Normalization of relative and incomplete temporal expressions in clinical narratives

RECEIVED 30 June 2014
REVISED 12 October 2014
ACCEPTED 20 October 2014
PUBLISHED ONLINE FIRST 13 April 2015

Weiyi Sun[1], Anna Rumshisky[2], Ozlem Uzuner[3]

ABSTRACT

Objective To improve the normalization of relative and incomplete temporal expressions (RI-TIMEXes) in clinical narratives.

Methods We analyzed the RI-TIMEXes in temporally annotated corpora and propose two hypotheses regarding the normalization of RI-TIMEXes in the clinical narrative domain: the anchor point hypothesis and the anchor relation hypothesis. We annotated the RI-TIMEXes in three corpora to study the characteristics of RI-TMEXes in different domains. This informed the design of our RI-TIMEX normalization system for the clinical domain, which consists of an anchor point classifier, an anchor relation classifier, and a rule-based RI-TIMEX text span parser. We experimented with different feature sets and performed an error analysis for each system component.

Results The annotation confirmed the hypotheses that we can simplify the RI-TIMEXes normalization task using two multi-label classifiers. Our system achieves anchor point classification, anchor relation classification, and rule-based parsing accuracy of 74.68%, 87.71%, and 57.2% (82.09% under relaxed matching criteria), respectively, on the held-out test set of the 2012 i2b2 temporal relation challenge.

Discussion Experiments with feature sets reveal some interesting findings, such as: the verbal tense feature does not inform the anchor relation classification in clinical narratives as much as the tokens near the RI-TIMEX. Error analysis showed that underrepresented anchor point and anchor relation classes are difficult to detect.

Conclusions We formulate the RI-TIMEX normalization problem as a pair of multi-label classification problems. Considering only RI-TIMEX extraction and normalization, the system achieves statistically significant improvement over the RI-TIMEX results of the best systems in the 2012 i2b2 challenge.

Keywords: temporal reasoning, medical language processing, temporal expression normalization

BACKGROUND AND SIGNIFICANCE

Temporal expressions (TIMEXes) are words or phrases that carry information about time. For example, the phrase “last Friday” in “the patient tripped and fell last Friday” is a TIMEX that indicates when the incident (the fall) occurred. Narrative texts rely on TIMEXes to present the timeline of a story. In clinical narratives, TIMEXes convey important clinical information, such as the duration of symptoms and the frequency of medication administration. Hence, understanding TIMEXes is a crucial part of any natural language processing task that deals with the temporal dimension.

TIMEXes come in various forms. For example, to express the same concept, one can say “Nov 29, 2013,” “last Friday,” “the 29th,” “the day after Thanksgiving,” or “Jenny’s birthday.” To understand such TIMEXes, natural language processing systems need to not only identify the TIMEXes, but also normalize all the various forms of expressing “Nov 29, 2013” into a canonical format.

In many cases, TIMEX text spans provide sufficient information for normalization (eg, “Nov 29, 2013” almost always means the same date in any context). However, this is not the case for two types of TIMEXes: relative and incomplete TIMEXes (hereafter collectively referred to as “**RI-TIMEXes**”). **Relative TIMEXes** are phrases whose temporal meanings are stated as relative values against other time points (eg, “two days before the fall”). **Incomplete TIMEXes** refer to TIMEXes that contain partial information toward their normalized value. For example, the TIMEX in “the lab result at 6 am” states the time, but we need to refer to the context to determine the TIMEX’s calendar date.

RI-TIMEXes are abundant in narrative texts – on average, they constitute 26% of TIMEXes in clinical narratives, 32% of TIMEXes in historical narratives, and 54% of TIMEXes in newswire articles (see Section “RI-TIMEX Annotation” for details). Normalizing RI-TIMEXes is a challenging task. In the 2012 i2b2 temporal relation challenge,[1] a shared-task challenge on clinical narratives with a TIMEX extraction and normalization track, the top 10 systems achieved an average normalization accuracy of 0.32 in RI-TIMEXes, in comparison to the overall TIMEX normalization accuracy of 0.67.

RI-TIMEX normalization requires two pieces of reference information, apart from the RI-TIMEX’s text span: an anchor point and an anchor relation. The **anchor point** is the TIMEX that the RI-TIMEX refers to. The **anchor relation** is the temporal relation between the RI-TIMEX and its anchor point that shows how the RI-TIMEX relates to the anchor point in the narrative timeline. We propose that the anchor point and anchor relation of RI-TIMEXes can be[2] formulated as two multi-label classification problems, which differentiates our method from the state-of-the-art rule-based approaches. The RI-TIMEX normalization system we present herein is fully informed by the context surrounding the RI-TIMEXes, and, thus, we expect it to perform better in the clinical narrative domain than existing methods.

EXISTING WORK

TIMEX standards

There are various standards in the general domain for representing TIMEXes.[3] The most frequently used standard, TIMEX3 (adopted in temporal specification languages such as TimeML,[4]) defines four types of TIMEXes: date, time, duration, and set (or frequency). It also normalizes the TIMEX’s value to ISO8601 format. The TIMEX3 standard uses temporal function to indicate RI-TIMEXes and marks the TIMEX IDs of their anchor points. For example, “two weeks from June 7, 2003” contains two TIMEXes: the duration “two weeks” and the date

Correspondence to Weiyi Sun, 1400 Washington Ave, Draper 114B, Albany, NY 12222, USA; wsun2@albany.edu; Tel: 860-534-1416

For numbered affiliations see end of article.

