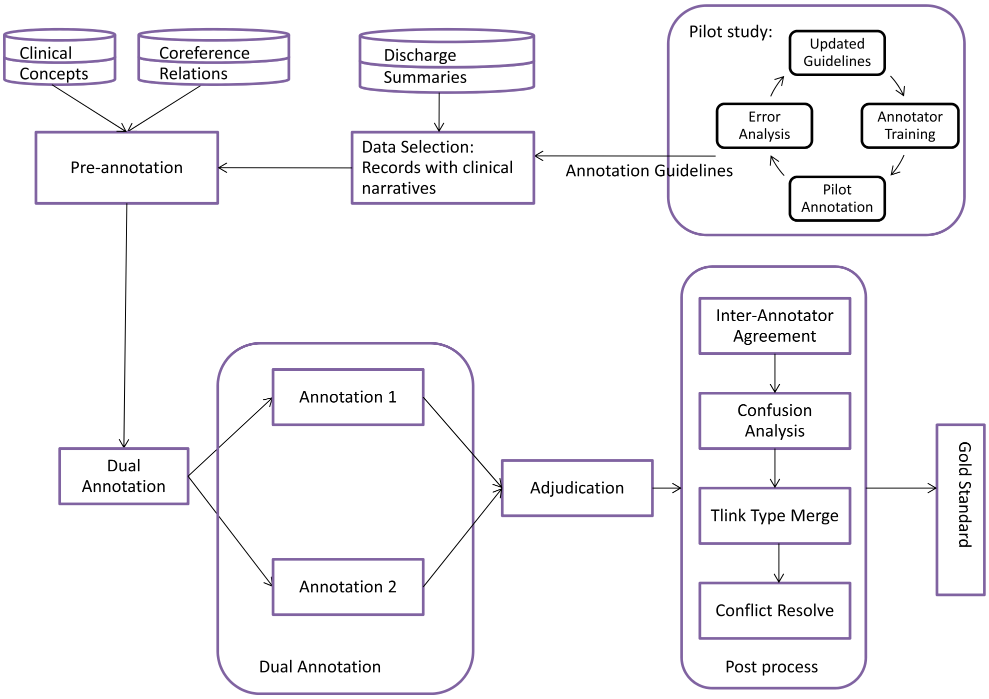

Fig. 3. Annotation process.

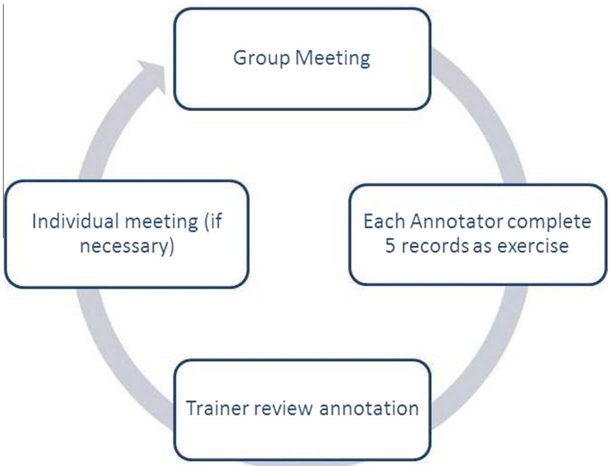
Group Meeting
Each Annotator complete 5 records as exercise
Trainer review annotation
Individual meeting (if necessary)

Fig. 4. Annotator training process.

ing. The adjudicators could edit or remove disputed annotations, or add new annotations, but could not edit or remove agreed annotations. The adjudicators participated in adjudication training before starting the task. Their training resembled annotation training. Average training time for one adjudicator was about 8 h. The average time for an adjudicator to complete the adjudication of one clinical record was about 50 min – not much less than the annotation time.

The long adjudication time is caused by the fact that in order to address the disagreements between the two annotations, the adjudicators have to do the temporal relation inference manually. The TLINK disagreements usually correspond to the more difficult and vague temporal relations in clinical narratives. Moreover, each addition, removal or modification of the problematic TLINK may cause a potential conflict with TLINKs that are already adjudicated. Hence, the adjudicator not only needs to more carefully examine the context of the temporal relation, but also needs to understand the thought processes in the two annotations to be able to address the differences.

Table 4
Average precision and recall on EVENT/TIMEX3 span compared against TimeBank.

