

TEMPTING system: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries



Yung-Chun Chang^{a,b}, Hong-Jie Dai^{c,*}, Johnny Chi-Yang Wu^a, Jian-Ming Chen^a,
Richard Tzong-Han Tsai^d, Wen-Lian Hsu^a

^a Institute of Information Science, Academia Sinica, Nankang, Taipei, Taiwan, ROC

^b Department of Information Management, National Taiwan University, Taipei, Taiwan, ROC

^c Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei, Taiwan, ROC

^d Department of Computer Science and Information Engineering, National Central University, Taoyuan, Taiwan, ROC

ARTICLE INFO

Article history:

Received 14 March 2013

Accepted 11 September 2013

Available online 20 September 2013

Keywords:

Temporal relation extraction

Natural language processing

Text mining

Hybrid method

Maximum entropy

ABSTRACT

Patient discharge summaries provide detailed medical information about individuals who have been hospitalized. To make a precise and legitimate assessment of the abundant data, a proper time layout of the sequence of relevant events should be compiled and used to drive a patient-specific timeline, which could further assist medical personnel in making clinical decisions. The process of identifying the chronological order of entities is called temporal relation extraction. In this paper, we propose a hybrid method to identify appropriate temporal links between a pair of entities. The method combines two approaches: one is rule-based and the other is based on the maximum entropy model. We develop an integration algorithm to fuse the results of the two approaches. All rules and the integration algorithm are formally stated so that one can easily reproduce the system and results. To optimize the system's configuration, we used the 2012 i2b2 challenge TLINK track dataset and applied threefold cross validation to the training set. Then, we evaluated its performance on the training and test datasets. The experiment results show that the proposed TEMPTING (TEMPoral relaTion extractING) system (ranked seventh) achieved an *F*-score of 0.563, which was at least 30% better than that of the baseline system, which randomly selects TLINK candidates from all pairs and assigns the TLINK types. The TEMPTING system using the hybrid method also outperformed the stage-based TEMPTING system. Its *F*-scores were 3.51% and 0.97% better than those of the stage-based system on the training set and test set, respectively.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Event and temporal information extraction from plain text is an essential task in natural language processing (NLP) research; and it has been used for many NLP applications, such as document summarization and question answering [4]. Document summarization must identify the key events in one or more stories to yield the best summary with the least extraneous information. Question answering must be able to answer queries about dates, the duration of events, and even relative times in the form of natural language or a set query type. The possible domains for temporal information extraction are numerous and varied, but it is especially useful for processing patient records in the medical domain. Such records contain valuable temporal data that is usually in the form of free text. Information about the sequence of events in patient records plays an essential role in clinical research because it can help solve diverse problems; for example, finding the correlations between

treatments and outcomes or extracting information about adverse drug reactions. However, the knowledge that certain items are correlated may be worthless without knowing their temporal order. For example, if researchers are interested in the efficacy of the medicine vancomycin, realizing that the symptom of a disease changes after a patient takes vancomycin may be insufficient without learning that the change only occurs after three days of treatment. To improve the temporal extraction of patient information in the clinical domain, the Sixth Informatics for Integrating Biology and the Bedside (i2b2) NLP challenge [27] was held in 2012. The challenge involved automatically determining the temporal relations between events described in de-identified patient discharge summaries. Based on the TLINK, EVENT and TIMEX3 tags defined in TimeBank [17],¹ the challenge released a dataset annotated by

¹ TimeBank is a hand-annotated corpus that conforms to the TimeML specification [18], which is an ISO standard for the annotation of events and temporal expressions, as well as anchoring and ordering the relations between them. In TimeBank, the TLINK tag represents the temporal relationship between events (the EVENT tag) or between an EVENT and a time expression (the TIMEX3 tag).

* Corresponding authors. Fax: +886 2 2739 2914 (H.-J. Dai).

E-mail address: hjdai@tmu.edu.tw (H.-J. Dai).

Partners HealthCare and Beth Israel Deaconess Medical Center [22], and proposed three tracks: (1) EVENT/TIMEX3 track: recognize event and time expressions; (2) TLINK track: based on the given EVENT and TIMEX3 tags, identify their temporal relations; (3) End-to-End track: perform the above two tasks on raw discharge summaries. In the framework of the TLINK track, the actions, situations and descriptions in a discharge summary that are related to the patient's timeline are annotated with the EVENT tag. Phrases in the summary that provide time-relevant knowledge are annotated with the TIMEX3 tags. Temporal links or TLINKs indicate the relations between pairs of EVENTS and TIMEX3s; and the attribute “type” of a TLINK indicates how temporal objects are related to each other. A TLINK track system must establish a TLINK link between the involved entities. It must also state their attribute type explicitly; that is, indicate if they are: “before”, “after”, “includes”, “is included”, “holds”, “simultaneous”, “immediately after”, “immediately before”, “begins”, “ends”, and so on.

Compared to general relation extraction tasks, identifying temporal relations in patient discharge summaries is more strenuous, and has motivated a great deal of NLP research in recent years [12,19,23]. Continuing with the vancomycin example above, the recognition of temporal phrases (e.g. after . . . three days) and event entities (e.g. vancomycin) are clearly prerequisites for understanding and interpreting the statement. However, the further temporal knowledge and inferences are more important and must be derived. First, the temporal aspects of the properties of entities must be properly assigned. Second, event descriptions involving entities and their time stamps must be extracted. Finally, the temporal order of the involved entities has to be inferred. In this paper, we focus on the TLINK track and propose a hybrid TEMPoral relation extractING (TEMPING) system that combines rule-based and machine learning approaches to capture the above knowledge and inferences. Temporal relation extraction rules and the pseudocode for integrating the rule-based and machine learning results are formally described. To the best of our knowledge, many leading systems in i2b2 [6,24] do not consider some of the proposed rules or state them formally.

2. Material and methods

Our objective is to extract temporally related entity pairs e_i and e_j , and their temporal relation, r , from a text as a TLINK tuple $[e_i, r, e_j]$. Fig. 1 shows the system architecture of TEMPTING, which is comprised of two key subsystems: the *multi-stage rule-based temporal relation extraction system* (the rule-based system) and the *maximum entropy-based temporal relation extraction system* (the machine learning-based system). For each discharge summary d_k in $D = \{d_1, \dots, d_l\}$, all entities $E = \{e_1, e_2, \dots, e_m\}$ in d_k are first paired to form the candidate TLINK pair set $P = \{p_1, \dots, p_n\}$, where $p = \langle e_i, e_j \rangle$ is a TLINK pair constructed by the entities e_i and e_j . The candidate pair set P is then processed by both subsystems and the TLINKs they extract are fused by a TLINK integration algorithm to generate the final results.