"June 7, 2003"; the temporal function here is that the duration "two weeks" starts on the date "June 7, 2003."

In the clinical domain, independent from the TimeML scheme, there are several temporal representation schemes tailored for clinical narratives, including studies based on temporal constraint model,[5,6] and OWL-based time ontology.[7] We refer readers to Sun et al.'s[3] survey paper for more details. The i2b2 temporal challenge developed a temporally annotated clinical corpus using the TIMEX3 standard.[2] To simplify annotation, i2b2 removed the temporal function attribute in TIMEX3 and, instead, used temporal relations to indicate the anchor points of relative TIMEXes. Another major change in the i2b2 version of TIMEX3 annotation is that, instead of using a single document creation time (DCT) for each document, i2b2 assigns a section creation date (SECTIME) for each section. In the i2b2 corpus, each discharge summary contains a clinical history section and a hospital course section. The SECTIMEs for clinical history sections are the admission dates, and the SECTIMEs for hospital course sections are the discharge dates.

TIMEX-annotated datasets

Temporally annotated corpora are available in newswire, historical narrative, and clinical domains. Our study used one widely-used representative corpus from each of these domains. In the newswire domain, we studied the widely adopted TimeBank dataset, consisting of 183 newswire articles, 63K tokens, and 1423 TIMEXes.[8] In the historical narrative domain, we studied the WikiWars corpus, consisting of 22 Wikipedia articles, 119K tokens, and 2681 TIMEXes.[9] In the clinical domain, we studied the i2b2 temporal challenge dataset, consisting of 310 discharge summaries, 178K tokens, and 3844 TIMEXes.[2]

TIMEX learning

Existing works show that TIMEX (especially RI-TIMEX) normalization is a challenging task. In both the TempEval challenge[10–12] and the i2b2 temporal reasoning challenge,[1] the best performing systems achieved an F1-measure in the 90s on TIMEX text span identification (TIMEX extraction), which approximated or exceeded the inter-annotator agreements in these corpora. In contrast, in the TIMEX normalization task in the same two challenges, the accuracy of the best performing systems remained in the mid-80s and lower 70s, much lower than the inter-annotator agreement in the same task in the respective corpora. The challenges' participants recognize that TIMEX normalization is a difficult task.[13] In particular, the authors of the best performing TempEval system, HeidelTime, concluded, in their error analysis for TempEval 3, that "wrongly detected reference times or relations" were one of "the main sources for incorrect value normalization of underspecified expressions."[14]

In the newswire domain, the common strategy for resolving RI-TIMEXes is to anchor every RI-TIMEX to the DCT and use verbal tense and/or lemma indicators (eg, "ago," "prior") to determine the anchor relation.[13–17] Few temporal information extraction works exist on general corpora, other than newswire. Strotgen et al.[18] analyzed the difference between RI-TIMEX anchor points in four domains: newswire, historical narratives, colloquial (a corpus of text messages), and scientific text. Their strategy was to anchor all RI-TIMEXes to DCT in the news, colloquial, and scientific corpora. In historical narratives, they anchored all RI-TIMEXes to previous TIMEXes. A **previous TIMEX** is a TIMEX that appears immediately before the RI-TIMEX in the narrative text. For instance, in the sentence "on December 7, 1941, the Japanese attacked Pearl Harbor, and the following day the United States declared war on Japan," the previous TIMEX of the RI-TIMEX, "the following day" is "December 7, 1941." To determine anchor relations, Strotgen et al. used tense as an indicator in news, colloquial, and scientific text (eg, the authors assigned the relation "before" for RI-TIMEXes in past-tense sentences and the relation "after" for present-tense and future-tense sentences) and always assigned "after" anchor relations for RI-TIMEXes in historical narratives. This strategy achieved value f-measures of 73.8 for the newswire domain, 74.5 for the historical narrative domain, 91.9 for the colloquial domain, and 64.7 for the scientific domain. However, we show in Section "RE-TIMEX Annotation" that these strategies are not suitable for the clinical domain.

In the clinical domain, the participants in the i2b2 challenge mainly deal with RI-TIMEXes in two ways. In some systems, the participants assign SECTIMEs as anchor points to RI-TIMEXes.[19,20] For instance, one would anchor an RI-TIMEX to the admission date if words such as "admission" and "operation" appear near the RI-TIMEX. In other systems, the participants define a list of keywords, such as "operation" and "date of birth." When any of the keywords appears in the text, the systems associate it with a nearby TIMEX. For a RI-TIMEX containing certain signal phrases (eg, "post-operative day #6" and "day of life #5"), the system uses the TIMEX associated with the relevant keyword as the anchor point of the RI-TIMEX.[21–24] These approaches seem intuitive, but they cannot handle relative TIMEXes that do not contain any signal phrases (eg, "two days later") or incomplete TIMEXes. These approaches are difficult to generalize. Indeed, the challenge results shows that RI-TIMEX normalization is an unsolved problem in the clinical domain.[1]

