Journal of Biomedical Informatics 46 (2013) S20–S28

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

MedTime: A temporal information extraction system for clinical narratives

Yu-Kai Lin [a,*], Hsinchun Chen [a], Randall A. Brown [b]

[a] Department of Management Information Systems, University of Arizona, Tucson, AZ 85721, USA
[b] Department of Medicine, College of Medicine, University of Arizona, Tucson, AZ 85724, USA

ARTICLE INFO

Article history:
Received 15 March 2013
Accepted 22 July 2013
Available online 31 July 2013

Keywords:
Temporal information extraction
Temporal expression recognition and normalization
Event recognition
i2b2

ABSTRACT

Temporal information extraction from clinical narratives is of critical importance to many clinical applications. We participated in the EVENT/TIMEX3 track of the 2012 i2b2 clinical temporal relations challenge, and presented our temporal information extraction system, MedTime. MedTime comprises a cascade of rule-based and machine-learning pattern recognition procedures. It achieved a micro-averaged *f*-measure of 0.88 in both the recognitions of clinical events and temporal expressions. We proposed and evaluated three time normalization strategies to normalize relative time expressions in clinical texts. The accuracy was 0.68 in normalizing temporal expressions of dates, times, durations, and frequencies. This study demonstrates and evaluates the integration of rule-based and machine-learning-based approaches for high performance temporal information extraction from clinical narratives.

1. Introduction

The 2012 i2b2 NLP challenge is on temporal relation identification [1]. The objective is to be able to construct patient's clinical timeline from text. To accomplish this end, the challenge comprises three tracks: (a) recognize the extents (text spans) and attributes of events and temporal expressions (TEs) given raw records (*the EVENT/TIMEX3 track*), (b) determine temporal relations given raw records and gold standard EVENT/TIMEX3 tags (*the TLINK track*), and (c) determine temporal relations given raw records (*the end-to-end track*).

We address the EVENT/TIMEX3 track and develop a temporal information extraction system. The reasons to focus on the EVENT/TIMEX3 track are two-fold. First, recognitions of events and TEs from text are the most fundamental tasks for temporal information extraction. Advanced analyses and applications on temporal NLP are not possible without having an event/TE recognition component to start with. Second, there exists no a systematic study on event and TE recognition from clinical narratives. Previous studies have shown that the system performance of TE tagging varies significantly from one document domain to another [2]. This may suggest that the usage of TEs in each domain presents certain unique traits. Most prior work was developed and evaluated on newswire articles, and thus little is known about the characteristics of events and TEs in clinical narratives.

The EVENT/TIMEX3 track is comprised of two tasks: EVENT annotation and TIMEX3 annotation. The first task, EVENT annotation, is to determine clinically relevant events from clinical narratives. The event recognition (ER) resembles the 2010 i2b2 challenges with regards to extracting medical concepts from clinical notes [3]. However, the events here have a much broader range of semantic and linguistic characteristics. Moreover, the events here are not limited to be just noun phrases as medical concepts are. For instance, "asleep" and "consult" are considered clinical events which are neither a medical concept nor a noun phrase. The second task, TIMEX3 annotation, is concerned with temporal expression recognition and normalization (TERN). It requires not only the recognition of clinical TEs but also the retrieval of temporal information from each TE. Several perplexing issues quickly arise, for example, as one tries to determine what date is referred to when "today" is arbitrarily used in a sentence. Similarly, confusion arises when the phrase "postoperative day #2" is used when a document is devoid of an overt operation date.

The paper is structured as follows. Section 2 introduces a brief research background of prior studies. Section 3 delineates the framework of our system, MedTime. Section 4 presents the evaluation results. Section 5 analyzes the effects of normalization strategies, offers an error analysis, and compares the performance of MedTime with other systems. Section 6 concludes this paper.

2. Research background

Temporal information processing has been an important area in biomedical and health informatics research. Temporal information systems are developed to facilitate healthcare management, predict disease risk or progression, and search for similar clinical

* Corresponding author. Fax: +1 520 6266499.
E-mail address: yklin@email.arizona.edu (Y.-K. Lin).

1532-0464/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jbi.2013.07.012

patterns [4–6]. Most work, however, has been confined to representing and analyzing numerical or categorical electronic health record (EHR) data. Despite the proliferation of medical NLP studies in the past decade, there has been no concerted effort to address the problem of temporal information extraction from clinical narratives [7]. On the other hand, temporal tagging of natural language text has gained considerable attention recently in computational linguistics and artificial intelligence. One major motivation is the practical need for temporal-aware NLP applications, e.g., event monitoring, temporal question answering, and document summarization.

As a central functionality in temporal information extraction, TERN plays a pivotal role in determining temporal relations and understanding messages. To be useful in temporal-aware NLP applications, TEs need to be recognized and normalized such that their temporal information is encoded explicitly in a standard format. However, the rich representations of temporal information in natural language make automatic TERN a challenging task. Because TEs are often vague or underspecified, TERN is difficult even for human annotators.

Based on how temporal information is represented, Alonso et al. categorized TEs into three groups: explicit, implicit, and relative [8]. Explicit TEs are the TEs that have fully specified and self-contained temporal information, such as "Nov. 24th, 2011" or "three times a week." Explicit TEs can be normalized without resorting to any external information, which is not the case in normalizing implicit and relative TEs. Implicit TEs use an alias to represent the actual temporal information, such as "Thanksgiving 2011" or "admission date." Normalizing implicit TEs involves knowledge about the aliases, e.g., the exact date of Thanksgiving or admission in the previous examples. Finally, relative TEs anchor their temporal information to a contextual reference point, such as "last evening" with respect to the present time. Note that relative TEs can anchor on implicit TEs. For instance, "postoperative day #3" is a relative TE anchored on an implicit TE "operation date" which needs to be determined from the document.