In the rule-based system, we group several heuristic rules into three different stages to determine the type r of each p in P . Depending on the setting of the TLINK track, the type should be one of the following values: {“BEFORE”, “OVERLAP” and “AFTER”}. First, we apply the intra-sentence rules to extract the TLINKs in a sentence. Then, we exploit the inter-sentence extraction rules along with the co-reference and timeline concept to extract cross-sentence TLINKs. Finally, TLINKs are retrieved according to the clinical note sections, which are different portions of clinical notes that contain rich information about the patient during different stages, such as admission, treatment and discharge. These can be used to determine the TLINK types that each EVENT/TIMEX3 be-

longs to. In the Results section, we will show the performance of the proposed rules based on the combination of different stages.

In the machine learning-based system, an additional phase, TLINK detection, is implemented to filter out EVENT/TIMEX3 pairs that do not possess a TLINK. A supervised learning-based TLINK type classifier then predicts the type r of p . For each p , the feature extraction component extracts representative text features that are used by the TLINK detector/classifier. In the following subsections, we describe our two temporal relation extraction systems and the rules/features developed for the TLINK track.

2.1. Multi-stage rule-based temporal relation extraction system

The rule-based system is implemented as a series of rule-based models that exploit a number of linguistic rules to capture various temporal relation patterns. One or several deterministic rules represent a processing stage in the system. In the following subsections, we use four symbols to represent each rule: constants, variables, functions, and predicates. *Constants* represent objects in a discharge summary, such as a time entity “per day” or an event entity “levofloxacin”. *Variables* (e.g., i, j) range over the objects. Constants and variables may belong to specific types. *Predicates* represent the relationships between objects, or the attributes of objects. We use logical connectives and quantifiers to construct the rules recursively from constants or variables. The Boolean operations of logical conjunction, disjunction and negation are denoted by “ $\wedge$ ”, “ $\vee$ ”, and “ $\neg$ ” respectively; and the symbol “ $\Rightarrow$ ” means “implies”. For instance, $A \Rightarrow B$ means that if A is true, then B is also true; however, if A is false, nothing is said about B . Parentheses may be used to enforce the precedence.

2.1.1. Stage 1: intra-sentence TLINK extraction

This stage determines the default TLINK types of the TLINK pairs found in the same sentence. Fig. 2 shows an example where two OVERLAP TLINKs should be extracted from the sentence. One is constructed by two EVENT entities, e_1 and e_2 ; and the other is comprised of e_2 and the time expression (TIMEX3) entity t_1 . Note that we use different variable symbols to represent different types of entities. For EVENT entities, the variable et_i is used; and the t_i symbol is used for TIMEX3 entities.

In the training set of the i2b2 TLINK track, we observed that the probability of two EVENT entities in the same sentence having a TLINK type of OVERLAP is 61.8%. Therefore, in the first stage, we developed the following rule to assign the OVERLAP link between two event entities found in the same sentence.

R.1.1 $isTheSameSentence(et_i, et_j) \Rightarrow TLINK_{OVERLAP}(et_i, et_j)$

Furthermore, based on our analysis on the training set, we believe there is a high probability that an EVENT entity will establish an OVERLAP TLINK with the nearest co-occurring TIMEX3 in the same sentence. Accordingly, the following rule is used to construct the OVERLAP link between an EVENT entity and its nearest TIMEX3 entity.

R.1.2 $isTheSameSentence(e_i, t_j) \wedge isTheNearestTime(t_j, e_i) \Rightarrow TLINK_{OVERLAP}(e_i, t_j)$

2.1.2. Stage 2: Inter-sentence TLINK extraction

To discover the relationships between time expressions or events across sentences, we employ two strategies in this stage. The first uses the co-reference resolution model developed in our previous work [8] to uncover EVENTS that literally imply the same concept. The second exploits timeline-dependent information to determine the type of the TLINK between the EVENT and TIMEX3 entities.

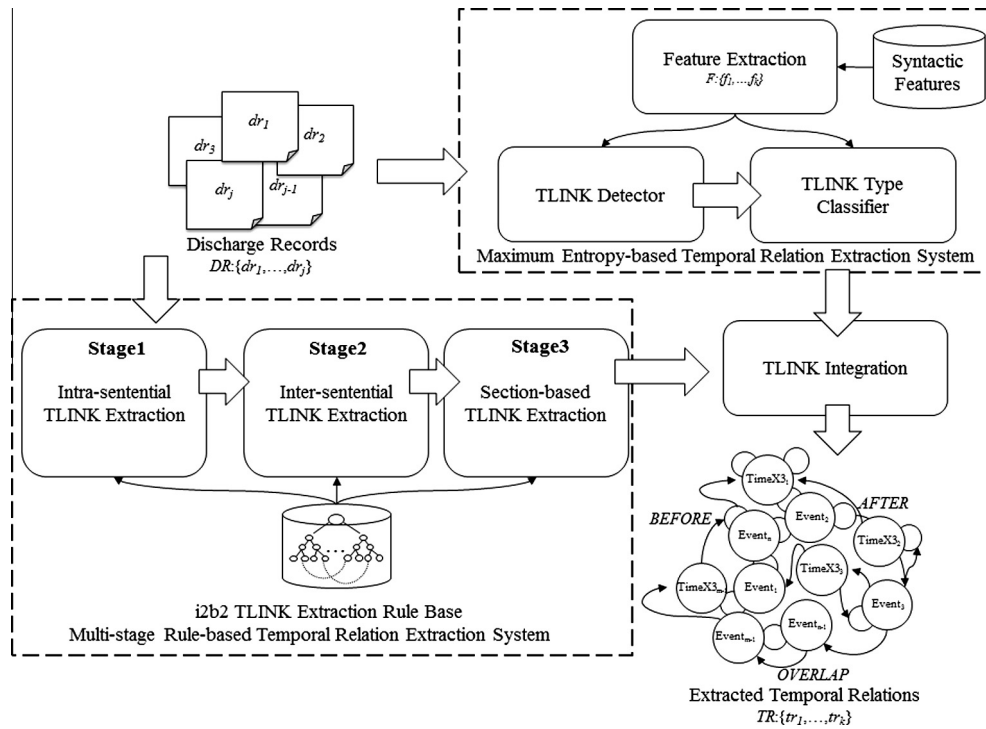


Fig. 1. The system architecture.