	Exact match		Partial match	
	i2b2	TimeBank	i2b2	TimeBank
EVENT	.83	.78	.87	.81
TIMEX3	.73	.83	.89	.96

As an effort to reduce the manual inference work required in the adjudication process, we experimented with presenting to the adjudicator differences between the complete TLINKs transitive closures in the two annotations instead of the differences between the raw TLINKs. However, the transitive closure process drastically increased the number of disagreed TLINKs and made the adjudication process even more difficult. Another effort to improve the adjudication efficiency was to add the adjudicator-requested highlighting and indexing features in the relation adjudication tool, MAI[2] [33]. The highlighting helps the adjudicators to easily locate related entities for a given TLINK, while the indexing feature helps them to browse other relations that involve a given entity. The adjudicators reported that these features were very helpful. Looking forward, we believe that an adjudication interface with embedded temporal closure and TLINK conflict detection components will benefit future temporal annotation efforts.

4.6. Post-processing

As mentioned in Section 2.1, some TLINKs may conflict others. Although the adjudicators did a good job of removing most of the conflicting TLINKS (e.g. ‘A > B’ against ‘A = B’), we found that adjudicated annotations contained an average of 5.24 conflicting

[2] The author of the MAE/MAI toolkit [33], Amber Stubbs, kindly provided these requested features for this project.

Table 5
Accuracy of EVENT/TIMEX3 attribute agreement, compared against TimeBank.

EVENT	i2b2		TimeBank	TIMEX3	i2b2		TimeBank
	Accuracy	Kappa	Accuracy		Accuracy	Kappa	Accuracy
TYPE	0.93	0.9	0.77	TYPE	0.9	0.37	1
MODALITY	0.96	0.37	1	VAL	0.75	–	0.9
POLARITY	0.97	0.74	1	MOD	0.83	0.21	0.95

TLINKs per record, amounting to 2.98% of all TLINKs. We manually corrected these conflicting TLINKs in post processing.

5. Annotation quality

5.1. EVENT/TIMEX3

To assess the quality of the EVENT/TIMEX3 annotations, we computed the average precision and recall between two annotators by holding one annotation as key and the other as response. Since precision and recall are symmetric, it does not matter which annotation is held as key. We reported both 'exact span match' and 'partial span match' results. In 'exact span match', two annotations are considered a match only if the text spans agree exactly. In 'partial span match', two annotations are considered a match if their text spans overlap; this includes exact span match. We choose to report average precision and recall as IAA for entity spans instead of kappa score [34] in order to make our result comparable to TimeBank's IAA. As shown by Hripcsak and Rothschild [35], in cases where the null labels are ubiquitous, the kappa score is comparable to average precision and recall. For attributes, we report the percentage of agreed attributes in partially matched EVENTs/TIMEX3s, and the kappa scores. We notice that the kappa scores for EVENT Modality, EVENT Polarity, TIMEX3 Type and TIMEX3 Modifier attributes are low. The reason for this is that each of these attributes has a dominant attribute value, for example, the majority of EVENTs have the Modality 'Factual', which increases the by-chance agreement score and thus lowers the kappa scores. The TimeBank 1.2 documentation[3] reports similar inter-annotator agreement (IAA) measures. But the reported TimeBank agreement was computed over 10 documents annotated by two expert annotators, while our agreement is reported over the entire corpus. As shown in Tables 4 and 5 below, the IAA of our entire corpus by all eight annotators is comparable to TimeBank's IAA on ten documents between two expert annotators.

5.2. TLINKs

Each TLINK connects two extents and specifies the TYPE of the TLINK. An extent can be an EVENT or a TIMEX3. We evaluate TLINK extent agreement and TYPE agreement separately using the three methods that have been reported in previous literature [16,36,37]:

- comparing the raw TLINKs: The 'raw against raw' evaluation does not require the computation of temporal closure. However, since the annotators can assign TLINKs to any two extents, there are many different ways to annotate the exact same timeline. For example, if we have three extents A, B and C happening at the exact same time, we may choose any two and assign a 'SIMULTANEOUS' relation. This explains the low agreement score on raw against raw TLINK extent match (see column 'Raw–Raw' in Table 6). TimeBank uses this IAA method and reports a 0.55 extent agreement and 0.77 in TYPE agreement.

Table 6
TLINK inter-annotator agreement.