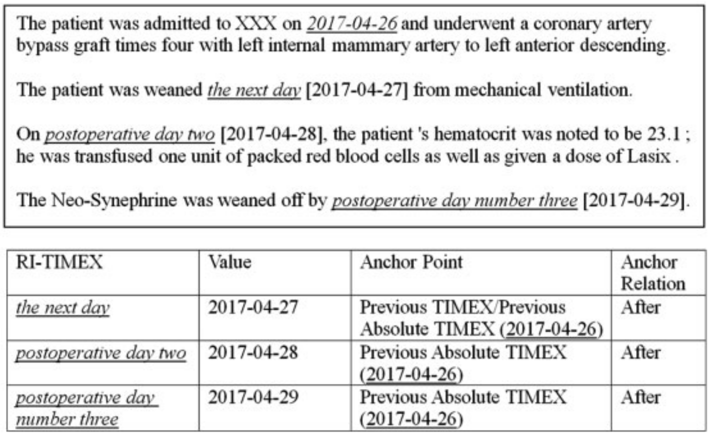
RI-TIMEX ANNOTATION

Assumptions

Finding the anchor points and determining the anchor relations of RI-TIMEXes are critical steps toward RI-TIMEX normalization. To understand and compare the characteristics of the anchor points and anchor relations of RI-TIMEXes in different narrative domains, we annotated RI-TIMEXes in three widely-used temporally annotated corpora: TimeBank, WikiWars, and the i2b2 temporal corpus. We designed our annotation guidelines based on the following assumptions for RI-TIMEX anchor points and anchor relations.

Anchor point assumptions: As discussed in the previous section, the existing methods for finding anchor points are limited to 1) assuming a default anchor point or 2) looking for RI-TIMEXes that contain certain signal phrases. The first method builds on the assumption that, when a writer narrates, he or she tends to follow the timeline of and speak in regards to the time when the writing occurs (ie, DCT). For narratives spanning a longer period, a writer tends to follow the timeline established by the previous TIMEXes. Only in rare cases does a writer move back and forth in the narrative timeline without explicitly stating the new time point, because this risks losing the reader. The second method suggests that, in clinical narratives, there are cases when a writer would use some significant clinical event as a temporal anchor point and keep referring to it. In the above "post-operative day #6" example, the anchor point is fixed to the operative day, even though other postoperative days may have been mentioned between the text that specifies "operation day" and the RI-TIMEX "post-operative day #6." Inspired by the implicit assumptions underlying the two existing methods of RI-TIMEX normalization, we propose the following hypothesis regarding RI-TIMEX anchor points: A RI-TIMEX in narrative text usually anchors to one of the following TIMEXes –DCT or SECTIME; previous TIMEX; or previous absolute TIMEX. A previous absolute TIMEX refers to the non-RI-TIMEX that appears immediately before the RI-TIMEX. Figure 1 shows a snippet of a de-identified discharge summary that serves as an illustration of our hypothesis.

Figure 1: RI-TIMEX Anchoring Example.



The patient was admitted to XXX on *2017-04-26* and underwent a coronary artery bypass graft times four with left internal mammary artery to left anterior descending.

The patient was weaned *the next day* [2017-04-27] from mechanical ventilation.

On *postoperative day two* [2017-04-28], the patient 's hematocrit was noted to be 23.1 ; he was transfused one unit of packed red blood cells as well as given a dose of Lasix .

The Neo-Synephrine was weaned off by *postoperative day number three* [2017-04-29].

RI-TIMEX	Value	Anchor Point	Anchor Relation
the next day	2017-04-27	Previous TIMEX/Previous Absolute TIMEX (2017-04-26)	After
postoperative day two	2017-04-28	Previous Absolute TIMEX (2017-04-26)	After
postoperative day number three	2017-04-29	Previous Absolute TIMEX (2017-04-26)	After

The TIMEXes in the text are shown in italic and underlined. The RI-TIMEX "the next day" is relative to the TIMEX that appears prior to it, the previous TIMEX "2017-04-26," and, thus, the value of the RI-TIMEX is "2017-04-27." In this case, since the previous TIMEX also happen to be an absolute TIMEX, the RI-TIMEX "the next day" is anchored to both the previous TIMEX and the previous absolute TIMEX. For the RI-TIMEX "postoperative day two," the previous TIMEX is "the next day," and the previous absolute TIMEX is "2017-04-26." The anchor point for this RI-TIMEX is the previous absolute TIMEX. This example shows that our previous absolute TIMEX hypothesis is based on the assumption that the time stamps of significant events in the timeline are usually explicitly stated as absolute TIMEXes, which later RI-TIMEXes often refer back to.

Anchor relation assumptions: When we model time in a continuous, linear fashion, we may view TIMEXes as temporal intervals (a period defined by two instantaneous time points on a timeline).[25] Therefore, there can be 13 possible temporal relations between two TIMEXes intervals. We propose that, for the purpose of date and time RI-TIMEX resolution, it is sufficient to treat TIMEXes as time points. Thus, we can assume that the anchor relation between a RI-TIMEX and its anchor point can be one of the following: before, after, or equal.