The antibiotic chosen was Levaquin 500 mg 1 tab per day.	(DocID: 2.xml)
$E = \{et_1, et_2, t_1\} = \{\text{"The antibiotic"}, \text{"Levaquin"}, \text{"per day"}\}$	
Extracted TLINKs:	
$TLINK_{OVERLAP}(et_1, et_2)$	
$TLINK_{OVERLAP}(t_1, et_2)$	

Fig. 2. An example of intra-sentence TLINK extraction.

2.1.2.1. Co-reference resolution. Co-reference resolution involves determining whether two noun phrases are used to refer to the same thing. We used the co-reference resolution method developed for the co-reference resolution track in the 2011 i2b2/VA/Cincinnati Challenge to determine the co-referential relationship between entities. The following rule captures the idea that if two EVENTS are co-reference pairs, a TLINK with type attribute “OVERLAP” should be established between them because they should indicate the same incident.

R.2.1 $isCoreference(e_i, e_j) \Rightarrow TLINK_{OVERLAP}(e_i, e_j)$

2.1.2.2. Timeline-dependent linking. To discover the relationships between time expressions or events across sentences, we developed a timeline-dependent algorithm that is implemented in two steps. First, all time expressions in a discharge summary are collected and denoted as $T = \{t_1, t_2, \dots, t_o\}$. Second, the TLINK types among the collected expressions are determined based on their descriptions.

Unfortunately, a wide variety of phrases and expressions can be used to state temporal information. For example, December 25, 2010 can be expressed in different ways, such as 2010-12-25, 12/25/2010, and 25/12/2010. Different temporal phrases that refer to the same event must be normalized so that they can be processed easily. In this work, the normalization information provided with the TLINK track dataset was used. For example, the normalized format for a calendar date is [YYYY]-[MM]-[DD]; and in the sentence “The patient was brought to the operating room on 03/

30/1999” the normalized time expression for “03/30/1999” is 1999-03-30. After using common sense rules to compare the normalized temporal description, the TLINK type is assigned to any t_i and t_j in T . For instance, the TLINK tuple $[t_i, \text{“AFTER”}, t_j]$ is generated if t_i ’s normalized value is after t_j ’s value, and the “OVERLAP” type is assigned if their values are the same.

After constructing the timeline for all t in T , the algorithm looks for the set of overlapping tuples. A set consists of entities whose temporal type is OVERLAP with their corresponding time expressions in T , e.g., R.1.2 may represent an overlapping tuple $\{et_1, \text{“OVERLAP”}, t_1\}, \{et_3, \text{“OVERLAP”}, t_2\}$. Using the constructed timeline, the algorithm then applies the following rule to determine the missing TLINK types among entities in the overlapping tuple.

R.2.2 $[TLINK_{OVERLAP}(e_i, t_k) \vee TLINK_{BEFORE}(e_i, t_k)] \wedge [TLINK_{OVERLAP}(e_j, t_l) \vee TLINK_{AFTER}(e_j, t_l)] \wedge TLINK_{BEFORE}(t_k, t_l) \Rightarrow TLINK_{BEFORE}(e_i, e_j)$

The above rule states that (1) if an entity e_i occurs on t_k or before t_k , (2) the occurrence time of another entity e_j is t_l (or after t_l), and (3) t_l is after t_k , then the TLINK tuple $[e_i, \text{“BEFORE”}, e_j]$ should be generated. We also use the following rule, which is based on a similar principle.

R.2.3 $TLINK_{OVERLAP}(e_i, t_k) \wedge TLINK_{OVERLAP}(e_j, t_l) \wedge TLINK_{OVERLAP}(t_k, t_l) \Rightarrow TLINK_{OVERLAP}(e_i, e_j)$

When it is known that an entity e_i overlaps with the time t_k , and the occurrence time of another entity e_j is t_l , with the additional knowledge that t_k and t_l overlap, R.2.3 establishes a TLINK tuple $\langle e_i, \text{"OVERLAP"}, e_j \rangle$.

R.2.4 $TLINK_{BEFORE}(e_i, t_k) \wedge TLINK_{AFTER}(e_j, t_l) \wedge TLINK_{OVERLAP}(t_k, t_l)$
 $\Rightarrow TLINK_{AFTER}(e_j, e_i)$

Based on the timeline concept and the logic shown on the left-hand side of R.2.4 it is reasonable to conclude that entity e_i occurred before entity e_j because of the overlap between t_k and t_l . Hence, a TLINK tuple $\langle e_j, \text{"AFTER"}, e_i \rangle$ is derived.

2.1.3. Stage 3: Clinical note-section-based TLINK extraction

In this stage, we use several rules to determine the type of TLINKs between the time expressions of “admission date” or “discharge date”, and the entities found in certain clinical note sections of a discharge summary. We divide a discharge summary into five parts: clinical history, admission, course of treatment, discharge, and other clinical note sections. Fig. 3 shows the timeline of the five parts.

We assign the section time of the admission note-section as the date of admission. In addition, the clinical history note-section, which usually contains information about the patient's physical status as well as his/her psychological, social, and sexual statuses, is set as the time before the admission date. The section times of the course of treatment, discharge and other note-sections are defined in a similar manner. The clinical note section-based timeline information can be used to determine the TLINK types of the enti-

ties in the different note-sections. We illustrate the concept in Fig. 4.

In this summary, it is obvious that the admission date t_1 is “05/09/2000”, and the discharge date t_2 is “05/30/2000” because they are listed in the “ADMISSION DATE” and “DISCHARGE DATE” note-sections, respectively. Therefore, the occurrence times of entities found in the “HISTORY OF PRESENT ILLNESS” note-section should be BEFORE t_1 and t_2 , because this note-section details the illness history of the patient. Similarly, the occurrence times of entities in the “COURSE OF TREATMENT” note-section, should be “AFTER t_1 , but BEFORE t_2 ”.

One of the key steps in this stage is to determine the clinical note sections of the discharge summary. Clinical records are usually in a consistent format. For example, a discharge summary is divided into different clinical note sections with each section title printed in upper case letters with a colon at the end [13]. We exploit this property and employ a string matching method to recognize the corresponding sections. Table 1 shows the keywords used and the corresponding clinical note sections.

Based on the clinical note sections in a discharge summary and the timeline information, we delineate the following rules. For example, R.3.1 sets the occurrence times of all EVENT entities in the course of treatment note-section as “AFTER” and “BEFORE” the admission date and the discharge date respectively.