TLINK	Raw–raw	Closure – closure	Raw – closure
Extents (average precision and recall)	0.39	0.37	0.86
TYPE (accuracy)	0.79	0.72	0.73

- comparing the temporal closures generated from two TLINK annotations [36]. To account for the non-uniqueness of raw TLINK annotations, we also experimented with comparing the temporal closures of the two sets of TLINK annotations. The drawback of this method is its sensitivity to small changes in the annotation. In certain cases, this method heavily penalizes the agreement score because of a difference in just one TLINK between the annotations. Consider the following case: one of the raw TLINK annotations contains two sets of EVENTs such that the EVENTs within the same set are temporally related to each other, but there is no TLINKs between EVENTs from different sets. The other annotation is exact the same except that it contains an additional TLINK that links an EVENT of one set to an EVENT in the other set. The agreement score between the two annotations will be very low because the additional link in the second annotation may create transitive TLINKs between every pair of EVENTS between the two sets. The 'Closure – Closure' column in Table 6 exhibits the evaluation score using this method.
- comparing the raw TLINK annotations against the temporal closure of the other annotation [37]: The 'raw against closure' evaluation computes the percentage of raw TLINKs in one annotation against the temporal closure of the other. It overcomes the drawbacks of the previous methods. The result of 'raw – closure' method is shown in the last column in Table 6.

Even though the overall TYPE accuracy looks acceptable, we noticed that the score is heavily influenced the dominating TLINK TYPEs. BEFORE_OVERLAP, DURING, BEGUN_BY and ENDED_BY TLINKs account for about 20% of the raw TLINKs, and only about 4% of the temporal closure. The accuracy for these TLINK TYPEs is much worse than those for the dominating TLINK TYPES (BEFORE/AFTER and OVERLAP/SIMULTANEOUS). Table 7 shows the raw against closure score breakdowns for each TLINK TYPE. Table 8 shows the TLINK confusion matrix and indicates that the minority TLINK TYPEs caused much confusion between annotators. Thus, in the i2b2 temporal relations corpus, we collapsed the 8 TLINK TYPEs into 3 major TLINK TYPEs:

- BEFORE: The original BEFORE, BEFORE_OVERLAP and ENDED_BY relations were merged as BEFORE relations.
- AFTER: The original AFTER and BEGUN_BY relations were merged as AFTER relations.
- OVERLAP: The original OVERLAP, SIMULTANEOUS and DURING relations were merged as OVERLAP relations.

The TLINK TYPE agreement (raw against closure) by merged TYPEs is shown in Table 9. The overall TLINK agreement increased

[3] http://timeml.org/site/timebank/documentation-1.2.html#iaa.

Table 7
TLINK accuracy score TYPE breakdown (before merging).

	Before/after	Overlap/simultaneous	During	Begun_by	Before_overlap	Ended_by	Overall
Accuracy	0.85	0.78	0.3	0.23	0.1	0.34	0.73

Table 8
TLINK confusion matrix.

	Before	After	Overlap/simultaneous	Before_overlap	During	Begun	Ended
BEFORE	7744	91	91	350	1	1	10
AFTER	130	261	51	11	4	15	2
OVERLAP/SIMULTANEOUS	963	205	6159	277	56	108	60
BEFORE_OVERLAP	918	39	312	398	33	3	2
DURING	1	21	244	6	103	5	3
BEGUN_BY	3	52	158	8	5	74	0
ENDED_BY	27	5	38	10	1	2	75

Table 9
TLINK accuracy score TYPE breakdown (after merging).

	Before/after	Overlap/simultaneous	Overall
Accuracy	0.86	0.81	0.84

from 0.73 before the merge to 0.84 after the merge. However, the merging process inevitably created conflicting TLINKs in the gold standard. For example, given EVENTs A, B and C with the original TLINKs ‘A DURING B’, ‘C DURING B’ and ‘A BEFORE C’, after merging, the DURING relations become SIMULTANEOUS, and thus creating conflicting TLINKs. Fortunately, such conflicts are infrequent. There are on average 6.5 such TLINKs in each document (amounting to 3.6% of the total number of raw TLINKs, or 0.55% of the TLINK closure). Since most of the machine learning systems train on the TLINK closure, the number of conflicting TLINKs can be considered negligible. These conflicts are an inevitable result of merging different TLINK types. One of the ways to obtain a non-conflicting TLINK corpus using the present annotation would be to use the un-merged raw annotation, and address each conflicting TLINK case by case as during the merging process.

6. i2b2 Temporal relations corpus

The i2b2 temporal relations corpus consists of 310 discharge summaries of more than 178,000 tokens. The annotated corpus includes both merged and unmerged TLINK annotations and can be obtained from https://www.i2b2.org/NLP. On average, each discharge summary in the corpus contains 86.6 EVENTs, 12.4 TIMEX3s and 176 raw TLINKs (1145.8 TLINKs in the temporal closure). The EVENT, TIMEX3, TLINK (before temporal closure) TYPE distributions are shown in Table 10. The TLINK TYPE distribution in temporal closures is shown in Table 11.