Annotation

We annotated the anchor points and anchor relations of RI-TIMEXes in three domains. Although the focus of this paper is RI-TIMEX normalization in the clinical domain, we also briefly describe our annotation results in the newswire and historical narrative corpora, to serve as a comparison to the clinical domain data. The comparison can illuminate the shared and unique characteristics of RI-TIMEXes among these domains.

We extracted RI-TIMEXes by filtering known formats of absolute TIMEXes (eg, mm/dd/yy format) from all annotated TIMEXes and manually reviewed whether or not the remaining TIMEXes are RI-TIMEXes. For newswire data, we annotated the TimeBank corpus.[8,12] Among the 1221 "date" or "time" type TIMEXes, 54% were RI-TIMEXes. For the historical narrative domain, we looked at the WikiWars corpus.[9] Among the 2387 "date" or "time" type TIMEXes, we found 861 (36%) RI-TIMEXes. For clinical narratives, we examined the i2b2 challenge corpus. The corpus contains 310 discharge summaries, split into a training set (190 documents) and a test set (120 documents). There are 1712 and 1282 "date" or "time" type TIMEXes in the training and test sets, respectively. We found 624 (36%) RI-TIMEXes in the training set and 481 RI-TIMEXes in the test set.

In the annotation process, we conducted single-pass annotation on the newswire, historical narratives, and i2b2 training set. To evaluate annotation quality, 50% of the i2b2 testing set was dual-annotated. We present our annotation results and inter-annotator agreement in Section "Results And Discussion".

METHODS

Our RI-TIMEX normalization strategy is to learn the RI-TIMEX anchor point and anchor relation using multi-label classifiers and to combine the anchor point and relation with the information extracted from the RI-TIMEX text span to form the final value of the RI-TIMEX. For instance, to normalize the RI-TIMEX "post-operative day # six," we first learned that the anchor point, the date of the operation, was September 16, 2006; the anchor relation is "after"; and, finally, the normalizing value is 6. We can then add 6 days to the operation date to obtain the final value for the RI-TIMEX, September 9, 2007.

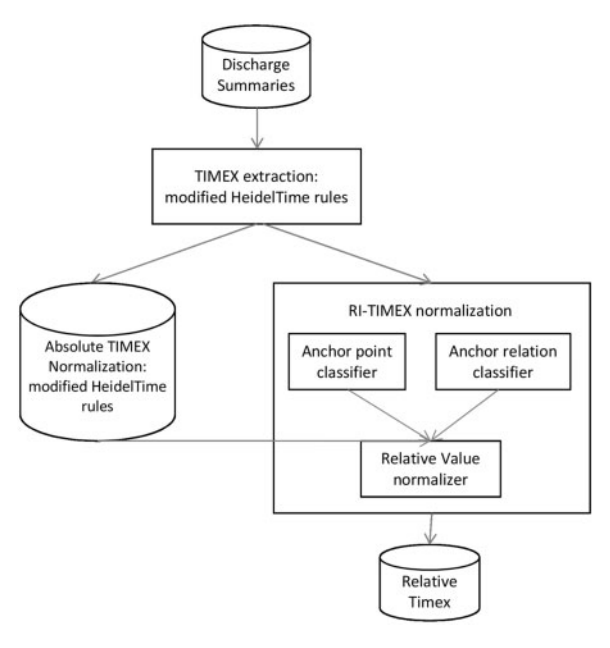
System structure

The structure of our proposed TIMEX extraction and normalization system is shown in Figure 2:

Since the TIMEX extraction and the normalization of absolute TIMEXes are not the main focus of this paper, we used the existing HeidelTime rule-based TIMEX extraction and normalization system for these subtasks.[14] The HeidelTime system was developed for the general domain. The out-of-the-box rules in the HeidelTime system are not suitable for the clinical domain. Without tuning, the TIMEX extraction F1-measure of the i2b2 corpus (10-fold cross validation) is 72, and the value accuracy of absolute TIMEXes is 39%. The value accuracy is the percentage of correctly normalized TIMEXes in all correctly extracted TIMEXes. We adapted the extraction and normalization rules using the training set of the i2b2 corpus. After tuning, the extraction F1-measure (10-fold cross validation) is 92, and the absolute TIMEXes value accuracy is 74%.

After extracting the TIMEXes and normalizing the absolute TIMEXes, we processed the RI-TIMEXes using the following steps:

Figure 2: Structure of the TIMEX Extraction and Normalization System.


Discharge Summaries
TIMEX extraction: modified HeidelTime rules
Absolute TIMEX Normalization: modified HeidelTime rules
RI-TIMEX normalization
Anchor point classifier
Anchor relation classifier
Relative Value normalizer
Relative Timex

- **Anchor point classifier.** We trained four binary SVM classifiers to learn each anchor point type: admission date, discharge date, previous TIMEX, and previous absolute TIMEX. When none of the classifiers returned positive classifications, we used admission date as the default anchor point. When the classifiers returned conflicting classifications, we selected the class label based on its prevalence in the training set, in the following order of preference: admission date, previous TIMEX, previous absolute TIMEX, discharge date. We used the LibSVM implementation of SVM for this classifier.[26]
- **Anchor relation classifier.** Similarly, we treated the anchor relation problem as another multi-label classification task, with the labels "before," "after," and "equal/during," using LibSVM.
- **Value normalization.** The last piece of information required to decipher the value of the RI-TIMEX is the meaning conveyed in the text span of the RI-TIMEX. We composed a set of rules to parse the RI-TIMEX spans. The rules parse the numbers (in both digit forms and word forms) and the expressions of units such as weeks, days, and months.
- **Integration.** This step combines the outputs of the above components to generate the final value of the RI-TIMEX. When a RI-TIMEX's anchor point is another RI-TIMEX, the integration step runs recursively to find the final value of the target RI-TIMEX.