Over the years several temporal information extraction systems have been developed. Most of them evolved from shared tasks such as MUC-6, MUC-7, ACE 2004, ACE 2007, and TempEval-2. Given the similarity between the TempEval-2 competition [9] and the 2012 i2b2 challenge, the results from the former offer important insights for the current study. The first implication is about extent detection. We have found that conditional random fields (CRF) are a very effective technique for detecting the extent of events and TEs. Two CRF-based machine learning applications, TIPSem [10] and TRIOS [11], both obtained high *f*-measures in recognizing events and TEs. The second implication is from the val score, which is an evaluation metric quantifying the performance of temporal normalization. Compared to the *f*-measures, the val scores reported by the teams in the TempEval-2 competition have greater variance and lower average value, with a mean of 0.57 and median of 0.59. The highest val score 0.85 was from a rule-based system named HeidelTime. This suggests that temporal normalization is a more difficult procedure than temporal recognition and that a rule-based approach is an effective design for temporal normalization tasks. Indeed, there is still no elegant machine learning approach that could normalize temporal expressions. Even the top machine learners in TempEval-2 need to develop rules to normalize TEs after the extents were determined by their supervised models.

3. System design: MedTime

This section describes the design of our proposed MedTime system (Fig. 1). Our design is a hybrid and cascade framework, interweaving rule-based and machine learning procedures for temporal information extraction in six major steps: (1) pre-processing, (2) temporal tagging by HeidelTime, (3) Clinical FREQUENCY TE tagging, (4) sequence labeling, (5) clinical temporal normalization strategies, and (6) post-processing.

3.1. Pre-processing

Pre-processing consists of subroutines that support the core information extraction procedures.

3.1.1. Section time extraction

The clinical narratives in the challenge corpus contain two types of section times: ADMISSION and DISCHARGE. The section times have important clinical implications and can be meaningful reference dates for temporal normalization. They are analogous to the document creation times in the TimeML corpora but with a less standardized format. The document creation times in the TimeML corpora are considered as metadata. They are in a uniform format, and each document must associate with one document creation time. By contrast, the section times in the clinical narratives are part of the text. They are expressed in diverse formats, and in some cases, may not even exist in a clinical narrative.

Given its importance to our temporal normalization procedure, we extract section times before we proceed to our regular temporal tagging. We observed that, when present, the section time expressions are placed in the first few lines of a clinical narrative, under the headings of "Admission Date" and "Discharge Date." In addition, we also found that section times are represented as explicit TEs. As such, we develop a simple regular expression algorithm to extract and normalize section times.

3.1.2. Text cleaning

One requirement from the challenge is that the system needs to be able to process different formatted clinical narratives, which could have originated from different health care institutions. The text cleaning subroutine aims to address two document formatting issues. First, many of the clinical narratives are appended with a section of electronic signature, which contains TEs irrelevant to the patient's clinical timeline. We formulate rules to remove these texts before proceeding to the core TERN procedures. The second issue is from the inconsistent XML encoding. A small portion of the clinical narratives are XML well-formed, which substitute (&, ' , >, <, ") in the text for predefined entities, e.g., & is replaced by &. To unify text representation, we convert these predefined entities back to their original characters, e.g., from & to &.

3.1.3. Feature generation

We extract morphological, syntactic, semantic, and composite features from clinical narratives to enable machine learning (Table 1). The morphological, syntactic, and some of the composite features, i.e., noun phrase (NP) chunks and adjacent features, are common in prior NLP studies. Many other participating systems in the 2012 i2b2 challenge also include these basic features [12,13]. Most of the morphological and syntactic features are generated using the Stanford CoreNLP package [14]. On the other hand, semantic features are domain-specific and knowledge-based. Medical abbreviations are commonly used in clinical narratives for both clinical events (e.g., HTN for hypertension) and clinical temporal expressions (e.g., BID for twice daily). As such, successful identification of medical abbreviations is indispensable to clinical NLP tasks. We incorporate a comprehensive list of medical abbreviations from Wikipedia (http://en.wikipedia.org/wiki/Category:Lists_of_medical_abbre-

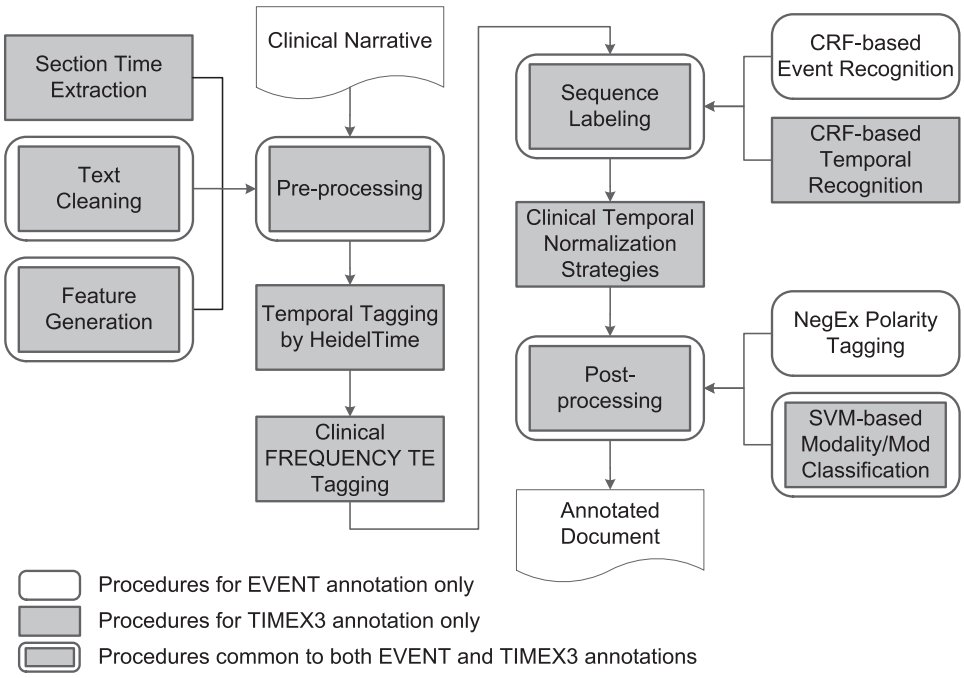

Fig. 1. MedTime system overview.

Table 1
Machine learning feature set.

