A hybrid system for temporal information extraction from clinical text

Buzhou Tang,[1,2] Yonghui Wu,[1] Min Jiang,[1] Yukun Chen,[3] Joshua C Denny,[3,4] Hua Xu[1,3]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, Texas, USA
[2]Department of Computer Science, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong, China
[3]Department of Biomedical Informatics, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA
[4]Department of Medicine, Vanderbilt University, School of Medicine, Nashville, Tennessee, USA

Correspondence to
Dr Hua Xu, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, USA; hua.xu@uth.tmc.edu

Received 9 January 2013
Revised 11 March 2013
Accepted 18 March 2013
Published Online First
9 April 2013

ABSTRACT

Objective To develop a comprehensive temporal information extraction system that can identify events, temporal expressions, and their temporal relations in clinical text. This project was part of the 2012 i2b2 clinical natural language processing (NLP) challenge on temporal information extraction.

Materials and methods The 2012 i2b2 NLP challenge organizers manually annotated 310 clinic notes according to a defined annotation guideline: a training set of 190 notes and a test set of 120 notes. All participating systems were developed on the training set and evaluated on the test set. Our system consists of three modules: event extraction, temporal expression extraction, and temporal relation (also called Temporal Link, or 'TLink') extraction. The TLink extraction module contains three individual classifiers for TLinks: (1) between events and section times, (2) within a sentence, and (3) across different sentences. The performance of our system was evaluated using scripts provided by the i2b2 organizers. Primary measures were micro-averaged Precision, Recall, and F-measure.

Results Our system was among the top ranked. It achieved F-measures of 0.8659 for temporal expression extraction (ranked fourth), 0.6278 for end-to-end TLink track (ranked first), and 0.6932 for TLink-only track (ranked first) in the challenge. We subsequently investigated different strategies for TLink extraction, and were able to marginally improve performance with an F-measure of 0.6943 for TLink-only track.

INTRODUCTION

Temporal information extraction (TIE) is a challenging area of natural language processing (NLP) research and is an important component of many NLP systems, such as question answering, document summarization, and machine translation. For clinical NLP systems that process medical narrative data, accurate recognition and interpretation of the timing of medical events is crucial for many medical reasoning tasks. To generate a complete timeline of medical events of a patient, a clinical TIE system must be able to identify events (eg, medical concepts), temporal expressions (eg, dates associated with events), and temporal relations between two events. Although significant efforts have been devoted to the representation, annotation, and extraction of temporal information in the general English domain (eg, the TimeML framework[1]), the performance of the state-of-the-art TIE systems is still not ideal (F-measures around 60–70%).[2 3] Moreover, extracting temporal information from clinical text can be more challenging than general English texts due to lack of formalism in writing quality.

To accelerate TIE research in the medical domain, the 2012 Informatics for Integrating Biology and Beside (i2b2) clinical NLP challenge focused on extraction of temporal information from hospital discharge summaries.[4] The challenge consists of three sub-tasks: (1) clinical event extraction with relevant attributes (eg, polarity) in clinic text; (2) temporal expression extraction, which requires both identification and normalization of text strings indicating date, time, and duration; and (3) temporal relation extraction, which determines if a temporal link ('TLink') exists between two events, two times, or one event and one time, and what type (*before*, *after*, or *overlap*) of TLink it is. The TLink extraction task was further divided into two tracks: 'end-to-end' (using system generated events and temporal expressions) and TLink-only (using gold standard events and temporal expressions). In this paper, we describe our TIE system developed for the i2b2 challenge. It is a comprehensive pipeline-based system that addressed all sub-tasks and was the top-ranked system in the i2b2 challenge in both the end-to-end TLink and TLink-only evaluations.

BACKGROUND

In the general English domain, many investigators have studied TIE from natural-language text corpora, such as newswires. TIE work began primarily with temporal representation in the 1980s. An important work was the interval-based algebra for representing temporal information in natural language, proposed by Allen in 1983.[5] Many early studies adopted Allen's representation, which promptly became a standard. In the 1990s, the widespread development of large annotated text corpora for NLP advanced TIE research dramatically. Community-wide information extraction tasks started to include TIE tasks. The message understanding conferences (MUCs) sponsored by the US government organized two consecutive temporal-related tasks: MUC-6 (1995)[6] and MUC-7 (1998).[7] In MUC-6, extracting absolute time information (ie, extracting exactly-specified times in the text) was a part of a general named entity recognition (NER) task. In MUC-7, the TIE task was expanded to include extraction of relative times also. These two tasks defined the Timex tags, which interpret time expressions into a normalized ISO standard form through the TIDES Timex2 guidelines.[8 9] In 2004, extracting and normalizing temporal expressions according to the Timex2 guidelines for both English and Chinese texts was part of the Time

To cite: Tang B, Wu Y, Jiang M, *et al. J Am Med Inform Assoc* 2013;**20**: 828–835.

Expression Recognition and Normalization Evaluation challenge, sponsored by the Automatic Content Extraction program.[10] These tasks provided preliminary but valuable contributions to TIE research.