In short, the system input is the discharge summary texts and the system output is marked up TIMEXes, with normalized value in the i2b2 xml format.

Feature sets

We experimented with several feature sets in the temporal anchor point and anchor relation classifiers. We used chi-square attribute selection in our experiments. To determine the chi-square threshold value, we performed 10-fold cross-validation on the training set using all the features. We chose to use 7.88 as the threshold for anchor point classification and 9.58 as the threshold for anchor relation classification. The feature sets are:

1. Bag of words in an N-token window before and after the RI-TIMEX, as well as the RI-TIMEX text span itself. Ten-fold cross-evaluation results indicate that, for this dataset, choosing N to be any number from 6 to 10 tokens yields statistically significantly improved results. We report the results with $N = 8$. The tokens are case normalized. Both unigrams and bigrams were included.
2. Same bag of words features as (A), with all numbers normalized, both in digit formats (eg, "2," "2nd," "#2") and in spelled-out word formats (eg, "two," "second"), to a unified number token. In other words, we did not distinguish any numbers from one another – "2" and "3" are considered to be the same token.
3. EVENT features. In i2b2 data, any text span that indicates a clinically significant event was annotated with an EVENT tag. There are six types of EVENTs: PROBLEM, TREATMENT, TEST, CLINICAL_DEPT, OCCURRENCE, and EVIDENTIAL.[2] Some RI-TIMEXes are closely related to EVENTs (eg, "the day of admission" where "admission" is an i2b2 EVENT). So, we included features of the EVENTs that exist in the same clause of a sentence as the RI-TIMEX. Both EVENT type and normalized EVENT text span were included.
4. Previous TIMEX features. The annotation showed that many RI-TIMEXes are anchored to the previous TIMEX or previous absolute TIMEX. It is natural to add them as features to inform the prediction of anchor points and relations. We further tested the following feature representations:

   a. TYPE attribute (DATE, TIME, DURATION, and FREQUENCY) of only the previous TIMEX.
   b. The bag of words of the previous TIMEX text span.
   c. The bag of words of the previous DATE- or TIME-type TIMEX text span.
   d. The bag of words of the previous absolute DATE- or TIME-type TIMEX text span.
   e. A flag indicating whether the previous TIMEX is the section time "admission date," the section time "discharge date," or neither.

5. Tense information of the sentence. Sentence tense may inform anchor relation classification. We used the verbal tense of the main verbs as features in our classifiers, using Stanford Parser.[27]

RESULTS AND DISCUSSION

Annotation results

Table 1 shows the distribution of anchor points and relations in the three domains. A RI-TIMEX may simultaneously anchor more than one anchor point, when the anchor points happen to have the same value. Hence, the percentages of anchor point distributions for each domain may not add up to 100%. We found that in TimeBank, 96% of all RI-TIMEXes anchor to the DCTs. In the WikiWars corpus, we observed 93% of the RI-TIMEXes anchoring to previous TIMEXes. In contrast, we found that the anchor points of RI-TIMEXes in the training set of the 2012 i2b2 corpus were more varied. The distribution indicates that the approaches adopted in the existing works will work well in the newswire and historical narrative domains; that is, by using DCT and previously mentioned TIMEXes as default anchor points, the systems can correctly anchor the majority of the RI-TIMEXes.[18] However, such a strategy will not work on clinical narratives data, because, by assuming the most frequent label in that data (ie, admission date), this strategy can, at most, correctly anchor 59% of the RI-TIMEXes (41%, if assuming section time, because some of the

Table 1: RI-TIMEXes Annotation Statistics

	Types	TimeBank (%)	WikiWars (%)	i2b2 (Training) (%)		i2b2 (Test) (%)
Anchor points annotation statistics	DCT	96	1	Admission	53	59
				Discharge	16	19
	Previous TIMEX	2	93		37	34
	Previous absolute TIMEX	1	2		35	28
	Not in the above types	1	4		5	2
Anchor relations annotation statistics	Before	38	12	11		12
	After	6	55	46		46
	Equal/During	55	30	41		41
	Not in the above types	1	3	2		1

admission date anchoring occurs in the hospital course section, and some of the discharge date anchoring occurs in the clinical history section; so, assuming section time would result in worse performance than assuming admission date). Interestingly, in all three datasets, at least 95% of all RI-TIMEXes were anchored to at least one of following time points: DCT (admission or discharge time), previous TIMEX, and previous absolute TIMEX. For the most part, the RI-TIMEXes not anchored to one of these time points are either: 1) annotation errors that assigned a wrong TIMEX value to the RI-TIMEX, which cannot be anchored to any anchor points; 2) instances in which the anchor point never appeared in the text (eg, “on Feb-8, 20 days after the accident,” in which the true anchor point of TIMEX, “20 days after,” should be the date of the accident, which never appeared in the text); (3) other, infrequent cases when the RI-TIMEXes is another time point mentioned in the text. This confirms our anchor point hypothesis.