R.3.1 $isHospitalCourseSection(s_i) \wedge isInSection(et_i, s_i)$
 $\Rightarrow TLINK_{AFTER}(et_i, t_{admission}) \wedge TLINK_{BEFORE}(et_i, t_{discharge})$

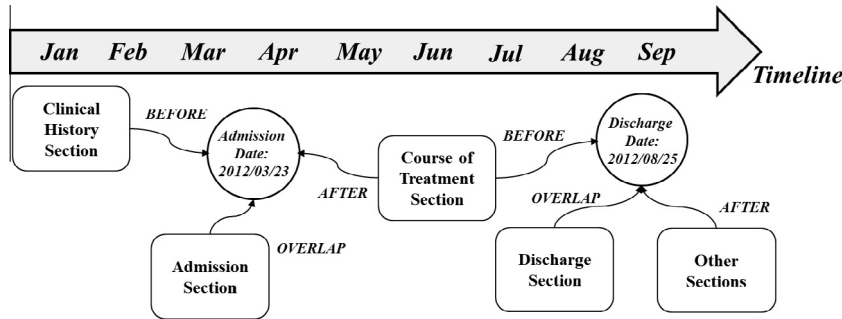


Fig. 3. The timeline of the five parts in a discharge summary.

ADMISSION DATE : 05/29/2000 DISCHARGE DATE : 05/30/2000 HISTORY OF <u>PRESENT</u> ILLNESS : The patient is a 51 year-old female, with a non-toxic multinodular goite. She has developed symptoms from a right lower dominant nodule with pressure discomfort in the neck. ... <u>HOSPITAL</u> COURSE : On 5-29-00, she underwent thyroidectomy with removal of a 3.3 centimeter right lobe microfollicular adenoma and left lobe 6 mm. papillary carcinoma. ... Entity set: $E = \{t_1, t_2, et_1, et_2, et_3, et_4, t_3, et_5, et_6, et_7, et_8\} = \{\text{"05/29/2000"}, \text{"05/30/2000"}, \text{"a non-toxic multinodular goite"}, \text{"symptoms"}, \text{"a right lower dominant nodule"}, \text{"pressure discomfort in the neck"}, \text{"5-29-00"}, \text{"thyroidectomy"}, \text{"removal"}, \text{"microfollicular adenoma"}, \text{"papillary carcinoma"}\}$ Extracted TLINKs: $TLINK_{BEFORE}(et_1, t_1)$ $TLINK_{BEFORE}(et_2, t_1)$ $TLINK_{BEFORE}(et_3, t_1)$ $TLINK_{BEFORE}(et_4, t_1)$ $TLINK_{BEFORE}(et_5, t_2)$ $TLINK_{BEFORE}(et_6, t_2)$ $TLINK_{BEFORE}(et_7, t_2)$ $TLINK_{BEFORE}(et_8, t_2)$
--

Fig. 4. An example of clinical note section-based TLINK extraction (DocID: 801.xml).

Table 1

Keywords used to determine the clinical note sections of a discharge summary.

Keyword	Clinical note section type
Admission	Admission
Discharge	Discharge
"procedure", "daily", "hospital"	Hospital Course
"present", "identifying", "history"	Clinical History
"additional"	Other

R.3.2 $isHospitalCourseSection(s_i) \wedge isInSection(et_i, s_i)$

$$\Rightarrow TLINK_{BEFORE}(et_i, t_{admission}) \wedge TLINK_{BEFORE}(et_i, t_{discharge})$$

R.3.3 $isAdmissionSection(s_i) \wedge isInSection(et_i, s_i)$

$$\Rightarrow TLINK_{OVERLAP}(et_i, t_{admission}) \wedge TLINK_{BEFORE}(et_i, t_{discharge})$$

isDischargeSection(s_i) $\wedge isInSection(et_i, s_i)$

$$\Rightarrow TLINK_{AFTER}(et_i, t_{admission}) \wedge TLINK_{OVERLAP}(et_i, t_{discharge})$$

R.3.4 $isOtherSection(s_i) \wedge isInSection(et_i, s_i)$

$$\Rightarrow TLINK_{AFTER}(et_i, t_{discharge})$$

2.2. Maximum entropy-based temporal relation extraction system

Fig. 5 shows the flow chart of the proposed supervised learning temporal relation extraction system, which is comprised of two key subsystems: the TLINK detector and the TLINK type classifier. In contrast to the rule-based system described in Section 2.1, which treats all pairs in the candidate TLINK pair set $P = \{p_1, \dots, p_n\}$ as possible candidates, the proposed system first filters the set with the TLINK detector. For each pair p' in the filtered pair set P' , the TLINK type classifier predicts its type to complete the TLINK tuple. For example, in the sentence "A peripheral intravenous line was started on Labor Day in the a.m. No respiratory distress was noted..." from 101.xml in the training dataset, three candidate TLINK pairs $\{\langle et_{24}, t_7 \rangle, \langle et_{24}, et_{221} \rangle, \langle et_{221}, t_7 \rangle\}$ are generated, where et_{24} , t_7 , and et_{221} represent "A peripheral intravenous line", "Labor Day", and "distress" respectively. After TLINK detection, only two TLINK pairs $\{\langle et_{24}, t_7 \rangle, \langle et_{221}, t_7 \rangle\}$ remain. Finally, the TLINK type

classifier infers the TLINK types, and generates the TLINK tuples $\{\langle et_{24}, "AFTER", t_7 \rangle, \langle et_{221}, "OVERLAP", t_7 \rangle\}$.

In this work, we formulate the TLINK detection and TLINK type classification tasks as binary and multi-class classification problems, respectively. Furthermore, we utilize the maximum entropy classification method [2] to construct a logistic regression-based statistical model for each task. Let TP denote a candidate TLINK pair. The maximum entropy method classifies a candidate TLINK pair in terms of the following conditional probability:

$$P(TP|cp_i) = \frac{1}{Z(cp_i)} \exp \left(\sum_j w_j * f_j(TP, cp_i) \right) \quad (1)$$

$$Z(cp_i) = \exp \left(\sum_j w_j * f_j(TP, cp_i) + \sum_k w_k * f_k(\neg TP, cp_i) \right), \quad (2)$$

where f_j is a feature function and w_j is its weight. A feature function indicates a specific condition between TP and cp_i ; and $Z(cp_i)$ is a smoothing factor that is used to normalize $P(TP|cp_i)$ within the range $[0, 1]$. Given a training dataset, the weights of the feature functions can be derived by the conditional maximum likelihood estimation method. The learned weights are then used by Eq. (1) for the TLINK detection and TLINK type classification tasks. Usually, the syntactic information between two entities is important for constructing a TLINK in a sentence. We utilize the subsets of lexical features proposed in [3,5,7] for TLINK detection and TLINK type prediction. Table 2 shows some of the lexical features. In the table, we use the candidate TLINK pairs, $p_1 = \langle et_{24}, t_7 \rangle = \langle \text{"A peripheral intravenous line", "Labor Day"} \rangle$, and $p_2 = \langle et_{24}, et_{221} \rangle$ shown in Fig. 5 as an example to explain the developed features.