Rapid development of TIE methods started in 2004 with the work of TimeML,[1] a robust specification language for events and temporal expressions in natural language. The TimeML schema mainly integrates two annotation schemes: TIDES (Translingual Information Detection, Extraction, and Summarization) TIMEX2 and STAG (Sheffield Temporal Annotation Guidelines).[11 12] It defined three elements of temporal information: events, temporal expressions, and temporal relations. Events, including verbs, adjectives, and nominals, corresponding to events and states are classified into different types, and have various attributes, including tense, aspect, and other features. Temporal expressions are token sequences that denote times with various attributes such as their normalized values. TimeML also represents temporal relations between events/times using an Allen-like format. It defines temporal relations using three types of links: TLinks (Temporal Links), SLinks (Subordination Links), and ALinks (Aspectual Links). TimeML has become an ISO standard for temporal annotation. Several TimeML-based annotated corpora have been created. The popular corpora include TimeBank1.2, AQUAIN, TempEval, and TempEval2. Among them, the TempEval corpus, based on TimeBank1.2, was created for the temporal relation task at TempEval1 in 2007.[2] For the Tempeval2 task in 2010, a multilingual corpus was created.[3 13] Detailed information about these corpora can be found at http://www.timeml.org/site/timebank/timebank.html. Many TIE systems have been developed based on these available corpora.[13]

Both machine learning and rule-based methods have been applied to TIE in the general English domain. For event extraction, machine learning methods widely used in NER have been adopted and have demonstrated good performance, including conditional random fields (CRFs) and supported vector machines (SVMs).[13] For temporal expression extraction, both machine learning and rule-based methods were investigated in TempEval2; in this test, rule-based methods slightly outperformed machine learning based methods.[13] All systems in TempEval2 identified temporal expressions attributes using rule-based methods.[13] HeidelTime, an open source system for temporal expression extraction, is a representative rule-based system that performed well in TempEval2.[14] Temporal relation extraction is typically divided into different sub-tasks. For example, in TempEval2, TLinks were divided into three different types: (1) TLinks between event and documentation time; (2) TLinks between events/times within the same sentence; and (3) TLinks between events/times across sentences. Both machine learning based or rule-based methods were used for different sub-tasks in TempEval2. To date, performance of temporal relation extraction systems has been less than optimal—the best system in TimeEval2 competition achieved F-measures of 82%, 65%, and 58% on three types of TLinks.[3] More recently, researchers have investigated methods that can integrate constraints among TLinks from all sub-tasks to further improve TIE performance. For example, Naushad *et al*[15] used Markov Logic networks to model the constraints in all TLinks and showed improved performance.

Temporal information is important for many medical applications. A number of studies[16–27] have addressed various topics of temporal representation and reasoning with medical data. Processing temporal events in medical text, however, has not been extensively studied. A few studies have developed different methods to extract temporal expressions from clinical narratives.[16 17 22] For example, Reeves *et al* extended the open-source temporal awareness and reasoning systems for question interpretation (TARSQI) toolkit, originally developed from news reports, to extract temporal expressions from veterans affairs (VA) clinical text. They found that temporal expressions in clinic notes were very different from those in the newswire domain, and the out-of-the-box implementation of the TARSQI toolkit performed poorly.[22] Some existing clinical NLP systems, such as ConText[25] and MedLEE,[26] also have the capability to recognize certain temporal expressions and link them to clinical concepts. More comprehensive systems such as developed by Zhou *et al*,[16 19 27] can not only extract temporal expressions associated with medical events, but also reason about temporal information in clinical narrative reports. For more details of studies in clinical TIE, see the review paper by Zhou and Hripcsak.[27] Nevertheless, very few studies have investigated the use of TimeML in the medical domain. Recent studies by Savova *et al*[17 21] have annotated clinical text using TimeML.

To advance the TIE research in the medical domain, organizers of the 2012 i2b2 clinical NLP challenge prepared an annotated clinical corpus based on TimeML and organized a clinical TIE competition similar to the TimeEval2 competition. The 2012 i2b2 challenge consisted of three subtasks: (1) *Event extraction*: six types of clinical events were extracted for the i2b2 challenge, including medical problems, tests, treatments, clinical departments, evidentials, and occurrences. Every event also has two attributes: polarity and modality. The polarity attribute marks whether an event is positive or negative, and the modality attribute is used to describe whether an event actually occurred or not. (2) *Temporal expression extraction*: the TIMEX3 tag was used to annotate temporal expressions, which has three main attributes: type (date/time/duration/frequency), value (normalized value of the TIMEX3), and modifier of a value (more, less, approximate, and so on). (3) *Temporal relation (TLink) extraction*: in this task, systems identified relations between events and times, and determined the type of relation. Three relation types (*before*, *overlap*, and *after*) were used in this challenge, as a simplification of the 13 more detailed ones specified in TimeML (simultaneous, before, after, immediately before, immediately after, including, being included, during, beginning, begun by, ending, identity, set/subset). All TLinks were further divided into three categories: (1) TLinks between events and section times (eg, admission or discharge time); (2) TLinks between events/times within one sentence; and (3) TLinks between events/time across sentences (eg, co-referenced entities).

METHODS

Figure 1 shows an overview of our clinical TIE system for the 2012 i2b2 NLP challenge. It consists of three components: event extraction, temporal expression extraction, and TLink extraction. Our TLink extraction module was further divided into three sub-classifiers for TLinks between events and section times, TLinks within one sentence ('sentence-internal'), and TLinks across sentences ('sentence-across'). Detailed descriptions are presented below.

Dataset

In the i2b2 challenge, organizers manually annotated 310 clinic notes by following the annotation guideline, of which 190 notes were used as the training set, and the remaining 120 notes were used as the test set. The number of events, temporal expressions, and TLinks, respectively, were (16 468; 2368; 33 781) in the training set and (13 594; 1820; 27 736) in the test set.