In the anchor relation annotation, we found that all three corpora show variety in the temporal anchor relations. The two narrative domain corpora, WikiWars and i2b2 discharge summaries, share similar distribution of anchor relation types, featuring more “After” and “Equal/During” anchor relations, while the newswire corpus differs by showing more “Before” and “Equal/During” relations.

We calculated the inter-annotator agreement between two annotators in 50% of the test set. For anchor point annotation, the two annotators agreed on 94.7% of the RI-TIMEX incidences. For the anchor relation annotation, the annotators agreed on 98.4% of the RI-TIMEXes with clear anchor points. The annotation and guidelines will be available to the research community as a part of the i2b2 data repository.

Feature selection results

Table 2 shows the feature selection experiment results, which are 10-fold cross-validation results obtained from the i2b2 training set. We first included all the features, and report those result in the first row. We reported classification accuracies by label, which are comparable to the end-to-end evaluation results we present in the next sub-section. We chose to the use the feature set (B) + (D1) + (D2) in the final system, because its improvement over feature sets (A) and (B) in discharge date anchor point classification and anchor relation classifications was statistically significant in the randomization test. While the results using (B) + (D3) + (D4), (B) + (D3) + (D4) + (D5) and (B) + (D3) + (D4) + (E) were not statistically different from (B) + (D1) + (D2), we chose the (B) + (D1) + (D2) set (bolded in Table 2), due to its simplicity.

We found that feature set (C), EVENT features, doid not inform the classification of either the anchor point or the anchor relation. Our error analysis showed that most of the events related to the RI-TIMEX are already included in the N-token window of the RI-TIMEX. Events further away from the RI-TIMEX add more noise than discriminating information to the classification. We also found that feature set (E), tense features, did not help improve the performance of anchor relation prediction, because, in the discharge summary, the majority of verbal tenses are past tense, with the exception of a few future tense cases in sentences mentioning follow-up plans. Thus, in most cases, tense information is only indicative of anchor relations when the anchor point is the discharge date, which is an infrequent occurrence. Signal words, such as “next” and “prior,” in or near the RI-TIMEXes are usually more informative than tense information in determining the temporal anchor relation.

Our results also showed that the normalized bag of words, feature set (B), informed the classification of anchor points, but doid not help with anchor relation classification. Our error analysis showed that numbers in the RI-TIMEX span are sometimes useful in predicting anchoring relations. For example, in some clinical documents, providers refer to any time after an event, but still within 24 h of the event, as “day #1.” If a baby was born on the morning of February 8, February 8 was considered “day of life #1” and February 9 was considered “day of life #2.” In this case, even though both TIMEXes anchor to the same anchor point, the RI-TIMEX “day of life #1” is considered to have an “Equal/During” anchor relation with the anchor point, the date of the birth, while the RI-TIMEX “day of life #2” is considered to have an “After” anchor relation with the anchor point. Thus, normalizing all the numbers to a unified token adds noise to the anchor relation classifier, even though it also improves the anchor point classifier. The previous TIMEX feature appears to be helpful in both anchor point and anchor relation detection.

End-to-end results

In this section, we report the end-to-end evaluation results on the held-out test set of the 2012 i2b2 corpus, specifically on the features set (B) + (D1) + (D2). The last column of Table 3 shows the accuracy of each step of the system. For each step, the accuracy is computed using the RI-TIMEXes that are correctly extracted or classified in the previous step(s). For instance, an anchor point classifier accuracy of 74.68% means that the anchor point was correctly identified for 74.68% of the RI-TIMEXes from which the system correctly extracted

Table 2: 10-Fold Cross Validation in the Training Set Accuracy Using Different Feature Sets

Feature Sets	Anchor Points				Anchor Relation		
	Admission (330)	Discharge (101)	Previous TIMEX (228)	Previous Absolute TIMEX (221)	After (287)	Before (69)	Equal/During (256)
All features	79.01	91.67	67.53	75.00	92	72.9	89.3
(A)	74.52	90.71	65.54	76.44	91.6	75.7	89.7
(B)	76.76	91.19	64.74	77.88	93.4	71.4	84.6
(B) + (C)	75.80	88.62	62.34	76.44	83.6	74.3	90.5
(B) + (D1) + (D2)	**77.56**	**92.47**	**68.91**	**75.16**	**93.4**	**81.4**	**92.1**
(B) + (D3) + (D4)	78.53	91.99	68.91	75.64	91.3	77.1	90.1
(B) + (D3) + (D4) + (D5)	78.85	91.99	68.75	76.60	90.9	77.1	90.1
(B) + (D1) + (D2) + (E)	78.85	91.67	67.47	75.16	93.4	80.0	90.1

The number of instances for each class is shown in parentheses in the first row.