Table 11
TLINK TYPE distribution in temporal closures.

Before/after	Overlap/simultaneous	During	Before_overlap	Begun_by	Ended_by
80.9%	14.7%	0.7%	0.4%	1.6%	1.7%

7. Conclusions

i2b2 created a temporally annotated discharge summary corpus that is accessible by the research community. The i2b2 temporal relations corpus provides a rich resource for temporal reasoning study in the clinical domain. The annotation quality of this corpus is on par with stable and proven temporal annotation corpora in the general domain. The temporal reasoning systems that perform well on this corpus can potentially support time-related downstream clinical applications on narrative discharge summaries, such as time-specific question answering, medication reconciliation, and computer assisted coding.

We identified several challenges in temporal annotation, including: the handling of TLINK conflicts in annotation time; the TLINK closure representation in adjudication and the trade-off between the administrative overhead in multi-pass annotation and the quality of the single-pass annotation. We believe that addressing these issues will help increase annotation efficiency and accuracy in future temporal annotation tasks.

8. Funding

This project was supported by Informatics for Integrating Biology and the Bedside (i2b2) award No. 2U54LM008748 from the National Institutes of Health (NIH)/National Q5 Library of Medicine (NLM), by the National Heart, Lung, and Blood Institute (NHLBI), and by award No. 1R13LM01141101 from the NIH NLM. The con-

Table 10
Annotation TYPE distribution.

Events		TIMEX3		TLINK (before TC)	
OCCURRENCE	17.9%	DATE	70.5%	BEFORE/AFTER	13.0%
EVIDENTIAL	4.1%	TIME	2.7%	OVERLAP/SIMULTANEOUS	66.6%
TEST	16.4%	DURATION	16.7%	DURING	4.5%
PROBLEM	32.4%	FREQUENCY	10.1%	BEFORE_OVERLAP	9.0%
TREATMENT	24.4%			BEGUN_BY	3.7%
CLINICAL DEPT	4.9%			ENDED_BY	2.7%

tent is solely the responsibility of the authors and does not necessarily represent the official views of the NLM, NHLBI, or NIH.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2013.07.004.

References

[1] Meystre S, Savova G, Kipper-Schuler K, Hurdle J, et al. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform 2008;35:128–44.

[2] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. Journal of the American Medical Informatics Association 2013.

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

[4] Stacey Michael, McGregor Carolyn. Temporal abstraction in intelligent clinical data analysis: a survey. Artif Intell Med 2007;39(1):1–24.

[5] Denny Joshua C, Peterson Josh F, Choma Neesha N, Hua Xu, Miller Randolph A, Bastarache Lisa, et al. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. J Am Med Inform Assoc 2010;17(4):383–8.

[6] Liu Mei, Jiang Min, Kawai Vivian K, Stein Charles M, Roden Dan M, Denny Joshua C, et al. Modeling drug exposure data in electronic medical records: an application to warfarin. In: AMIA annual symposium proceedings, volume 2011. American Medical Informatics Association; 2011. p. 815.

[7] Irvine Ann K, Haas Stephanie W, Sullivan Tessa. Tn-ties: A system for extracting temporal information from emergency department triage notes. In: AMIA Annual symposium proceedings, volume 2008. American Medical Informatics Association; 2008. p. 328.

[8] Zhou Li, Parsons Simon, Hripcsak George. The evaluation of a temporal reasoning system in processing clinical discharge summaries. J Am Med Inform Assoc 2008;15(1):99–106.

[9] Clark Kimberly K, Sharma Deepak K, Chute Christopher G, Tao Cui. Application of a temporal reasoning framework tool in analysis of medical device adverse events. In: AMIA annual symposium proceedings, volume 2011. American Medical Informatics Association; 2011. p. 1366.

[10] Li M, Patrick J. Extracting temporal information from electronic patient records. AMIA Annu Sympos Proc 2012;2012:542.

[11] Pustejovsky J, Castano J, Ingria R, Sauri R, Gaizauskas R, Setzer A, et al. Timeml: Robust specification of event and temporal expressions in text. New Direct Question Answer 2003;3:28–34.

[12] Augusto JC. The logical approach to temporal reasoning. Artif Intell Rev 2001;16(4):301–33.