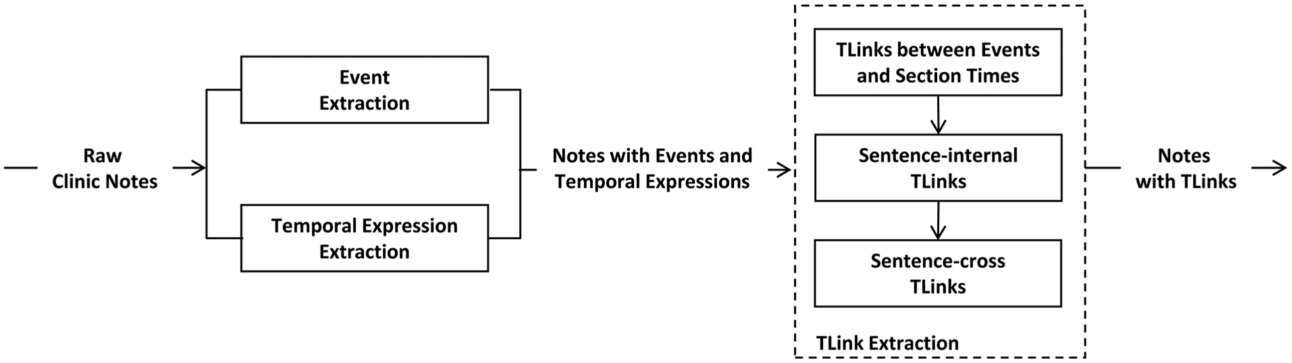

Figure 1 Overview of our clinical temporal information extraction system for the 2012 i2b2 natural language processing challenge.

Event extraction

The 2012 i2b2 event extraction task was very similar to the 2010 i2b2 NLP challenge for NER. There were two differences: (1) the event types were expanded from three in the 2010 challenge to six in this challenge; and (2) the assertion task in the 2010 challenge was replaced by two sub-tasks: polarity and modality. We participated in the 2010 challenge and have successfully developed a clinical NER tagger[28] that was ranked second in the 2010 challenge. In this study, we leveraged the previous system and developed a cascaded classification system for extracting events and their attributes. Figure 2 shows the architecture of the cascaded classifier for event extraction. For entity recognition, we developed a two-stage classifier based on CRFs: the first-stage classifier recognized medical problems, tests, and treatments events; the second-stage classifier recognized the remaining three types of events: clinical departments, occurrences, and evidentials. The results of the first-stage classifier were used as additional features for the second stage. This design allowed us to utilize the existing annotated corpus (826 notes) from the 2010 challenge for medical problems, tests, and treatments. After using our CRF classifiers to recognize events, SVM-based classifiers were used to assign polarity and modality for each event. We used similar features for all classifiers as described in the previous study.[29]

Temporal expression extraction

We developed a rule-based system in Python to extract temporal expressions, normalize their values, and identify the modifier attributes simultaneously, based on predefined regular expressions. We adopted many of the rules from the Java-based HeidelTime system, the best performing system in TimeEval2. The following is an example showing how the system works. In the sentence 'She was admitted on the morning of Feb 25, 2009', our system first extracts 'the morning of Feb 25, 2009' as a temporal expression with a value of 'Feb 25, 2009' and the modifier word 'morning' by the rule 'the (%PartOfDate) (%WordMonth %DigitDay, %FourDigitYear)', where PartofDate is a lexicon of all terms representing parts of dates (eg, 'morning', 'afternoon', and 'evening'), WordMonth is a lexicon of all English words representing months (eg, 'Jan', 'January', 'February'), DigitDay is a lexicon of valid calendar day numbers (ie, 1–31), and FourDigitYear is a lexicon or regular expression for valid four-digit year numbers (eg, '1984', '2009'). The value 'Feb 25, 2009' is then normalized to '2009-02-25,' and 'morning' is mapped to the modifier 'START'. Finally, a temporal expression is successfully extracted as follows: TIMEX3='the morning of Feb 25, 2009' 10:3 10:8||type='DATE'||val='2009-02-25'||mod='START'. One difference between our system and HeidelTime was that the post-processing in our system was a separate module, while that of HeidelTime was integrated into recognition rules.

TLink extraction

As mentioned above, we divided the TLink extraction task into three sub-tasks:

1. *TLinks between events and section time*: In this challenge, each note has two section times: admission time and discharge time. For each event, we need to identify the TLink type (Before, Overlap, and After) between the event and its corresponding section time (admission or discharge time). We developed two classifiers for determining event-admission and event-discharge time, respectively.
2. *TLinks between events/times within one sentence (sentence-internal TLinks)*: Those are TLinks found between two events, two times, or one event and one time that are all located in the same sentence. We developed two classifiers for this TLink type: (1) a classifier for TLinks between two events; and (2) a classifier for TLinks between one event and one time. TLinks between two times were very rare and were ignored by our algorithm.
3. *TLinks between events/times across sentences (sentence-across TLinks)*: This subtask is potentially the most challenging. In our review of the training data, we noticed two primary types of sentence-across TLinks: (1) TLinks

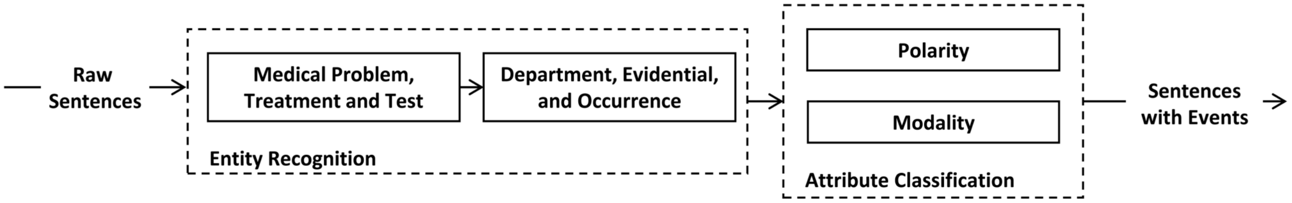

Figure 2 Architecture of the event extraction system.