2.3. TLINK integration algorithm

Fig. 6 shows the proposed TLINK integration algorithm, which is designed to favor the TLINK tuples generated by the rule-based system over those of the machine learning-based system because cross validation indicates that the former system outperformed the latter on the training set. We discuss the performance of both systems in the next section. In the first loop, the algorithm checks the tuples generated by the machine learning-based system to

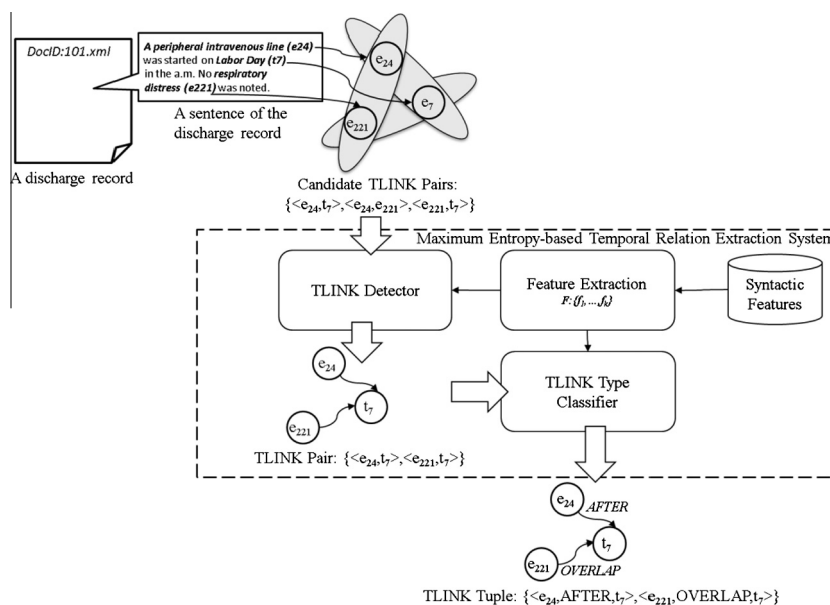
**Fig. 5.** System architecture of maximum entropy-based temporal relation extraction system.

Table 2

The lexical features used in the machine learning-based system.

Feature function	Description	Example
First type (<i>ft</i>)	The entity type of e_1	$ft(p_1) = \text{EVENT}$
Last type (<i>lt</i>)	The entity type of e_2	$lt(p_1) = \text{TIMEX3}$
Middle entities (<i>me</i>)	The number of entities between the given candidate TLINK pair	$me(p_2) = 1$
Distance (<i>dis</i>)	The number of tokens between the given candidate TLINK pair	$dis(p_2) = 9$
Density (<i>den</i>)	The probability of a TLINK pair establishing a TLINK decreases as the distance and the number of entities between them increases. The density feature is the same as to the inverse of the <i>dis</i> product to <i>me</i> . The logarithms and Laplace smoothing are used to avoid overflow and zero probability respectively $den(e_i, e_j) = -\log\left(\frac{1}{(dis(e_i, e_j)+1) \times (me(e_i, e_j)+1)}\right)$	$den = 1.301$

```

INPUT:
 $R_{MRTS} = \{ \langle e_l, t_l, e_m \rangle, \dots \}$ : the TLINK tuples from the Multi-stage Rule-based Temporal relation extraction System (MRTS)
 $R_{METS} = \{ \langle e_l, t_l, e_m \rangle, \dots \}$ : the TLINK tuples from the Maximum Entropy-based Temporal relation extraction System (METS)
BEGIN
  FOR EACH  $tuple_o = \langle e_l, t_l, e_m \rangle$  in  $R_{METS}$ 
    IF  $tuple_o = \langle e_l, t_l, e_m \rangle$  in  $R_{MRTS}$  # the pair  $\langle e_l, e_m \rangle$  is also in  $R_{MRTS}$  with the TLINK type  $t_l$ 
      IF  $t_l \neq t_o$  # if the types determined by MRTS and METS are different
        IF  $t_l$  is updated in Stage 2 or 3 # if the TLINK type  $t_l$  is determined in Stage 2 or Stage 3 of MRTS, accept the type,
          ADD  $tuple_o$  into  $R_{INTEGRATED}$  # accept  $tuple_o$ 
        ELSE # otherwise,
          ADD  $tuple_o$  into  $R_{INTEGRATED}$ 
        ENDIF
      ELSE
        ADD  $tuple_o$  into  $R_{INTEGRATED}$ 
      ENDIF
    ELSE
      ADD  $tuple_o$  into  $R_{INTEGRATED}$ 
    ENDIF
  ENDFOR
  FOR EACH  $tuple_p = \langle e_l, t_l, e_m \rangle$  in  $R_{MRTS}$ 
    IF the TLINK detector score of  $\langle e_l, e_m \rangle$  is larger than  $\tau$  # accept the  $tuple_p$  only if the score for  $\langle e_l, e_m \rangle$  generated by the TLINK
      ADD  $tuple_p$  into  $R_{INTEGRATED}$  # detector is large than a threshold  $\tau$ 
    ENDIF
  ENDFOR
END
OUTPUT:
 $R_{INTEGRATED}$ : the final TLINK tuples

```

Fig. 6. The TLINK integration algorithm. Statements behind the symbol “#” are comments.**Table 3**

Comparison of the performance of the machine learning-based system and each stage of the rule-based system on the training set.

Components	Precision (%)	Recall (%)	F-measure (%)
MRTS _{stage1}	42.04	30.47	35.33
MRTS _{stage2}	55.43	9.46	16.16
MRTS _{stage3}	78.87	9.87	17.54
METS	51.73	24.27	33.03

determine whether the entity pair $\langle e_l, e_m \rangle$ in the tuple also exists in the tuple set R_{MRTS} generated by the rule-based system. If the pair is found in R_{MRTS} , but with a different TLINK type, the algorithm will accept the TLINK type determined by the machine learning-based method. The algorithm will only accept the rule-based system's TLINK type when the system updates the TLINK type value after Stage 1 (see Section 2.1.1). We designed the algorithm in this way because the precision of Stages 2 and 3 in the rule-based system (see Sections 2.1.2 and 2.1.3) are better than those of the machine learning-based system (see Table 3 in the Results and Discussion section). In addition, if the entity pair is not involved in any tuples generated by the rule-based system, the tuple containing the pair will be accepted and added to the final tuple set.