Type	Feature	Source
Morphological	Part-of-speech (POS) tag	Stanford CoreNLP
	Word lemma	Stanford CoreNLP
	Stemmed string	Porter stemmer (http://snowball.tartarus.org)
Syntactic	Token lexical class	Apache OpenNLP (http://opennlp.apache.org)
	Token shape	Stanford CoreNLP
	Token 2–4 character prefix	Stanford CoreNLP
	Token 2–4 character suffix	Stanford CoreNLP
Semantic	Medical abbreviation	Wikipedia
	SPECIALIST lexicon	MetaMap
	Semantic types	MetaMap
Composite	NP chunk	POS tags
	Diagnosis NP chunk	Semantic type (diap) + NP chunk
	Finding NP chunk	Semantic type (fndg) + NP chunk
	Temporal NP chunk	Semantic type (tmco) + NP chunk
	Adjacent features	Previous/Next 2 tokens in the same sentence

viations) to identify medical abbreviation features from clinical text. The unified medical language system (UMLS) provides standardized medical lexicon and semantic type for every medical concept. It has been known that incorporating domain specific features can improve the performance of NLP tasks. Hence, we use MetaMap [15], a front-end API of UMLS, to extract features of medical lexicon and semantic types. Finally, the diagnosis/finding/temporal NP chunks are composite features integrating semantic types and NP chunks, which aim to bring additional semantic information to NP chunks. Table 2 illustrates the core features. All these features are shared by the subsequent CRF-based event and temporal recognizers.

3.2. Temporal tagging by HeidelTime

While automatic TERN is still an open and emerging area, sophisticated, publicly available tools exist for processing news articles. As the best performing system in TempEval-2, HeidelTime was able to achieve a very high val match score in tagging news articles. Given the high demand for accurate val prediction in creating meaningful clinical timelines, we decided to incorporate HeidelTime as our initial temporal tagger. Because HeidelTime was already tuned towards high precision, no deleterious effect on noise containment was expected, even though HeidelTime did not have rules and patterns specific to clinical narratives.

3.3. Clinical FREQUENCY TE tagging

FREQUENCY is a type of TEs unique to the clinical domain. Many of the FREQUENCY TEs are clinical abbreviations, such as BID (twice a day) or q.8.h (every eight hours), which are not well-treated by HeidelTime. We identify two important characteristics from the FREQUENCY TEs in the corpus. First, the FREQUENCY TEs are often explicit. That is, normalizing these TEs does not require any reference information. Second, the FREQUENCY TEs generally have very regular syntactic patterns. For example, **[q.6.h]**, **[Q.8.h.]**, and **[q 12 h]** are patterns for every 6 h, every 8 h, and every 12 h, respectively. As another example, the patterns **[X 2]**, **[3 x]**, and **[x four]** denote, respectively, 2 times, 3 times, and 4 times. Given these two characteristics, we choose to develop a set of recognition patterns and normalization rules specially tailored for the FREQUENCY TEs.

We use the following clinical FREQUENCY TE pattern to illustrate how we recognize and normalize FREQUENCY TEs. Notice that instead of showing the actual regular expression, we demonstrate the pattern in a more readable format for didactical purposes. Now consider a simple pattern:

Table 2
Illustration of the core features.

TEXT	POS_TAG	LEMMA	STEM	TOKEN_CLASS	SHAPE	PREFIX_2	SUFFIX_2	MED_ABBR	LEXICON	SEMANTIC_TYPE	NP	Diagnosis_NP	Finding_NP	Temporal_NP
Her	PRP$	she	Her	ic	Xxx	He	er	F	O	O	B-NP	F	F	F
steroids	NNS	steroid	steroid	lc	xxxxx	st	ds	F	B-lexicon	B-strd	I-NP	F	F	F
were	VBD	be	were	lc	xxxx	we	re	F	O	O	O	O	O	O
tapered	VBN	taper	taper	lc	xxxxx	ta	ed	F	B-lexicon	B-hlca	O	O	O	O
and	CC	and	and	lc	xxx	an	nd	F	O	O	O	O	O	O
Pulmonary	JJ	pulmonary	Pulmonari	ic	Xxxxx	Pu	ry	F	B-lexicon	B-qlco	O	O	O	O
was	VBD	be	wa	lc	xxx	wa	as	F	O	O	O	O	O	O
consulted	VBN	consult	consult	lc	xxxxx	co	ed	F	B-lexicon	B-hlca	O	O	O	O
who	WP	who	who	lc	xxx	wh	ho	F	O	O	O	O	O	O
recommended	VBD	recommend	recommend	lc	xxxxx	re	ed	F	B-lexicon	B-idcn	O	O	O	O
a	DT	a	a	lc	x	a	a	F	O	O	B-NP	T	F	F
CT	NN	ct	CT	ac	XX	CT	CT	T	B-lexicon	B-diap	I-NP	T	F	F
scan	VB	scan	scan	lc	xxxx	sc	an	F	I-lexicon	I-diap	O	O	O	O
of	IN	of	of	lc	xx	of	of	F	O	O	O	O	O	O
the	DT	the	the	lc	xxx	th	he	F	O	O	B-NP	F	F	F
chest	NN	chest	chest	lc	xxxxx	ch	st	F	B-lexicon	B-blor	I-NP	F	F	F
to	TO	to	to	lc	xx	to	to	F	O	O	I-NP	F	F	F
evaluate	VB	evaluate	evalu	lc	xxxxx	ev	te	F	B-lexicon	B-ftcn	I-NP	F	F	F
the	DT	the	the	lc	xxx	th	he	F	O	O	B-NP	F	F	F
lung	NN	lung	lung	lc	xxxx	lu	ng	F	B-lexicon	B-blor	I-NP	F	F	F
parenchyma	NN	parenchyma	parenchyma	lc	xxxxx	pa	ma	F	I-lexicon	I-blor	I-NP	F	F	F
.	.	.	.	other	.	.	.	F	O	O	O	O	O	O