Table 3: Accuracy of RI-TIMEX Normalization at Each Step

	By Type	By Type Accuracy (%)
Extraction	Overall	82.12*
Anchor point	Overall	74.68
	Admission date (286)	91.56
	Discharge date (95)	50.67
	Previous Timex (166)	64.42
	Previous Absolute Timex (136)	77.78
Anchor relation	Overall	87.71
	Before (58)	68.97
	After (221)	90.06
	Equal/During (197)	90.29
Normalization	Overall	57.2

The number of cases in each class is shown in parentheses next to its class label. (The * value for TIMEX extraction is a recall measure result).

the text spans. Since improving TIMEX extraction was not our main focus in this work, we refer the readers to the 2012 i2b2 participants' papers for a more detailed description of fine-tuning and adapting HeidelTime rules for better TIMEX extraction.[19,28]

Table 3 also shows the break-down statistics of the anchor point and anchor relation type for correctly extracted RI-TIMEXes. The anchor point classifier performed similarly on the training and test set for the previous TIMEX and previous absolute TIMEX types. However, the performance on admission date and discharge date anchor points in the test data were quite different. Our strategy of preferring the admission date label, in the cases of missing labels, and discriminating against discharge date labels in cases of conflict labels increased the accuracy of admission date anchor points and brought down the accuracy of discharge date anchor points.

Overall, the trend of anchor relation and anchor point accuracy in the end-to-end test set was consistent with that in the training set cross-validation. Recall that 46% and 41% of all RI-TIMEX examples in the training set are "after" and "equal/during" relations, and only 11% are "before" relations. Unsurprisingly, the "before" accuracy was much lower than the accuracy of the other two relations. The anchor relation classifier achieved an accuracy of 87.71% over all the RI-TIMEXes with correctly identified anchor points.

The normalization step in our system used a set of rules to parse the TIMEX text span and generates a final ISO8601 standard value based on the anchor point, anchor relation prediction, and the parsed results. We noticed that the performance of this step was poor: only 57.2% of all the RI-TIMEXes with correctly identified text spans, anchor points, and anchor relations received a correct TIMEX value. Our error analysis showed that inconsistency in annotation resulting from the ambiguity of human language contributed to the poor performance. More specifically, many RI-TIMEXes in the corpus are of the format "No. X day after" a certain clinical event (e.g., "Post-operative Day #" "Day of Life #"). Occasionally, the provider will refer to the day of the clinical event (operation or labor, in the above examples) as post-event day #1, while other times they refer to the day after the clinical event as post-event day #1. A few of these ambiguities can be clarified when there is an absolute date assigned to the RI-TIMEX (eg, "Post-operative day #3, 03-14-1998"), but, most of the time, determining the date depends on the annotators' interpretation. To gauge the effect of this inconsistency, we relaxed the evaluation criteria so that it allowed a ± 1 day deviation of the post-event day type RI-TIMEXes. Under this criterion, the normalization step accuracy goes up to 82.09%.

Table 4 shows the F1-measure comparison between our system and the 2012 i2b2 participants results under strict (i2b2 evaluation method) and relaxed evaluation. Note that the i2b2 participating systems were not built to optimize the extraction and normalization of RI-TIMEXes, but RI-TIMEXes constitute more than 1/3 of the i2b2 TIMEX track. Table 4 shows how our system compares to these state-of-art systems. When using the relaxed method, our system has a larger gain than existing methods, which suggests that, for

Table 4: i2b2 Evaluation and Relaxed Evaluation Comparison Between Our Method and the i2b2 Top Performing Systems (RI-TIMEX only)

	Extraction Recall	Value Accuracy	Value Recall
i2b2 Evaluation Method			
i2b2 Avg	86.07	32.04	27.64
i2b2 top score	92.31	34.62	31.96
Our result	90.17	37.47	33.79
Relaxed Evaluation Method			
i2b2 Avg	86.07	43.54	37.51
i2b2 top score	92.31	45.42	41.93
Our result	90.17	53.77	**48.48**

these ambiguous RI-TIMEXes, our results are closer to the truth as a result of the anchor point and anchor relation classification. A randomization test showed that the improvement is statistically significant under the relaxed evaluation method, with a *P*-value of 0.0001.

CONCLUSIONS

Relative and incomplete temporal expressions present significant challenges in temporal reasoning, due to their dependence on context. In this paper, we presented a novel approach to normalizing the value of RI-TIMEXes, based on our analyses of temporally annotated narrative corpora from several domains. We showed that this approach provides statistically significant improvement of normalization results over the existing methods on the held-out test data of the i2b2 corpus. One limitation of this approach is the requirement of RI-TIMEX anchor point and anchor relation annotation. Additionally, this approach should be further tested on larger TIMEX annotated clinical corpora, other than the i2b2 corpus. Due to its simple structure, our method can be readily extended to other corpora or other domains.

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 Library of Medicine, by the National Heart, Lung, and Blood Institute, and by award No. 1R13LM01141101 from the NIH National Library of Medicine. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Library of Medicine, National Heart, Lung, and Blood Institute, or the NIH.

COMPETING INTERESTS

None.

CONTRIBUTORS

W.S. is the primary author. W.S. played the key role in this research, which included leading the data analysis and managing the annotation, designing, and conducting the experiment. A.R. and O.U. offered insights and guidance from data preparation to result analysis, and provided substantial edits to the manuscript. O.U. is the principle investigator.