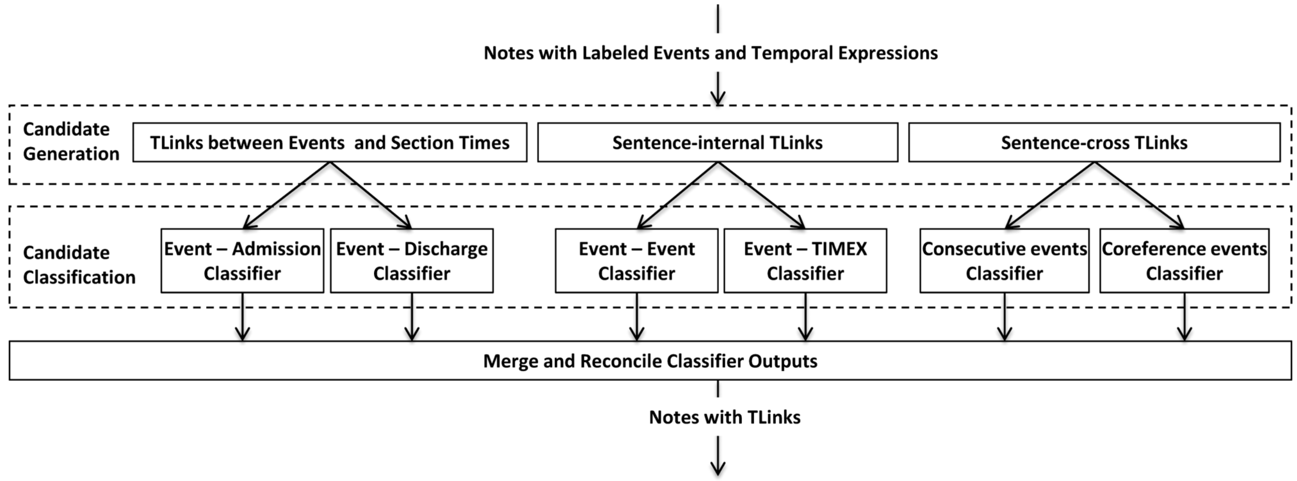
Notes with Labeled Events and Temporal Expressions
Candidate Generation
TLinks between Events and Section Times
Sentence-internal TLinks
Sentence-cross TLinks
Candidate Classification
Event – Admission Classifier
Event – Discharge Classifier
Event – Event Classifier
Event – TIMEX Classifier
Consecutive events Classifier
Coreference events Classifier
Merge and Reconcile Classifier Outputs
Notes with TLinks

Figure 3 Architecture of the TLink extraction system, a three-layer cascaded classification system.

between main events in consecutive sentences; and (2) TLinks between events that are co-referenced. We trained a classifier for each type of sentence-across TLink.

The i2b2 TLink extraction task did not specify the candidate pairs for TLinks (as did TempEval2). Thus, in theory, any two events/times can be a candidate pair to train a classifier. This approach would generate a huge number of candidate TLinks and may not be ideal for training a classifier. Therefore, selection of 'likely' TLink candidates from all possible pairs became very important to a high-functioning algorithm. We called this procedure 'TLink candidate generation'. This step was executed before training the classifier. Once candidate TLinks were selected, we built classifiers for each of the three categories of TLinks described above. More specifically, we built a cascaded classification system in which outputs of a classifier for one type of TLink were supplied as input features for the classifier of next category. Finally, all TLinks from each category were combined to generate the final output. Any conflicts among TLinks from different categories were resolved in a final, merging step. Figure 3 shows the architecture of the complete TLink extraction system. More details of TLink candidate generation, candidate classification, and merging are described below.

TLink candidate generation

The strategies for candidate generation were different for each of the three categories of TLinks described above. For TLinks between events and section times, we included all (event, section time) pairs as candidates. For TLinks within one sentence, any candidate pair that met one of the following criteria was included: (1) any consecutive event/time pair in the sentence; or (2) any event/time pair that has a dependency relation, based on the dependency parse tree of the sentence from the Stanford Parser.[30] Figure 4 shows an example of dependency relations of words in one sentence, where events are in bold, italics and underlined. The consecutive event/time pairs {(transferred, here), (here, further mgmt of airway), (further mgmt of airway, large thyroid mass)} were selected as candidates. Moreover, there are three dependency-based candidate pairs: (1) between 'transferred' and 'here', due to a dependency path from 'here' to 'transferred'; (2) between 'further mgmt of