Pattern : = ({Letter "q"}|{Word"every"})
+ ({Dot}|{Hyphen}|{Space}|{Empty String})
+ ({Digit}|{Number Text})
+ ({Dot}|{Hyphen}|{Space}|{Empty String})
+ ({Time Unit}|{Date Unit})

This pattern comprises five portions, one in each line. This pattern recognizes FREQUENCY TEs such as **[q-6-h]**, **[Q.8.H]**, **[Q1D]**, and **[every eight hours]**. In addition, since that the third portion captures the quantity and that the fifth portion captures the unit, this pattern can be further utilized to normalize these FREQUENCY TEs. For example, after recognizing that **[every eight hours]** is a FREQUENCY TE and knowing that the quantity is eight and the unit is hour, the normalized value "RPT8H" can be trivially derived, in which (1) the RP are the designated leading symbols for FREQUENCY TEs if they involve repetitions, (2) the T symbol comes from the fact that the FREQUENCY TE use a time unit (hour), (3) the 8 is the quantity been captured, and (4) the H symbol is for the actual time unit—hour.

3.4. Sequence labeling

As shown in TempEval-2, CRF is an effective technology in recognizing extents of events and TEs in newswire articles. CRF has also been demonstrated to be very effective in extracting clinical concepts, including medical problems, treatments, and tests [16]. We adopted MALLET [17] to train two CRF models, one for EVENT annotation and the other for TIMEX3 annotation. We encoded labels in IOB2 format [18] with type information. That is, the label B-TIMEX3-DATE represents a token which is the beginning of a date TE, and I-EVENT-TREATMENT represents a token which is inside a treatment event. Through this approach, the CRF models predict the extent and the type of an annotation simultaneously.

With the initial tagging from HeidelTime and our FREQUENCY TE tagger, the CRF-based temporal recognition aims at extending the coverage to domain specific TEs, e.g., "post-op day four" and "one day prior to admission." This is achieved by using domain specific documents, i.e., the clinical narratives, to train the CRF models. Note that the CRF-based temporal recognizer may recognize existing TIMEX3 annotations from HeidelTime as well as the FREQUENCY TE Tagger. In this case, we keep the original annotations, given that the existing annotations were originated from rule-based procedures tuned towards high precision.

3.5. Clinical temporal normalization strategies

Our rule-based temporal normalizer is built upon JChronic, an open source date parser in Java. We extend and modify JChronic to better handle the implicit and relative TEs in the clinical narratives. The JChronic program requires "present time" and "direction of offset" as parameters to calculate the time of an input TE. Our three novel normalization strategies guide JChronic's behavior by resolving the required parameters, i.e., the present time and the direction of offset.

Algorithm 1 delineates our temporal normalization steps. The recognized TEs from the previous sequence labeling procedure are stored in a list, sorted by their appearance order in the document. That is, the first TE in the list is the first TE mentioned in the document, and the last TE in the list is the last TE mentioned in the document. The nested for-loops iterate all sentences and TEs sequentially, and try to pair TEs with their corresponding sentences. A sentence can provide contextual cues for the encompassed TEs. Therefore, the corresponding sentence of a TE is considered in normalizing the TE. DATE and TIME TEs are normalized with the procedure **normalizeDateTime** while DURATION TEs are normalized by another procedure **normalizeDuration**. Notice that here we do not normalize FREQUENCY TEs since they should be normalized by our clinical FREQUENCY TE tagger in a prior step.

Even with these temporal normalization steps, some of the TEs passed from the CRF model may still not be normalized—either because that these are false positive TEs or that the TEs have the unusual patterns that have not been captured by our existing normalization rules. In any case, we choose to drop these TEs to ensure the produced TIMEX3 annotations all have values for their val attribute.

Our three strategies take place at various procedures in Algorithm 1. As an overview, we propose *contextual alias registry* (Strategy CAR) and *chronological order of TEs* (Strategy COTE) to resolve reference time for implicit TE, e.g., **[postoperative day #3]** and relative TEs **[last evening]**, respectively. For underspecified TEs, e.g., **[Tuesday]**, we propose *distance-based direction determination* (Strategy DDD), paired with lexical markers, to identify the direction of offset. The design rationales of each strategy are discussed separately in the following.

Algorithm 1. Temporal Normalization Steps

Input: A clinical narrative document D and a list of TIMEX3 tags L from D, each possessing values of their id, start, end, text, and type attributes (from our CRF-based temporal recognition procedure).
Output: A list of TIMEX3 tags with all their val attributes resolved.

sort L by the start attribute values in ascending order;
let $Start_\alpha$,End_α,$Text_\alpha$,$Type_\alpha$ and Val_α be the respective attribute values of α, $\forall\alpha \in L$;
let R be a map with keys ADMISSION, DISCHARGE, OPERATION, BIRTH, and NOW;
initializeContextualAliasRegistry(R,D)
for each sentence s in D **do** /* *Consider sentence as the context for the TEs within the sentence* */
  let $Start_s$,End_s be the start and end positions of s in D;
  for each TIMEX3 tag α in L do
    if ($Start_\alpha$ > End_s) or (End_α < $Start_s$) **then**
      continue on the next TIMEX3 tag; /* *Ignore TIMEX3 tags if they are not in the sentence s* */
    end
    if ($Type_\alpha$ is DATE) or ($Type_\alpha$ is TIME) **then**
      Val_α := normalizeDateTime(α,s,R); /* *Normalize TEs with types DATE or TIME* */
      updateContextualAliasRegistry(R,s,Val_α);
    else if ($Type_\alpha$ is DURATION) and ($Text_\alpha$ contains " to ") **then**
      Val_α := normalizeDuration($_\alpha$); /* *Normalize TEs with type DURATION* */
    end
    if Val_α is *null* **then**
      remove α from L; /* *Remove TIMEX3 tags if the TEs could not be normalized* */
    end
  end
end
return L;
Procedure initializeContextualAliasRegistry(R,D) : /* *Strategy CAR* */
  R.ADMISSION := R.NOW := *D.SECTIME.ADMISSION*;
  R.DISCHARGE := *D.SECTIME.DISCHARGE*;
  R.OPERATION := *null*; R.BIRTH := *null*;
Procedure updateContextualAliasRegistry(R,s,Val_α) : / * *Strategy CAR* */
  if R.OPERATION is *null* and s has a procedure word **then** /*MetaMap semantic type =topp */
    R.OPERATION := Val_α ; /*Found Operation Date */
  end
  if R.BIRTH is *null* and s contains "born" or "birth" **then**
    R.BIRTH := Val_α ; /*Found Date of Birth */
  end
  if R.NOW $\leqslant$ Val_α **then** /* *Strategy COTE* */
endFunction normalizeDateTime(α,s, R) : **if** ($Text_\alpha$ matches any triggers) or (s matches any triggers) **then** Val_α' := *normalizeWithAlias*($Text_\alpha$, s, R);
  /* *Strategy CAR;see Algorithm 2* */ **else** Val_α' := *normalizeWithJchornic*($Text_\alpha$, s, R);
  /* *Strategy COTE&DDD;see Algorithm 3* */ **end return** Val_α';Function normalizeDuration(α) : (TE_1, TE_2) := split $Text_\alpha$ by " to "; Val_α' := normalizeDuration(TE_1, TE_2); /* In that it determines the difference and time granularity */ **return** Val_α';