In the second loop, the algorithm checks for the remaining tuple set of the rule-based system, which was not added to the final tuple set. From the perspective of the machine learning-based system, the tuples in the remaining tuple set should be filtered out by the TLINK detector. However, because we prefer to accept the results of the rule-based system, the algorithm only rejects the

entity pair $\langle e_l, e_m \rangle$ whose TLINK detector score is less than the threshold τ . In this work, we set the value of τ at 0.3.

3. Results

We used the training set released by the i2b2 TLINK track as our dataset to develop the rules used by the rule-based system (MRTS). For the machine learning-based system (METS), we performed threefold cross validation on the same dataset to select efficient features for TLINK detection and TLINK type classification in the sequential forward feature selection method [26]. We also performed threefold cross validation on the proposed stage-based system and the ME-based system to compare their performance. Then, we averaged the results of the three runs to obtain the global performance. Table 3 shows the performance of each stage of MRTS and METS using the official evaluation script provided by the i2b2 TLINK track. The evaluation metrics are the Precision, Recall, and F-measure.

The first three rows show the performance of the stages of MRTS. It is noteworthy that Stage 3 achieves the highest precision, which demonstrates the capability of the proposed clinical note section-based TLINK extraction method. The precision of METS is lower than that in MRTS-Stage 2 and Stage 3, but it is better than the score of MRTS-Stage 1. Accordingly, our TLINK integration algorithm only accepts the TLINK type determined by METS if its type remains the same after Stage 1 of MRTS (indicated by the third IF in Fig. 6). The results in the fourth row show that, the performance of METS was not significantly better than that of MRTS. This is

Table 4
Performance comparison of different configurations.

Configuration	Precision (%)	Recall (%)	F-measure (%)
(1) MRTS _{stage1}	42.04	30.47	35.33
(2) MRTS _{stage1+2}	43.77	46.66	45.17
(3) MRTS _{stage1+2+3}	54.87	49.44	52.01
(4) MRTS _{stage1+2+3} + METS	55.96	55.07	55.51
Baseline _{random}	27.98	18.02	21.92

Table 5
Overall results derived by the compared methods on the test dataset.

Components	Precision (%)	Recall (%)	F-measure (%)
Run1 (MRTS _{stage1+2})	57.58	50.94	54.05
Run2 (MRTS _{stage1+2+3})	55.07	55.55	55.31
Run3 (MRTS _{stage1+2+3} + METS)	56.70	55.86	56.28
Baseline _{random}	25.42	17.81	20.95

probably due to the low inter-annotator agreement (kappa value: 0.3 reported by the organizers [27]) and the inability of syntactic features alone to discriminate between temporal relations. To achieve better results, supplementary context information and semantic features should also be considered.

Table 4 compares the results derived by different configurations on the training set. The first configuration only uses the rules defined in MRTS-Stage 1. Based on the results of MRTS_{stage1}, the second configuration integrates the rules in MRTS-Stage 2 to modify the generated TLINK types. MRTS_{stage1+2+3} includes all the stages developed for MRTS. Finally, the MRTS_{stage1+2+3} + METS configuration uses the proposed TLINK integration algorithm to fuse all the TLINK tuples generated by MRTS and METS. In addition, as a baseline for the evaluation, we use a naive method (Baseline_{random}), which randomly selects TLINK candidates from all pairs and assigns the TLINK types. Comparison of MRTS_{stage1} with MRTS_{stage1+2+3} shows that the proposed stages improve the overall performance of temporal relation extraction, and they clearly outperform the baseline method. Incorporating Stage 2 into MRTS improves the F-score by 9.84%. The full MRTS system achieves an F-score of 52.01% on the cross validation dataset. The final hybrid system, which uses the TLINK integration algorithm to incorporate the results of METS with MRTS, improves the F-score by 20.18%.

Finally, Table 5 shows the average scores of the three runs we submitted for the i2b2 TLINK test dataset. In line with our observation in the training set, the hybrid system (Run 3) achieves the best performance, which demonstrates that the proposed TLINK integration algorithm is robust. The TEMPTING system also outperforms the baseline, which applies a random method on the test dataset.

4. Discussion

Compared to general relation extraction tasks, identifying temporal relations in discharge summaries requires temporal knowledge and inference. To assess the effectiveness of the proposed rules for different TLINK types, we preserve the TLINK type that we want to evaluate for each discharge summary, and select rules based on their target TLINK types for evaluating the individual TLINK type performance on the training and test sets. For instance, we remove TLINK types AFTER and OVERLAP from the dataset, and only use rules related to the BEFORE link (denoted as MRTS_{BEFORE}) when we evaluate the BEFORE TLINK. The results show that the TEMPTING system achieves the best performance on the OVERLAP TLINKs, and the worst performance on the TLINK type AFTER. The reason that the AFTER TLINK is much harder than others is that

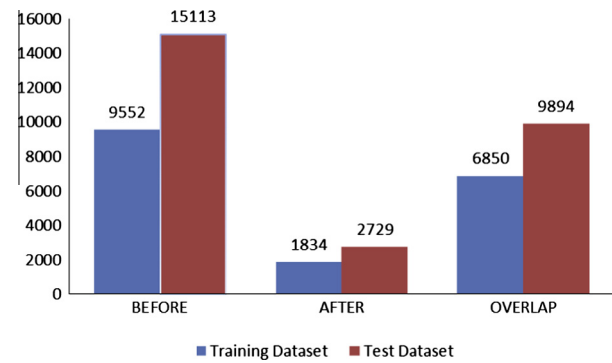


Fig. 7. The distribution of each TLINK type in the training and test datasets.

the distribution of AFTER TLINKs in the i2b2 dataset is much less than that of the others, as shown in Fig. 7. The phenomenon prevents us from devising useful rules and learning meaningful feature weights; hence the recognition task is more complicated.

Furthermore, after carefully examining the cases with incorrect AFTER TLINKs, we found that the developed rules based on the identified clinical note sections and their timeline information of MRTS_{AFTER} are incapable of discriminating between EVENT entities. For instance, R.3.1 sets the occurrence time of all EVENT entities in the course of treatment note-section as “AFTER” the admission date due to the difficulty in identifying the clinical note sections, leading to the unsatisfactory performance by MRTS_{AFTER}.