[13] Mani I, Pustejovsky J, Gaizauskas R. The language of time: a reader. Oxford University Press; 2005.

[14] Allen JF. An interval-based representation of temporal knowledge. In: Proc. 7th international joint conference on artificial intelligence. Canada: Vancouver; 1981. p. 221–6.

[15] Mani I, Pustejovsky J, Sundheim B. Introduction to the special issue on temporal information processing. ACM Trans Asian Lang Inform Process (TALIP) 2004;3(1):1–10.

[16] Pustejovsky J, Hanks P, Sauri R, See A, Gaizauskas R, Setzer A et al. The TimeBank corpus. In: Corpus linguistics, volume 2003; 2003. p. 40.

[17] Verhagen M, Gaizauskas R, Schilder F, Hepple M, Katz G, Pustejovsky J. Semeval-2007 task 15: Tempeval temporal relation identification. In: Proceedings of the 4th international workshop on semantic evaluations; 2007. p. 75–80.

[18] Verhagen M, Sauri R, Caselli T, Pustejovsky J. Semeval-2010 task 13: Tempeval-2. In: Proceedings of the 5th international workshop on semantic evaluation. Association for, Computational Linguistics; 2010. p. 57–62.

[19] UzZaman N, Llorens H, Allen J, Derczynski L, Verhagen M, Pustejovsky J. Tempeval-3: Evaluating events, time expressions, and temporal relations. arXiv, preprint arXiv:1206.5333; 2012.

[20] Roberts A, Gaizauskas R, Hepple M, Demetriou G, Guo Y, Setzer A et al. Semantic annotation of clinical text: the clef corpus. In: Proceedings of building and evaluating resources for biomedical text mining: workshop at LREC; 2008.

[21] Galescu L, Blaylock N. A corpus of clinical narratives annotated with temporal information. In: Proceedings of the 2nd ACM SIGHIT symposium on international health informatics. ACM; 2012. p. 715–20

[22] Castaño J, Gaizauskas R, Ingria B, Katz G, Knippen B, Littman J et al. 1.2.1 a formal specification language for events and temporal expressions; 2005. p. 10.

[23] UzZaman N, Allen JF. Trips and trios system for tempeval-2: Extracting temporal information from text. In: Proceedings of the 5th international workshop on semantic evaluation. Association for, Computational Linguistics; 2010. p. 276–83

[24] Llorens H, Saquete E, Navarro B. Tipsem (English and Spanish): evaluating crfs and semantic roles in tempeval-2. In: Proceedings of the 5th international workshop on semantic, evaluation; 2010. p. 284–91.

[25] Graff D, Cieri C. English Gigaword. Philadelphia: Linguistic Data Consortium; 2003.

[26] Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc 2010;17(5):514–8.

[27] Savova G, Bethard S, Styler W, Martin J, Palmer M, Masanz et al. Towards temporal relation discovery from the clinical narrative. In: AMIA annual symposium proceedings, volume 2009. American Medical Informatics Association; 2009. p. 568.

[28] Jordan PW, Mowery DL, Wiebe J, Chapman WW. Annotating conditions in clinical narratives to support temporal classification. Proc Am Med Inform Assoc Sympos 2010;2010:1005.

[29] Bramsen P, Deshpande P, Lee YK, Barzilay R. Finding temporal order in discharge summaries. In: AMIA annual symposium proceedings, volume 2006. American Medical Informatics Association; 2006. p. 81.

[30] Uzuner O, South B, Shen S, DuVall S. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.

[31] Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South B. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc 2012;19(5):786–91.

[32] Temporal histories of your medical events, thyme.

[33] Stubbs A. Mae and mai: lightweight annotation and adjudication tools. In: Proceedings of the 5th linguistic annotation workshop. Association for, Computational Linguistics; 2011. p. 129–33.

[34] Fleiss Joseph L, Levin Bruce, Paik Myunghee Cho. Statistical methods for rates and proportions. John Wiley & Sons; 2013.

[35] Hripcsak George, Rothschild Adam S. Agreement, the f-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12(3):296–8.

[36] Setzer A, Gaizauskas R, Hepple M. Using semantic inferences for temporal annotation comparison. In: Proceedings of the fourth international workshop on inference in computational semantics (ICOS-4); 2003. p. 25–6

[37] UzZaman N, Allen J. Temporal evaluation. In: Proceedings of the 49th annual meeting of the association for computational linguistics. Human Language Technologies; 2011. p. 351–6.