3.5.1. Strategy of contextual alias registry (CAR)

In contrast to newswire articles, there are significant dates that we normally expect in a clinical document: *Admission Date*, *Discharge Date*, *Operation Date*, and/or *Date of Birth* (for childbearing). We observe from the corpus that it is common for implicit and relative TEs to anchor on these domain-specific contextual alias dates. As such, we maintain a contextual alias registry in the temporal normalization process (Table 3). The actual dates/times for the aliases are initialized by the **initializeContextualAliasRegistry** procedure and updated by the **updateContextualAliasRegistry** procedure in Algorithm 1. To use the aliases in the normalization, each alias has a set of triggering rules such that if a TE or its adjacent words match one of the rules, the respective time of the alias will be considered as the reference time (Algorithm 2).

Algorithm 2. Temporal Normalization Using Contextual Alias Registry

Input: A temporal expression text $Text_\alpha$, the sentence s which contains $Text_\alpha$, and our contextual alias registry R.
Output: A normalized value of $Text_\alpha$.
RefTime := retrieve the alias from R by matching $Text_\alpha$ or s the with the triggers defined in Table 3;
TempShift := determine the temporal shift of $Text_\alpha$ with respect to *RefTime*;
Val_α' := apply *TempShift* on *RefTime*;
return Val_α';

3.5.2. Strategy of Chronological Order of Temporal Expressions (COTE)

It is common and intuitive to narrate stories in a chronological order. From the training documents, we observed a very high correlation between the appearance order of DATE/TIME TEs in a document and their timestamps (the average spearman rank correlation across all training texts is around 0.80). However, occasionally there are retrospective statements inserted within the main storyline. We posit that anchoring non-explicit TEs to a reference point in the retrospective statements is prone to be erroneous. Therefore, we propose a novel heuristic to improve the resolution of reference time which embodies a chronological constraint.

Suppose that the temporal normalization is executed sentence by sentence and from left to right in a sentence, we maintain a reference time denoting the largest timestamp normalized from the TEs thus far. We allow absolute as well as relative TEs as long as the type is DATE or TIME. Fig. 2 illustrates the difference between the chronological time heuristic and another common heuristic, the "previously mentioned time." Suppose that all the TEs in Fig. 2 are DATE TEs and that TE_4 is an underspecified TE requiring a reference time as an anchor. The chronological time heuristic chooses TE_2 as the reference time because for the content parsed thus far, TE_2 represent the rightmost time point on the timeline axis (comparing to TE_1 and TE_3). The previous mentioned time heu-

ristic, on the other hand, will use TE_3 as a reference time for TE_4 because TE_4 follows TE_3.

3.5.3. Strategy of distance-based direction determination (DDD)

Verb tense and lexical markers (e.g., "past" or "next") are considered effective devices to determine the direction of implicit and relative TEs in newswire articles. However, we find that tense is not an appropriate strategy in the domain of clinical narratives. This is due to the facts that the statements in clinical narratives are mostly retrospective, which means that the texts are predominately past-tensed. As such, we do not employ tense in determining the direction. We, instead, combine lexical markers with a novel distance-based strategy. That is, when lexical markers are absent from the context, we resolve the direction problem by choosing whichever direction gives a closer temporal distance to the discharge date (Algorithm 3). Akin to the COTE strategy, it is more likely to have a temporal direction which results in a shorter temporal distance to the discharge date. For example, suppose that the normalizing TE is **[Tuesday]** and that the reference time and the discharge date are known to be "20121015" (Monday) and "20121017" (Wednesday), respectively. If the direction of offset is past, the normalized value for **[Tuesday]** will be "20121009." On the other hand, if the direction of offset is future, the normalized value will be "20121016." Between the two possibilities, the second one seems more likely because it gives a time point that is closer to the discharge date. However, we ignore the direction that will lead to a normalized time which is greater than the discharge date. Continue with the settings in the previous example, but substitute the normalizing TE with **[Friday]**. We ignore the future direction because it gave a time point "20121019," which is after the actual discharge date.

Algorithm 3. Temporal Normalization Using JChronic

Input: A temporal expression text $Text_\alpha$, the sentence s which contains $Text_\alpha$, and our contextual alias registry R.
Output: A normalized value of $Text_\alpha$.
RefTime := retrieve *R.NOW* from *R*; /* *Strategy CAR* */
if $Text_\alpha$ or s contains a lexical marker of direction **then**
  direction := determine the direction according to the lexical marker in $Text_\alpha$ or s;
  Val'_α := *jchronic*(presentTime=*RefTime*, direction=*direction*, temporalExpression=$Text_\alpha$);
else
  Val^1_α:= *jchronic*(presentTime=*RefTime*, direction=PAST, temporalExpression=$Text_\alpha$);
  Val^2_α:= *jchronic*(presentTime=*RefTime*, direction=FUTURE, temporalExpression=$Text_\alpha$);
  Distance^1 := (R.DISCHARGE − Val^1_α);
  Distance^2 := (R.DISCHARGE − Val^2_α);
  if (Distance^1 > Distance^2) and (Distance^2 > 0) **then** /* *Strategy DDD* */
    Val'_α := Val^2_α;
  else
    Val'_α := Val^1_α;
  end
end
return Val_α';

3.6. Post-processing

The post-processing is a set of classification routines for the remaining EVENT/TIMEX3 attributes. Specifically, there are polarity and modality attributes for EVENT annotation and Mod attribute for TIMEX3 annotation that need to be determined.