REFERENCES

1. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc.* 2013;20(5):806–813
2. Sun W, Rumshisky A, Uzuner O. Annotating temporal information in clinical narratives. *J Biomed Inform.* 2013;46,S5–S12.
3. Sun W, Rumshisky A, Uzuner O. Temporal reasoning over clinical text: the state of the art. *J Am Med Inform Assoc.* 2013;20(5):814–819.
4. Pustejovsky J, Lee K, Bunt H, *et al.* ISO-TimeML: An International Standard for Semantic Annotation. In: *LREC 2010.* European Language Resources Association (ELRA);2010.
5. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform.* 2006;39(4):424–439.
6. Lai AM, Parsons S, Hripcsak G. Fuzzy temporal constraint networks for clinical information. In: *Proceedings of the American Medical Informatics Association 2008 Annual Symposium*;2008.Vol. 2008:374.
7. Tao C, Wei WQ, Solbrig HR, Savova G, Chute CG; American Medical Informatics Association. CNTRO: A semantic web ontology for temporal relation inferencing in clinical narratives. In: *Annual Symposium Proceedings (AMIA)*;2010.Vol. 2010:787.
8. Pustejovsky J, Hanks P, Sauri R, *et al.* The TimeBank Corpus. In: *Corpus Linguistics*, Vol. 2003;2003:40.
9. Mazur P, Dale R. WikiWars: A new corpus for research on temporal expressions. In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing.* Association for Computational Linguistics;2010: 913–922. http://dl.acm.org/citation.cfm?id=1870747&picked=formats&CFID=480149059&CFTOKEN=38608567.
10. Verhagen M, Gaizauskas R, Schilder F, *et al.* Semeval-2007 task 15: Tempeval temporal relation identification. In: *Proceedings of the 4th International Workshop on Semantic Evaluations*;2007:75–80. http://dl.acm.org/citation.cfm?id=1621488.
11. 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;2010:57–62. http://dl.acm.org/citation.cfm?id=1859674.
12. UzZaman N, Llorens H, Derczynski L, *et al.* Semeval-2013 task 1: Tempeval-3: Evaluating time expressions, events, and temporal relations. In: *Second Joint Conference on Lexical and Computational Semantics (* SEM)*;2013:Vol. 2: p. 1–9.
13. Derczynski L, Gaizauskas R. Usfd2: Annotating temporal expressions and TLINKs for tempeval-2. In: *Proceedings of the 5th International Workshop on Semantic Evaluation.* Association for Computational Linguistics;2010:337–340.
14. Strötgen J, Gertz M. HeidelTime: high quality rule-based extraction and normalization of temporal expressions. In: *Proceedings of the 5th International Workshop on Semantic Evaluation*;2010:321–324. http://dl.acm.org/citation.cfm?id=1859735.
15. 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*;2010:284–291. http://dl.acm.org/citation.cfm?id=1859727.
16. 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;2010:276–283. http://dl.acm.org/citation.cfm?id=1859726.
17. Chang AX, Manning CD. SUTIME: a library for recognizing and normalizing time expressions. *Language Resources and Evaluation.* 2012:3735–3740.
18. Strötgen J, Gertz M. *Temporal tagging on different domains: challenges, strategies, and gold standards.* In: LREC;2012, pp. 3746–3753. http://www.lrec-conf.org/proceedings/lrec2012/index.html.
19. Sohn S, Wagholikar KB, Li D, *et al.* Comprehensive temporal information detection from clinical text: medical events, time, and TLINK identification. *JAm Med Inform Assoc.* 2013;20(5):836–842.
20. DeSouza J, Ng V. Classifying temporal relations in clinical data: a hybrid, knowledge-rich approach. *J Biomed Inform.* 2013;46:S29–S39.
21. Xu Y, Wang Y, Liu T, *et al.* An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *J Am Med Inform Assoc.* 2013;20(5):849–858.
22. Kovacevic A, Dehghan A, Filannino M, Keane JA, Nenadic G. Combining rules and machine learning for extraction of temporal expressions and

events from clinical narratives. *J Am Med Inform Assoc.* 2013;20(5):859–866.

23. Lin YK, Chen H, Brown RA. MedTime: A temporal information extraction system for clinical narratives. *J Biomed Inform.* 2013;46:S20–S28.
24. Roberts K, Rink B, Harabagiu SM. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *J Am Med Inform Assoc.* 2013;20(5):867–875.
25. Allen JF. Towards a general theory of action and time. *Artif Intell.* 1984;23(2):123–154.
26. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol (TIST).* 2011;2(3):27.
27. De Marneffe MC, MacCartney B, Manning CD, *et al*. Generating typed dependency parses from phrase structure parses. *In: Proc LREC.* 2006;6:449–454.
28. Tang B, Wu Y, Jiang M, *et al*. A hybrid system for temporal information extraction from clinical text. *J Am Med Inform Assoc.* 2013;20(5):828–835.

AUTHOR AFFILIATIONS

[1]Department of Informatics, University at Albany, SUNY. Albany, NY

[2]Department of Computer Science, University of Massachusetts Lowell. Lowell, MA

[3]Department of Information Studies, University at Albany, SUNY. Albany, NY