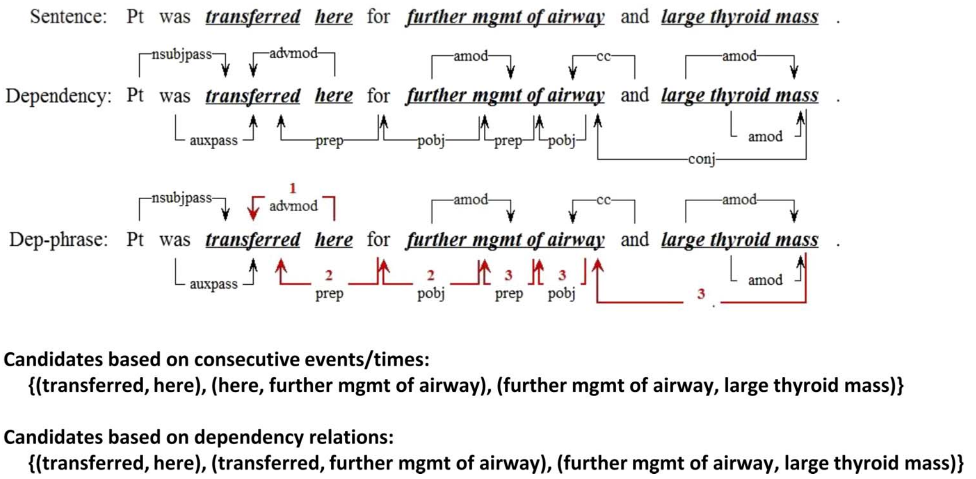
Sentence: Pt was transferred here for further mgmt of airway and large thyroid mass .
nsubjpass
advmod
amod
cc
amod
Dependency: Pt was transferred here for further mgmt of airway and large thyroid mass .
auxpass
prep
pobj
prep
pobj
amod
conj
1
nsubjpass
advmod
amod
cc
amod
Dep-phrase: Pt was transferred here for further mgmt of airway and large thyroid mass .
auxpass
2
prep
2
pobj
3
prep
3
pobj
3
amod
Candidates based on consecutive events/times:
{(transferred, here), (here, further mgmt of airway), (further mgmt of airway, large thyroid mass)}
Candidates based on dependency relations:
{(transferred, here), (transferred, further mgmt of airway), (further mgmt of airway, large thyroid mass)}

Figure 4 An example of candidate selection for sentence-internal TLinks. Consecutive entities or entities connected by a dependency path were selected as candidates. In this sentence, the consecutive events/times pairs included (transferred, here), (here, further mgmt of airway), and (further mgmt of airway, large thyroid mass); and dependency pairs included (transferred, here), (transferred, further mgmt of airway), and (further mgmt of airway, large thyroid mass). After removing duplicates, the final TLink candidates from this sentence were (transferred, here), (here, further mgmt of airway), (further mgmt of airway, large thyroid mass), and (transferred, further mgmt of airway).

Pt ***reports*** that he noticed ***a right neck mass*** last October.

...

Pt was ***transferred*** ***here*** for ***further mgmt of airway*** and ***large thyroid mass*** .

IP found ***endotracheal obstructing tumor*** .

...

Candidates between main events:

{(transferred, endotracheal obstructing tumor), (large thyroid mass), (further mgmt of airway, endotracheal obstructing tumor)}

Candidates based on co-reference:

{(a right neck mass, large thyroid mass)}

Figure 5 An example of generating TLink candidates across sentences, based on main events or co-references.

airway' and 'transferred' because of a dependency path from 'mgmt' to 'transferred' (mgmt → for → transferred); (3) between 'large thyroid mass' and 'further mgmt of airway' because of a dependency path from 'mass' to 'airway'. These dependent pairs ([transferred, here], [transferred, further mgmt of airway] and [further mgmt of airway, large thyroid mass]) are also selected as candidates.

Two strategies were used to generate candidates for sentence-across TLinks. The first dealt with candidate TLinks between 'main' events in consecutive sentences, which we defined as the first and last events in each sentence. For two consecutive sentences, we would collect all possible pairs between first and last events in both sentences as candidates. For the example in figure 5, the second sentence has 'transferred' as the first event and 'large thyroid mass' as the last event and the third sentence only has one event 'endotracheal obstructing tumor' as both the first and last event. Therefore we would generate two candidate pairs (transferred, endotracheal obstructing tumor) and (large thyroid mass, endotracheal obstructing tumor) from these two sentences. The second strategy is for co-references across multiple sentences. We defined a simple rule for identifying possible pairs based on co-references: any two events with the same semantic type and share the same head noun were treated as a potential co-reference pair. For the example in figure 5, the pair (a right neck mass, large thyroid mass) would be a candidate based on co-references because of the same head noun 'mass'.

TLink candidate classification

After candidate TLinks were selected, the next step was to link them with appropriate labels (Before, Overlap, After, or None) based on the gold standard to form the training matrix for building the machine learning classifier. A straightforward method is to label candidate TLinks without any expansion—we assign the candidates that appear in the gold standard with their corresponding TLink types, and assign 'None' to all others. However, our evaluation on the training set showed that this approach performed poorly, because a large number of implicit TLinks that could be inferred from existing TLinks were labeled as 'None'. In order to take implicit TLinks into account, we expanded candidate TLinks by calculating the closure sets of all TLinks in one document. We expanded both the candidate TLinks and the gold standard TLinks and then assigned corresponding labels. The closure set was calculated based on the rules listed in table 1, where an alphabetic letter in {A, B, C, …} represents an event or a temporal expression, '→', '←', and '↔' represent 'Before', 'After', and 'Overlap' relations, respectively. A→B, A←B, and A↔B represent TLinks of 'A Before B', 'A After B', and 'A Overlap B', respectively.

After candidate TLinks were expanded and assigned with corresponding labels based on the expanded gold standard, we extracted the following features for different TLink classifiers.