3.6.1. NegEx polarity tagging

Polarity is one of the required attributes for EVENT annotations. It specifies whether the described clinical event is negated or not. The polarity tagging in MedTime is performed by NegEx [19]. NegEx is a simple regular expression algorithm for negation determination in clinical documents. It has been widely adopted in clinical NLP systems. We noted that there are some recent studies that apply CRF in determining the scope of negations [20]. We found there are several issues that merit further exploration. To keep this study manageable, we chose to use the well-known NegEx, and plan to examine the efficacy of machine-learning-based polarity tagging in the future.

3.6.2. SVM-based modality/Mod classification

The cascade of CRF-based sequence labeling and SVM-based classification has been demonstrated as an effective design in previous studies of medical information extraction [21]. Here we train an SVM classifier for EVENT modality classification and another for TIMEX3 mod classification using the LIBSVM library [22]. The modality/mod classifications share the same feature set for the CRF-based sequence labeling, except that here we also include token N-grams ($1 \leqslant N \leqslant 3$) as new features. The reason to have an extended feature set is that the original feature set only captures features at the token level. Although a token-level feature set is sufficient for sequence labeling, we suspect that it may not be adequate for the modality/mod classification because the training and prediction entity here is an event or a TE which can have an arbitrary number tokens. In

Table 3
Contextual alias registry.

Alias	Source	Triggers	Example
Admission Date	Section time	(hospital|admission|AD|HD| emergency|presentation)	hospital day two (*Admission Date* + 1D)
Discharge Date	Section time	(discharge)	the night before discharge (*Discharge Date* – 1D)
Operation Date	The date of *Now* when the first surgical concept occurs (word with semantic type "topp") Default: *Admission Date*	(procedure|surgery|POD|op)	postoperative day #4 (*Operation Date* + 4D)
Date of Birth	The date of *Now* when the keywords "born" or "birth" occur Default: *Now*	(life)	day of life 3 (*Date of Birth* + 2D)
Now	The normalized date/time of the TE maintained by Strategy 2 Default: *Admission Date*	When none the above are applicable	last evening (*Now* – 1D)

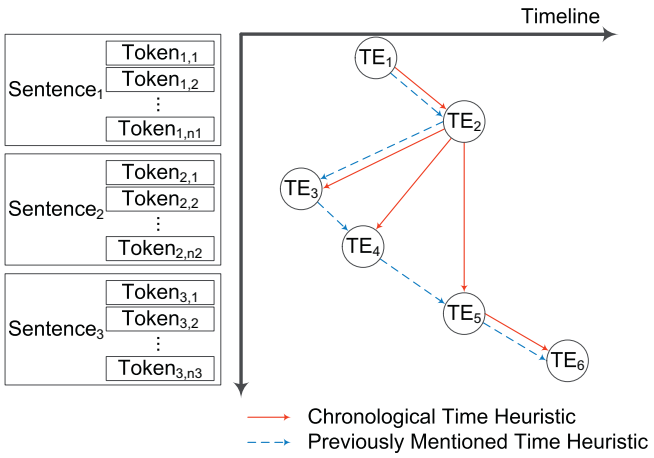

Fig. 2. Reference time identification by two different heuristics: the chronological time heuristic and the previously mentioned time heuristic.

other words, the token *N*-grams features of event expressions or TEs can be considered as annotation-level features in building the SVM classifiers.

4. Experiment results

There are principally two categories of evaluation for the EVENT/TIMEX3 task: extent and attribute. We first provide a brief summary of each, followed by the evaluation results of MedTime.

There are two ways to define accuracy of an extent prediction: exact extent match and partial extent match. For exact extent match, a prediction is considered accurate only when the predicted extent is exactly the same as the gold standard extent. On the other hand, for partial extent match, we consider a prediction accurate as long as the predicted extent overlaps the gold standard extent. The 2012 i2b2 challenge uses partial extent match as the default extent evaluation criterion. With this definition of correctness, the evaluation metrics for EVENT/TIMEX3 extent recognitions are micro-averaged precision, recall, and *f*-measure. Given a test corpus D and let TP_d,FP_d,FN_d denote, respectively, the numbers of true positive, false positive, and false negative EVENT/TIMEX3 extent predictions in a document d, where $d \in D$:

$$\text{Micro-Averaged Precision } (P) = \frac{\sum_{d\in D} TP_d}{\sum_{d\in D}(TP_d + FP_d)}$$

$$\text{Micro-Averaged Recall } (R) = \frac{\sum_{d\in D} TP_d}{\sum_{d\in D}(TP_d + FN_d)}$$

$$\text{Micro-Averaged } F\text{-Measure } (F) = \frac{2(P\times R)}{P+R}$$

The evaluation for attribute match consists of two steps. It first matches the system predictions with the gold standards by their extents. Then, among the total matched predictions, attribute score is calculated as the percentage of correct attribute predictions. Among the attribute match scores, the TIMEX3 val match score is arguably the most critical one as it signifies the efficacy of a temporal information extraction system in recovering temporal information from TEs.

For event recognition, MedTime achieved an 87.94% accuracy (*F*-measure) against hand-annotated data in the testing corpus (Table 4). The performance is comparable with the best performing system TIPSem in the TempEval-2 event recognition task.