Our analysis of the test dataset shows that the TEMPTING system can achieve approximately 76.0% accuracy in identifying the TLINK between EVENT and TIMEX3 entities. Specifically, the accuracy are 90.5% for “BEFORE”, 71.4% for “OVERLAP”, and 66.7% for “AFTER”. However, the system’s performance in recognizing the TLINK between EVENT and SECTIME is poor with only 20% accuracy on average. In our system, the development of this type of TLINK is based solely on the clinical note section that the EVENT belongs to. Because the chronological relations between note-sections are not closely examined, some of the TLINKs are incorrect. In addition, through the intense discussion of the TLINK track of the i2b2 2012 challenge Google group, the cognition of existing TLINKs varied among the annotators and participants. Disagreements between the annotators caused some complications, especially when applying the machine learning approach. Similarly, determining the cases where two EVENTS/TIMEX3s should be linked with a TLINK is also a challenging task, and may result in lost or redundant TLINKs. Based on the above analysis, we should prioritize improving the performance of detecting TLINKs between EVENT and SECTIME in the future and improve the efficiency of TEMPTING in retrieving TLINKs.

In addition, our results show that MRTS_{BEFORE} achieves a high precision, but a low recall. Our analysis of the experimental data shows that BEFORE links are usually constructed by an EVENT entity and a TIMEX3 entity; hence, the rule that sets the occurrence time of all EVENT entities in the clinical history note-section to “BEFORE” the admission date and the discharge date (see R.3.1, R.3.2, and R.3.3) yields a very high detection precision because the clinical history note-section is much easier to identify by using the proposed keywords (see Table 1). However, because our clinical note section identification method based on string matching has difficulty determining the corresponding note-sections, its recall is low.

For the machine learning-based system comprised of two subsystems, i.e., the TLINK detector and the TLINK type classifier, we conducted an additional experiment to verify the usability of the TLINK detector. The results are shown in Table 6. The first configuration (denoted as METS_{one-stage}) excludes the TLINK detector,

Table 6Performance comparison on the training set for METS_{one-stage} and METS_{two-stage}.

Components	Precision (%)	Recall (%)	F-measure (%)
METS _{one-stage}	38.98	40.22	39.59
METS	51.73	24.27	33.03
METS _{one-stage} + MRTS	49.84	59.25	53.57
METS + MRTS	55.96	55.07	55.51

but the second configuration adds it in MRTS_{stage1} (denoted as METS). In MRTS_{stage1+2+3} and MRTS_{stage1+2+3} + METS, based on the results of MRTS_{stage1} and MRTS_{stage1+2}, we use the TLINK integration algorithm to fuse all TLINK tuples generated by the rule-based system.

By comparing different configurations, we find that adding the TLINK detector improves the precision at the expense of the recall. Although adding the TLINK detector yields the lowest *F*-score, it improves the precision and *F*-score when we use the fusion algorithm to combine the outputs of the rule-based and machine learning-based systems through the fusing algorithm, while it still maintained a satisfied recall. Based on the results, we conclude that the TLINK detection step must be executed before predicting TLINK types.

5. Related work

Regardless of the specific details of target temporal relations, most researchers have approached the problem in a rule-based manner or a pattern-based manner. Many existing temporal relation extraction systems link events directly to associated time-stamps [9,21,16]. This is a simple and practical approach, but it is used at the expense of the recall because many events do not have an associated timestamp. The principle behind rule-based methods is to design a number of rules for classifying the types of temporal relations as shown by our rule-based system. In most works, these manually defined rules are based on Allen's interval algebra [1], which is a calculus developed for temporal reasoning. One usage of these rules is to enlarge the training set [14], and another is to ascertain the predicted temporal relations [21,25]. Pattern-based methods extract some generic lexical syntactic patterns for the co-occurrence of events. Extraction of such patterns can be performed manually or automatically. The major drawback of extracting patterns manually is that it tends to result in a high recall but a low precision, as shown by our results in Tables 3 and 4. Several heuristics that are similar to our hybrid method have been proposed to resolve the low precision problem [5,25].

Our maximum entropy-based system is an example of a machine learning-based approach that attempts to learn a classifier from an annotated corpus, and tries to improve the classification accuracy by feature engineering. Another example is Mani et al.'s ME classifier [14], which assigns each pair of events to one of the six relations in an augmented TimeBank corpus. The classifier relies on perfect features that were hand-tagged in the corpus, namely, the tense, aspect, modality, polarity and event class. Pair-wise agreement on grammatical tenses and aspect were also included. Lapata and Lascarides [11], whose learner is based on syntax and clausal ordering features, trained an event classifier for inter-sentence events and built a corpus by saving sentences that contained two events, one of which is triggered by a key time word (e.g. after or before).

There was a large variation in the machine learning methods explored by the i2b2 TLINK track participants, ranging from maximum entropy, Bayesian, and support vector machine methods to conditional random field approaches. For example, Roberts et al. [20] proposed two supervised methods. The first uses conditional

random fields to find relevant word sequences; and the second uses support vector machines to perform binary and multi-class classification. Their official results reported an overall temporal link detection *F*-score of 55.94%. Cherry et al. [6] proposed several features, such as syntactic, semantic, and structural features, and developed a temporal reasoning module to further infer the temporal relations between entities. Some participants also incorporated heuristics and rule-based components into their systems. For example, Tang et al. [24] used heuristics to select candidate entity pairs for assigning TLINKs. Nikfarjam et al. [15] also combined machine learning and a graph driven approach to extract TLINKs. The graph-based approach creates a temporal graph of a sentence based on parse tree dependencies of the simplified sentences, along with some additional frequent patterns. Some teams divided the tasks into more specific sub-tasks to improve the performance [10,28]. Other hybrid systems in the 2012 i2b2 TLINK track used simple rules to merge the TLINK results of multiple classifiers, e.g., by removing contradictory sentence-across TLINKs from the same sentences [24]. By contrast, we developed a fusion algorithm that integrates the results of two systems based on their performance on the dataset. The best configuration of our hybrid system achieved an *F*-score of 0.56, which ranked in the seventh place out of the top 10 teams.

6. Concluding remarks

Temporal information is crucial in electronic medical records and related clinical manuscripts. Hence, discovering the temporal relationships between entities in patient's medical data may help medical personnel make clinical decisions, as well as facilitate data modeling and biomedical research. Most of the approaches used in the 2012 i2b2 TLINK track were hybrid methods that combined supervised, unsupervised, and rule-based methods to extract temporal relations. In this paper, we describe several rules and present an algorithm that integrates the results of rule-based and supervised learning systems. Our experiment results demonstrate the efficacy of the proposed approach. Nevertheless, we believe our system could be improved by incorporating of additional linguistic and domain knowledge. In the future, we will employ more sophisticated features to strengthen the system. We will also investigate advanced machine learning algorithms, such as Markov logic network [29], which can extract temporal relation tuples accurately. Our objective is to further enhance the performance of temporal representation and reasoning in medical natural language processing.