TLinks between events and section time:

- Event position information: The position information is useful for events at the beginning or end of sections. For example, the events at the end of discharge section usually mention treatments after the discharge time. Specifically, we noted whether the event was: (1) in the first/last five sentences, (2) one of the first/last three events in its section, and (3) one of the first/last five events in the note.
- Bag-of-words: Treating each event as a word, the unigrams, bigrams, and trigrams of context within a window of [−2, 2] were extracted by our system.
- Part-of-speech (POS) tags: The unigrams, bigrams, trigrams of POS within a window of [−2, 2].
- Tense: Verbs appearing in the sentence containing the event were used to represent the tense of the sentence.
- Dependency-related features: The presence (or absence) of a 'time' or 'date' dependent on the event, the dependency relation type, and the order between it and the section time,
- Time-related features: The presence (or absence) of any 'time' or 'date' in the sentence containing the event, and the order between the nearest 'time' or 'date' and the section time.
- Event-related features: All attributes of the event, and whether or not the event contains a verb.

TLinks between events/times within one sentence used the above features plus several additional ones:

- Dependency-related features: Whether there is a dependency relation between two entities of the TLink? What is the dependency relation? The path of word and the path of POS in this relation.
- Distance: The distance between two candidate events in number of words, and the count of the number of events/temporal expressions between two candidates.
- Conjunction: The conjunctions between two entities of the TLink.

Table 1 Rules for TLink expansion

Unary formals	Binary formals
If A→B, then B←A	If A→B and B→C, then A→C
If A←B, then B→A	If A→B and B↔C, then A→C
If A↔B, then B↔A	If A→B and A↔C, then C→B
	If A↔B and B↔C, then A↔C

Table 2 Our system's best results reported in the i2b2 challenge

Task	Precision	Recall	F-measure	Type	Polarity	Modality	Value	Modifier
Event	0.9374	0.8679	0.9013	0.8360	0.8478	0.8312	–	–
TIMEX	0.8296	0.9055	0.8659	0.8500	–	–	0.7000	0.8462
TLink End-to-end	0.7006	0.5688	0.6278	–	–	–	–	–
TLink-only	0.7143	0.6733	0.6932	–	–	–	–	–

TIMEX, temporal expression extraction.

- TLinks between the events and section times, as determined by the previous step.

TLinks between events/times across sentences:

- For main events in two consecutive sentences, the following features were used: the presence of times/dates in the same sentence as the event, the words in each event, verb tense, and attributes of each event as identified in the earlier stages in the algorithm.
- For co-referenced events, the following features were used: the token length of each event, the number of overlapped tokens between the two candidate events, whether an event contains determiners such as 'his', the line distance of two events, whether the last tokens of two events match, the semantic type of an event, whether two events contain the same positional words such as 'left' and 'right', and whether two events contain the same anatomic words such as 'arm' and 'leg'.

We used CRF++ (http://crfpp.googlecode.com/svn/trunk/doc/index.html) as the implementation of CRFs, and LIBLINEAR (http://www.csie.ntu.edu.tw/~cjlin/liblinear/) as the implementation of SVMs. Parameters of all classifiers were optimized by 10-fold cross-validation on the training dataset.

Merging TLinks

Conflicts exist when merging TLinks generated from different classifiers. In this step, we defined a simple rule to resolve conflicts. Our assumption was that TLinks from event-section times and same sentences were more reliable than those from sentence-across TLink classifiers. Therefore, any sentence-across TLinks that contradicted TLinks in the closure set from the other two categories were removed from the final output.

Evaluation

All our evaluations were performed on the independent test dataset using the evaluation scripts provided by the i2b2 organizers. The evaluation programs for time/event extraction outputted micro-average precision, recall, and F-measure. In addition, accuracy was also reported for attributes. For TLink extraction, the performance was measured by precision, recall, and F-measure based on the reduced graph of all TLinks. Participating i2b2 teams had two options for submitting TLink results: end-to-end and TLink-only. For the end-to-end track, participating systems take raw clinical text as the input and generated TLinks using only the events and temporal information identified by their algorithms. For the TLink-only track, manually annotated events and temporal expressions from the gold standard annotations were used as the inputs of a system. In addition to our results reported in the i2b2 challenge, we also conducted additional experiments to assess the contributions of different strategies that we developed for TLink extraction using the test dataset.

RESULTS

For each track, a participating team could submit three runs. The results from the best run were used to rank participating systems in the challenge. Table 2 shows our best results for different sub-tasks in the challenge, as reported by the i2b2 challenge organizers. For event extraction, our best F-measure was 0.9013, ranked second among all participating teams. The corresponding accuracies for three event attributes: type, polarity, and modality were 0.8360, 0.8478, and 0.8312, respectively. For temporal expression extraction, our best run achieved an F-measure of 0.8659, with accuracies of 0.8500, 0.7000, and 0.8562 for type, value, and modifier attributes. The corresponding primary score was 0.6061 for temporal expression extraction, which was ranked fourth in the challenge. For TLink extraction, our system was ranked first for both end-to-end and TLink-only tracks, with F-measures of 0.6278 and 0.6932, respectively.

Table 3 shows the systematic evaluation of contributions of the different TLink extraction strategies used by our algorithm for the TLink-only track. We started with TLinks between

Table 3 Evaluation of contributions of different TLink extraction strategies for TLink-only track on test dataset

Setting	Precision	Recall	Average P&R	F-measure
Baseline	0.8839	0.2754	0.5353	0.4199
Baseline+consecutive	0.7855	0.5562	0.6825	0.6513
Baseline+dependency	0.8317	0.3951	0.6097	0.5357
Baseline+consecutive+dependency	0.7789	0.5664	0.6845	0.6558
Baseline+consecutive+dependency+main-event	0.7030	0.6615	0.6865	0.6816
Baseline+consecutive++dependency+co-reference	0.7030	0.6614	0.6865	0.6816
Baseline+consecutive+dependency+main-event+co-reference	0.7143	0.6733	0.6982	0.6932
Baseline+consecutive+dependency+main-event+co-reference+merge	0.7227	0.6681	0.7011	0.6943