Through our stepwise evaluations on MedTime's TIMEX3 tagging pipelines, we decompose the contribution of each major procedure in Table 5. The first major procedure was HeidelTime tagging. We observed a very high precision from HeidelTime. However, the recall and val match score at this step were both considerably low. In the second step, the parser tailored for FREQUENCY TEs increased both recall and val match score by about 8%. Given that FREQUENCY only accounts for about 10% of total TIMEX3 tags in the testing corpus, the FREQUENCY TE tagger provided considerable enhancement of the result. The CRF-based sequence labeling in the third step significantly improved the overall recall. This major improvement brought the *f*-measure to 0.88. Finally, the rule-based temporal normalizer increased the val score from 0.41 to 0.68. This demonstrates the efficacy of our rule-based temporal normalizer as well as our normalization strategies.

5. Discussion

Temporal normalization is arguably the most challenging part of temporal information extraction, and hence warrants additional discussion. With the earlier system evaluation, two remaining questions are: a) how normalization strategies affect the performance of temporal normalization, and b) what kinds of normalization errors were made in our current design. In Section 5.1, we analyze and discuss different normalization strategies for clinical narratives. The objective is to uncover the efficacy of each strategy and identify the TE characteristics in the domain of clinical narrative. Following which is an error analysis in Section 5.2. We investigate the major categories of normalization errors, illustrate the errors with examples, and discuss potential solutions for each of them. Finally, Section 5.3 compares the performance of MedTime with other participating systems in the 2012 i2b2 NLP Challenge.

5.1. Analysis of normalization strategies

Normalization strategies should be considered according to the TE characteristics of each document domain [23]. We analyze MedTime's temporal normalization performance on the test corpus using different settings of normalization strategies. In this analysis, we remove HeidelTime from our pipeline in order to be able to isolate the net effects of the strategies.

Table 6 presents how different normalization strategies affect the val match score. We find that the normalization strategies combining chronological time and contextual alias registry achieved the best performance (0.661). Given that admission date and operation date are two most common referred aliases and that many patients had the operation on the date of admission, they may explain why admission date is a reasonable reference time and outperformed the chronological time strategy at the absence of contextual alias registry. The contextual alias registry is shown

Table 4
Performance of MedTime's EVENT annotation.

Metric	Score
Precision	0.924
Recall	0.839
F-measure	0.879
Polarity match score	0.793
Modality match score	0.801

Table 5
Stepwise performance analysis on MedTime TIMEX3 annotations.

	(Step 1) HeidelTime	(Step 2) Frequency TE tagging	(Step 3) Sequence labeling	(Step 4) Normalization strategies	(Step 5) Post-processing
Precision	0.941	0.863	0.879	0.879	0.879
Recall	0.406	0.485	0.884	0.884	0.884
F-measure	0.567	0.621	0.881	0.880	0.880
Type match score	0.376	0.453	0.822	0.821	0.821
Mod match score	0.370	0.449	0.809	0.809	0.828
Val match score	0.334	0.415	0.453	0.688	0.688

to be an important strategy in MedTime. Incorporating the contextual alias registry is consistently better than without it for the three reference time identification strategies. This suggests that in clinical narratives, it is common to express time with respect to the salient clinical dates, e.g., one day prior to admission, postoperative day number four [24].

5.2. Error analysis of temporal normalization

From the evaluation results, we identify three major categories of temporal normalization errors made by MedTime. The first error category is about mistake DURATION TEs as DATE TEs. One example illustrating this type of errors is the following:

He was essentially bed bound for **[3 days]** *prior to admission.*

The system annotates **[3 days]** as a DATE TE while the gold standard is DURATION. The error is due to erroneous framing of context as "3 days prior" instead of "for 3 days." However, this is a rather tricky case because "3 days prior" is a more likely a grammatical unit than "for 3 days" (see Fig. 3). To correct this category of errors, it may require a more sophisticated parser that considers a broader context in both directions and perhaps with some overwriting rules or a disambiguation procedure.

The second category of errors is also related to the mis-categorization of DURATION TEs. There are DURATION TEs formatted like a FREQUENCY TE. For example,

Ms. Crossman is an 84 year old female with complaints of abdominal pain, diarrhea, nausea and vomiting **[x 1 week]**, ...

The system considers **[x 1]** to be a FREQUENCY TE with val attribute value "R1," while the gold standard is a DURATION **[x 1 week]** with val value "P1W." The patterns (x/d) and (/d x), where /d represents a digit, are commonly used in describing the frequency of medication, and are included in the FREQUENCY TE tagger. This type of errors can be easily fixed by testing the context after (x/d), and change the annotation to a DURATION TIMEX3 tag if the pattern is followed by a time unit, e.g., week.

The third category is errors in the contextual alias registry. It is essential to have accurate dates and times for the contextual aliases because these implicit TEs are repeatedly referred and anchored by relative TEs. In building the contextual alias registry, the system occasionally got the wrong date for the alias *Date of Birth*. These errors occur when the content has vague expressions modifying the alias date. For example, a clinical narrative starts with the following sentence:

Baby Daniel Holman was born at 39 and 3/7 weeks gestation (EDC 2015-04-25).

The system incorrectly considers "2015-04-25" as the *Date of Birth*. In this particular example, the *Date of Birth* is the admission date, but such information is not explicit in the document. Given this mistake made by MedTime, all the relative TEs in this document anchored on *Date of Birth*, e.g., "16 hours of life" and "day of life 2," are erroneously normalized. This suggests that significant attention should be devoted in building the contextual alias registry because the error could propagate to other TEs.

```
(ROOT
    (S
        (NP (PRP He))
        (VP (VBD was)
            (ADJP
                (ADJP (RB essentially) (RB bed) (VBN bound)
                    (PP (IN for)
                        (ADVP
                            (NP (CD 3) (NNS days))
                            (RB prior))))
                (PP (TO to)
                    (NP (NN admission)))))
        (. .)))
```

Fig. 3. Parsed tree of the example sentence.

Table 6
Comparison of normalization strategies.