Acknowledgments

This research was supported by Informatics for Integrating Biology and the Bedside (i2b2), award number 2U54LM008748 from the NIH/National Library of Medicine (NLM); by the National Heart, Lung and Blood Institute (NHLBI); and by award number 1R13LM01141101 from the NIH NLM. The content is solely the responsibility of the authors and does not necessarily reflect the official views of the NLM, NHLBI, or the National Institutes of Health. The study was conducted under the "III Innovative and Prospective Technologies Project" of the Institute for Information Industry, which is subsidized by the Ministry of Economic Affairs of the Republic of China. Moreover, this research was also supported by the National Science Council of Taiwan under Grant NSC 101-2319-B-010-002, NSC 102-2319-B-010-002, and NSC-102-2218-E-038-001, the research Grant TMU101-AE1-B55 of Taipei Medical University.

References

- [1] Allen J. Towards a general theory of action and time. *J Artif Intell* 1984;23(2):123–54.
- [2] Berger AL, Pietra SAD, Pietra VJD. A maximum entropy approach to natural language processing. *Comput Linguist* 1996;22(1):39–71.
- [3] Bethard S, Martin JH. CU-TMP: temporal relation classification using syntactic and semantic features. In: Proceedings of the 4th international workshop on semantic evaluations. Prague, Czech Republic: Association for Computational Linguistics; 2007. p. 129–32.
- [4] Boguraev B, Ando RK. Timeml-compliant text analysis for temporal reasoning. In: Proceedings of the 19th international joint conference on artificial intelligence; 2005. p. 997–1003.
- [5] Chang YC, Chuang PH, Chen CC, Hsu WL. FISER: an effective method for detecting interactions between topic persons. In: Proceedings of the 8th Asia information retrieval societies conference; 2012. p. 267–77.
- [6] Cherry C, Zhu X, Martin J, et al. A la Recherche du Temps Perdu—extracting temporal relations from medical text in the 2012 i2b2 NLP challenge. *J Am Med Inform Assoc*. doi:<http://dx.doi.org/10.1136/amiajnl-2013-001624> [published online first: 23.03.13].
- [7] Chklovski T, Pantel P. Global path-based refinement of noisy graphs applied to verb semantics. In: Proceeding of international joint conference on natural language processing; 2005. p. 792–803.
- [8] Dai HJ, Chen CY, Wu CY, Lai PT, Tsai RT, Hsu WL. Coreference resolution of medical concepts in discharge summaries by exploiting contextual information. *J Am Med Inform Assoc*; 2012 May 3.
- [9] Filatova E, Hovy E. Assigning time-stamps to event-clauses. In: Proceedings of the workshop on temporal and spatial information processing, vol. 13; 2001. p. 88–95.
- [10] Grouin C, Grabar N, Hamon T, et al. Eventual situations for timeline extraction from clinical reports. *J Am Med Inform Assoc* 2013;20(5):820–7. <http://dx.doi.org/10.1136/amiajnl-2013-001627>.
- [11] Lapata M, Lascarides A. Learning sentence-internal temporal relations. *J AI Res* 2006;27:85–117.
- [12] Ling X, Weld DS. Temporal information extraction. In: Proceedings of association for the advancement of artificial intelligence; 2010. p. 1385–90.
- [13] Long W. Extracting diagnoses from discharge summaries. In: Proceedings of the AMIA Annu Symp. American Medical Informatics Association; 2005. p. 470–4.
- [14] Mani I, Verhagen M, Wellner B, Lee CM, Pustejovsky J. Machine learning of temporal relations. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics; 2006. p. 753–60.
- [15] Nikfarjam A, Emadzadeh E, Sutton N, Gonzalez G. Temporal relationship extraction from clinical-notes using SVM and graph reasoning. In: Proceedings of 2012 i2b2 shared-tasks and workshop on challenges in natural language processing for clinical data; 2012.
- [16] Pasca, M. Answering definition questions via temporally-anchored text snippets. In: Proceedings of international joint conference on natural language processing; 2008. p. 411–7.
- [17] Pustejovsky J, Hanks P, Sauri R, et al. The TimeBank corpus. In: Proceedings of corpus linguistics. UK: Lancaster University; 2003. p. 647–56.
- [18] Pustejovsky J, Castano J, Ingria R, et al. TimeML: Robust specification of event and temporal expressions in text. In: Proceedings of the fifth international workshop on computational semantics (IWCS-5); 2003.
- [19] Pustejovsky J, Knippen R, Littman J, Sauri R. Temporal and event information in natural language text. *Lang Resour Eval* 2005;39(2):123–64.
- [20] Roberts K, Rink B, Harabagiu SM. UTD: hybrid methods for temporal relation identification in clinical text. In: Proceedings of 2012 i2b2 shared-tasks and workshop on challenges in natural language processing for clinical data; 2012.
- [21] Schilder F, Habel C. From temporal expressions to temporal information: semantic tagging of news messages. In: Proceedings of the workshop on temporal and spatial information processing, vol. 13; 2001. p. 65–72.
- [22] Sun W, Rumshisky A, Uzuner O, Szolovits P, Pustejovsky J. 2012 i2b2 Temporal relations challenge annotation guidelines; 2012.
- [23] Verhagen M, Sauri R, Caselli T, Pustejovsky J. SemEval-2010 task 13: TempEval-2. In: Proceedings of the 5th international workshop on semantic evaluation; 2010. p. 57–62.
- [24] Tang B, Wu Y, Jiang M, et al. A hybrid system for temporal information extraction from clinical text. *J Am Med Inform Assoc* 2013;20(5):828–35. <http://dx.doi.org/10.1136/amiajnl-2013-001635>.
- [25] Torisawa K. Acquiring inference rules with temporal constraints by using japanese coordinated sentences and noun-verb co-occurrences. In: Proceedings of the main conference on human language technology conference of the North American chapter of the association of computational linguistics; 2006. p. 57–64.
- [26] Tsai RTH, Sung CL, Dai HJ, Hung HC, Sung TY, Hsu WL. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics* 2006;7(Suppl. 5):S11.
- [27] Weiyl S, Anna R, Ozlem U. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J Am Med Inform Assoc*. doi:<http://dx.doi.org/10.1136/amiajnl-2013-001628> [published online first: 05.04.13].
- [28] Xu Y, Wang Y, Liu T, et al. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *J Am Med Inform Assoc* doi:<http://dx.doi.org/10.1136/amiajnl-2012-001607> [published online first: 06.03.13].
- [29] Yoshikawa K, Riedel S, Asahara M, Matsumoto Y. Jointly identifying temporal relations with Markov logic. In: Proceedings of the joint conference of the 47th annual meeting of the ACL and the 4th international joint conference on natural language processing of the AFNLP, vol. 1; 2009. p. 405–13.