This table shows the contributions of different TLink strategies on the test dataset, where 'baseline', 'consecutive', 'dependency', 'main-event', and 'co-reference' denote TLinks between events and section times, sentence-internal TLinks based on consecutive events/times or dependency relations, sentence-across TLinks between main events in consecutive sentences or based on co-references, respectively.

events and section times as a baseline, and sequentially added TLinks from different sentence-internal (including consecutive and dependency pairs), sentence-across TLinks (including main events and co-reference pairs) strategies, and the TLink merging step. Each of the TLink strategies we employed improved the F-measure of our system. When sentence-internal TLinks ('consecutive + dependency') were added, the F-measure increased from 0.4199 to 0.6558. When two types of sentence-across TLinks ('main-event + co-reference') were further added to the system, the F-measure improved to 0.6932. The merging step also slightly improved the F-measure. The best performance of the system in this experiment was 0.6943, which was slightly higher than our best result reported in the challenge (0.6932). The increase in F-measure was due to the increase in recall; the system's precision actually decreased when different types of TLinks were added. However, the merging step improved precision instead of recall.

DISCUSSION

We developed a hybrid TIE system for clinical text by combining rule-based and machine learning based approaches. We participated in all three tracks of the 2012 i2b2 NLP challenge using this system. Evaluation using the independent test dataset by the challenge organizers showed that our system achieved results among the best in all three tracks: F-measures of 0.8659 for temporal expression extraction (ranked fourth), 0.6278 for end-to-end TLink track (ranked first), and 0.6932 for TLink only track (ranked first). Further systematic evaluation showed that our ad hoc strategies for temporal relation extraction were beneficial to the task, indicating the success of our approaches.

Our system had a lower F-measure on temporal expression detection (0.8659) and a lower accuracy on the value attribute (0.7000) when compared with the top system in the challenge (F-measure 0.9003 and value accuracy 0.7291). The difference was mainly caused by lower recall—our system missed some temporal expressions in the test dataset. Overall, the best clinical TIE system performed much worse than the best system of TempEval2[3] on attribute identification, especially on the value attribute (accuracy: 73% vs 85%). We found that many errors associated with value normalization were related to inappropriate inferences. For example, determination of the value of the date expression 'three days after operation' requires knowing the operation date/time, which may also be an explicit time expression. Our current strategy for implicit date/time inference was very limited—we just used nearest explicit date/time by default. If no date/time can be found nearby, we then set the value to the corresponding section time. Take the date expression 'three days after operation' as an example again, if the date '2008-09-01' is the nearest date before it, its value is set to '2008-09-04' (2008-09-01 + three days, 2008-09-01 is recognized as the date of 'operation'). This experiment demonstrated that this method for implicit date/time could be improved upon. In addition, we also noticed some annotation errors for temporal expression attributes in the gold standard set. For example, 'two days ago', which should be a date, was annotated as a duration with the value 'P2D', for '2/25–2/27/00', which should be a duration with the value 'P2D', it was separated into two dates '2/25' (2000-02-05) and '2/27/00' (2000-02-07) in the gold standard. Such errors highlight the challenge in creating a human annotation set.

As explained in the Methods, generation of a reduced set of candidate pairs was important for TLink extraction in this challenge. We proposed different strategies for TLink within one sentence and across multiple sentences, which likely was a component of our system's performance compared with other i2b2 systems. To further assess the usefulness of our candidate generation strategies, we conducted an additional experiment to study the upper-bound performance of our methods on the training dataset. We collected all TLink candidates generated by our strategies and assigned correct relation types (based on the gold standard) to them. We then compared these results with the gold standard using the evaluation program and we obtained the following precision, recall, and F-measure: 0.9772, 0.7640, and 0.8576, respectively. Compared to the results from the competition (TLink only F-measure of 0.6932), these are very promising. However, the difference between the precision and recall was large (0.9772 vs 0.7640). In the future, we plan to focus on the following two aspects for further improvement: (1) seek new strategies for TLink candidate generation, which may further improve recall; and (2) explore new algorithms for TLink classification to take advantage of constraints among TLinks.

CONCLUSIONS

In this study, we developed a hybrid clinical TIE system that can extract events, temporal expressions, and temporal relations from hospital discharge summaries. Our system used both rule-based and machine learning based approaches and competed in all three tracks in the 2012 i2b2 clinical NLP challenge on clinical TIE. The system achieved top-ranked F-measures of 0.6278 for the end-to-end track and 0.6932 for the TLink-only track, indicating promise for TIE from clinical text.

Contributors The work presented here was carried out in collaboration between all authors. BT and HX designed the methods and experiments. BT, YW, MJ, and YC carried out the experiments. BT, JCD, and HX analyzed the data, interpreted the results, and wrote the paper. All authors have attributed to, seen, and approved the manuscript.

Funding This study is supported in part by grants from the National Library of Medicine (NLM) R01-LM007995 and National Cancer Institute (NCI) R01CA141307. We also thank the 2012 i2b2 NLP challenge organizers who were supported by Informatics for Integrating Biology and the Bedside award number 2U54LM008748 from the NLM and the National Heart, Lung and Blood Institute (NHLBI), and by award number 1R13LM01141101 from the NLM.