Reference time	w/o Contextual alias registry	w/ Contextual alias registry
Admission date	0.639	0.653
Previously mentioned time	0.591	0.653
Chronological time	0.604	0.661

5.3. Comparison with other participating systems in the 2012 i2b2 NLP challenge

To triangulate the ER and TERN performance of MedTime with the state of the art, it may be best to compare MedTime with the other participating systems in the 2012 i2b2 Challenge. Table 7 summarizes various EVENT and TIMEX3 evaluation metrics in the Challenge. The primary score for the EVENT task is *f*-measure while the one for the TIMEX3 task is the product of *f*-measure and val match score. Among a total of 14 teams participated in the EVENT/TIMEX3 track, MedTime is ranked the fourth on both EVENT and TIMEX3 tasks based on the respective primary scores. Interested readers are pointed to Sun et al. [1] for a systematic comparison among all participating systems.

Table 7
Comparing MedTime with other participating systems in the 2012 i2b2 challenge.

	Primary score	Precision	Recall	*F*-measure	Type	Polarity	Modality
(a) EVENT							
MedTime	0.879 (4th place)	0.924	0.839	0.879	0.735	0.793	0.803
Average	0.792	0.822	0.775	0.792	0.674	0.738	0.685
Medium	0.857	0.902	0.824	0.857	0.738	0.786	0.770
Maximum	0.917	0.942	0.893	0.917	0.857	0.859	0.856
	Primary score	Precision	Recall	*F*-measure	Type	Val	Modifier
(b) TIMEX3							
MedTime	0.606 (4th place)	0.879	0.884	0.880	0.821	0.688	0.828
Average	0.473	0.807	0.759	0.773	0.705	0.563	0.695
Medium	0.529	0.863	0.847	0.873	0.781	0.603	0.788
Maximum	0.656	0.951	0.949	0.914	0.893	0.729	0.891

6. Conclusions and future work

Temporal information extraction from clinical narratives is of critical importance to many clinical applications. For the 2012 i2b2 clinical temporal relations challenge, we demonstrated an effective solution to the EVENT/TIMEX3 track and presented a temporal information extraction system, MedTime.

Rule-based systems tend to have very high precision, but often with relatively low recall. On the other hand, machine learning approaches enable reasonable treatments of unanticipated and novel cases. For complex problems, such as temporal information extraction, it seems to be a reasonable design to combine the two approaches. Our experiments demonstrate the efficacy of this hybrid design.

MedTime is still under a continuous development towards a comprehensive temporal information extraction platform. Our error analysis suggested several directions for further enhancement. One limitation of current MedTime is the lack of temporal relation identification (i.e., TLINK annotation). We are working on building such component, which could enable MedTime to provide a broader range of practical applications in these clinical contexts in the future.

Acknowledgments

This research is supported by the US National Science Foundation Grant CBET-0730908, Defense Threat Reduction Agency Grant HDTRA10910058, Taiwan National Science Council Grant NSC101-3114-Y-002-003, and the Project "Patient@home – Innovative Welfare Technology for the 21st Century" funded by the Danish Agency for Science, Technology and Innovation. The 2012 i2b2 NLP challenge is supported by the US National Institutes of Health Grants NIH NLM 2U54LM008748 (PI: Isaac Kohane) and NIH NLM 1R13LM011411-01 (PI: Ozlem Uzuner).

References

[1] Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc 2013.
[2] Mazur P, Dale R. The Dante temporal expression tagger. In: Vetulani Z, Uszkoreit H, editors. Human language technology. Challenges of the information society, vol. 5603. Berlin/Heidelberg: Springer; 2009. p. 245–57.
[3] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552–6.
[4] Klimov D, Shahar Y, Taieb-Maimon M. Intelligent visualization and exploration of time-oriented data of multiple patients. Artif Intell Med 2010;49(1):11–31.
[5] Ebadollahi S, Sun J, Gotz D, Hu J, Sow D, Neti C. Predicting patient's trajectory of physiological data using temporal trends in similar patients: a system for near-term prognostics. AMIA Annu Symp Proc 2010;2010:192–6.
[6] Post AR, Harrison JH. Protempa: a method for specifying and identifying temporal sequences in retrospective data for patient selection. J Am Med Inform Assoc 2007;14(5):674–83.
[7] 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.
[8] Alonso O, Gertz M, Baeza-Yates R. On the value of temporal information in information retrieval. SIGIR Forum 2007;41(2):35–41.
[9] Verhagen M, Saurí R, Caselli T, Pustejovsky J. Semeval-2010 task 13: tempeval-2. In: Proceedings of the 5th international workshop on semantic evaluation. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 57–62.
[10] 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. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 284–91.
[11] 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. Stroudsburg, PA, USA: Association for Computational Linguistics; 2010. p. 276–83.
[12] Kovačević 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.
[13] 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.
[14] Klein D, Manning CD. Accurate unlexicalized parsing. In: Proceedings of the 41st annual meeting on association for, computational linguistics; 2003. p. 423–30.
[15] Aronson AR, Lang F-M. An overview of metamap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17(3):229–36.
[16] Xu Y, Tsujii J, Chang EI-C. Named entity recognition of follow-up and time information in 20000 radiology reports. J Am Med Inform Assoc 2012;19(5):792–9.
[17] McCallum A. Mallet: a machine learning for language toolkit; 2002. Available at: http://mallet.cs.umass.edu.
[18] Sang EFTK, Veenstra J. Representing text chunks. In: Proceedings of the ninth conference on European chapter of the association for computational linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics; 1999. p. 173–9.
[19] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001;34(5):301–10.
[20] Agarwal S, Yu H. Biomedical negation scope detection with conditional random fields. J Am Med Inform Assoc 2010;17(6):696–701.
[21] Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc 2010;17(5):524–7.
[22] Chang C-C, Lin C-J. Libsvm: a library for support vector machines. ACM Trans Intell Syst Technol 2011;2(3):1–27.
[23] Strötgen J, Gertz M. Temporal tagging on different domains: challenges, strategies, and gold standards. In: Proceedings of the eight international conference on language resources and evaluation (lrec'12). Istanbul, Turkey: European Language Resources Association (ELRA); 2012. p. 3746–53.
[24] Hripcsak G, Zhou L, Parsons S, Das AK, Johnson SB. Modeling electronic discharge summaries as a simple temporal constraint satisfaction problem. J Am Med Inform Assoc 2005;12(1):55–63.