Competing interests None.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1 Pustejovsky J, Castaño J, Ingria R, *et al*. {TimeML:} Robust specification of event and temporal expressions in text. *Fifth International Workshop on Computational Semantics {(IWCS-5)}*, Tilburg, The Netherlands, 2003: 1–11.
2 Verhagen M, Gaizauskas R, Schilder F, *et al*. The TempEval challenge: identifying temporal relations in text. *Lang Resources Eval* 2009;43:161–79.
3 Verhagen M, Sauri R, Caselli T, *et al*. SemEval-2010 task 13: TempEval-2. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Los Angeles, California: Association for Computational Linguistics, 2010: 57–62.
4 Sun W, Rumshisky A, Uzuner O. Evaluating Temporal Relations in Clinical Text: 2012 i2b2 Challenge Overview. *J Am Med Inform Assoc* 2013;20:806–13.
5 Allen JF. Maintaining knowledge about temporal intervals. *Commun ACM* 1983;26:832–43.
6 Sundheim BM. Overview of results of the MUC-6 evaluation. *Proceedings of the 6th conference on Message Understanding*. Columbia, Maryland: Association for Computational Linguistics, 1995: 13–31.
7 Chinchor N. Overview of MUC-7/MET-2. *Proceedings of the 7th Message Understanding Conference*, Fairfax, Virginia, 1998.
8 Mani I, Wilson G, Ferro L, *et al*. Guidelines for annotating temporal information. *Proceedings of the First International Conference on Human Language Technology Research*. San Diego: Association for Computational Linguistics, 2001:1–3.
9 Ferro L, Kozierok R, Gerber L, *et al*. Annotating temporal information: from theory to practice. *Proceedings of the Second International Conference on Human Language Technology Research*. San Diego, California: Morgan Kaufmann Publishers, 2002: 226–30.

10 Doddington G, Mitchell A, Przybocki M, *et al*. {The Automatic Content Extraction (ACE) Program—Tasks, Data, and Evaluation}. *Proceedings of LREC 2004*, Lisbon, Portugal, 2004: 837–40.

11 Setzer A. *Temporal information in newswire articles: an annotation scheme and corpus study*. University of Sheffield, UK, 2001.

12 Setzer A, Gaizauskas R. A pilot study on annotating temporal relations in text. *Proceedings of the Workshop on Temporal and Spatial Information Processing—Volume 13*. Association for Computational Linguistics, Toulouse, France, 2001: 1–8.

13 Pustejovsky J, Verhagen M. SemEval-2010 task 13: evaluating events, time expressions, and temporal relations (TempEval-2). *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*. Boulder, Colorado: Association for Computational Linguistics, 2009: 112–16.

14 Strötgen J, Gertz M. HeidelTime: high quality rule-based extraction and normalization of temporal expressions. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Los Angeles, California: Association for Computational Linguistics, 2010: 321–4.

15 UzZaman N, Allen JF. TRIPS and TRIOS system for TempEval-2: extracting temporal information from text. *Proceedings of the 5th International Workshop on Semantic Evaluation*. Los Angeles, California: Association for Computational Linguistics, 2010: 276–83.

16 Zhou L, Friedman C, Parsons S, *et al*. System architecture for temporal information extraction, representation and reasoning in clinical narrative reports AMIA Annu Symp Proc 2005, Austin, TX, USA, 2005:869–873.

17 Savova G, Bethard S, Styler W, *et al*. Towards temporal relation discovery from the clinical narrative. *AMIA Annual Symposium Proceedings/AMIA Symposium AMIA Symposium 2009*, Washington DC, USA, 2009: 568–72.

18 Galescu L, Blaylock N. A corpus of clinical narratives annotated with temporal information. *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*. Miami, Florida, USA: ACM, 2012: 715–20.

19 Zhou L, Melton GB, Parsons S, *et al*. A temporal constraint structure for extracting temporal information from clinical narrative. *J Biomed Inform* 2006;39:424–39.

20 Gaizauskas R, Harkema H, Hepple M, *et al*. Task-oriented extraction of temporal information: the case of clinical narratives. *Proceedings of the Thirteenth International Symposium on Temporal Representation and Reasoning. IEEE Computer Society*, Washington DC, USA, 2006: 188–95.

21 Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010;17:507–13.

22 Reeves RM, Ong FR, Matheny ME, *et al*. Detecting temporal expressions in medical narratives. *Int J Med Inform* 2013;82:118–27.

23 Denny JC, Peterson JF, Choma NN, *et al*. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *J Am Med Inform Assoc* 2010;17:383–91.

24 Patnaik D, Butler P, Ramakrishnan N, *et al*. Experiences with mining temporal event sequences from electronic medical records: initial successes and some challenges. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Diego, California, USA: ACM, 2011: 360–8.

25 Chapman WW, Chu D, Dowling JN. ConText: an algorithm for identifying contextual features from clinical text. *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*. Prague, Czech Republic: Association for Computational Linguistics, 2007: 81–8.

26 Friedman C, Alderson PO, Austin JH, *et al*. A general natural-language text processor for clinical radiology. *J Am Med Inform Assoc* 1994;1:161–74.

27 Zhou L, Hripcsak G. Temporal reasoning with medical data—a review with emphasis on medical natural language processing. *J Biomed Inform* 2007;40:183–202.

28 Jiang M, Chen Y, Liu M, *et al*. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc* 2011;18:601–30.

29 Tang B, Cao H, Wu Y, *et al*. Clinical entity recognition using structural support vector machines with rich features. *ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics*, Maui, HI, USA, 2012: 13–20.

30 Cer D, De Marneffe M-c, Jurafsky D, *et al*. Parsing to stanford dependencies: trade-offs between speed and accuracy. *LREC*, Floriana, Malta, 2010.