Temporal Relation Extraction in Clinical Texts

Yohan Bonescki Gumiel, Lucas Emanuel Silva E Oliveira, Vincent Claveau, Natalia Grabar, Emerson Cabrera Paraiso, Claudia Moro, Deborah Ribeiro Carvalho

Temporal Relation Extraction in Clinical Texts: A Systematic Review

YOHAN BONESCKI GUMIEL and LUCAS EMANUEL SILVA E OLIVEIRA, Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil
VINCENT CLAVEAU, IRISA - CRNS, Université de Rennes 1, Rennes, Rennes, France
NATALIA GRABAR, CRNS, Univ. Lille, Lille, 59000 Lille, France
EMERSON CABRERA PARAISO, Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil
CLAUDIA MORO and DEBORAH RIBEIRO CARVALHO, Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Curitiba, Paraná, Brazil

Unstructured data in electronic health records, represented by clinical texts, are a vast source of healthcare information because they describe a patient's journey, including clinical findings, procedures, and information about the continuity of care. The publication of several studies on temporal relation extraction from clinical texts during the last decade and the realization of multiple shared tasks highlight the importance of this research theme. Therefore, we propose a review of temporal relation extraction in clinical texts. We analyzed 105 articles and verified that relations between events and document creation time, a coarse temporality type, were addressed with traditional machine learning–based models with few recent initiatives to push the state-of-the-art with deep learning–based models. For temporal relations between entities (event and temporal expressions) in the document, factors such as dataset imbalance because of candidate pair generation and task complexity directly affect the system's performance. The state-of-the-art resides on attention-based models, with contextualized word representations being fine-tuned for temporal relation extraction. However, further experiments and advances in the research topic are required until real-time clinical domain applications are released. Furthermore, most of the publications mainly reside on the same dataset, hindering the need for new annotation projects that provide datasets for different medical specialties, clinical text types, and even languages.

CCS Concepts: • **Computing methodologies** → **Information extraction**; *Artificial intelligence*; *Natural language processing*; • **Applied computing** → **Health informatics**; *Life and medical sciences*;

Additional Key Words and Phrases: Temporal relation extraction, natural language processing, clinical data

This work was supported by the Brazilian Government Agency Coordination for the Improvement of Higher Education Personnel (CAPES).

Authors' addresses: Y. B. Gumiel (corresponding author), L. E. S. e Oliveira, C. Moro, and D. R. Carvalho, Graduate Program in Health Technology, Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição 1155, 80215-901, Curitiba, Brazil; emails: {yohan.gumiel, lucas.oliveira, c.moro, ribeiro.carvalho}@pucpr.br; V. Claveau, IRISA – CRNS, Université de Rennes 1, 263 Avenue Général Leclerc, 35000, Rennes, France; email: vincent.claveau@irisa.fr; N. Grabar, CRNS, Université of Lille, 42 Rue Paul Duez, 59000, Lille, France; email: natalia.grabar@univ-lille3.fr; E. C. Paraiso, Graduate Program in Informatics, Pontifícia Universidade Católica do Paraná, Rua Imaculada Conceição 1155, 80215-901, Curitiba, Brazil; email: paraiso@ppgia.pucpr.br.

0360-0300/2021/09-ART144 $15.00
https://doi.org/10.1145/3462475

ACM Reference format:
Yohan Bonescki Gumiel, Lucas Emanuel Silva e Oliveira, Vincent Claveau, Natalia Grabar, Emerson Cabrera Paraiso, Claudia Moro, and Deborah Ribeiro Carvalho. 2021. Temporal Relation Extraction in Clinical Texts: A Systematic Review. *ACM Comput. Surv.* 54, 7, Article 144 (September 2021), 36 pages.
https://doi.org/10.1145/3462475

1 INTRODUCTION

Whenever patients seek medical care data, their information is recorded in electronic health records (EHRs) and stored in either a structured or unstructured format. Structured data include medication information, laboratory data, and radiologic images [1]. In contrast, unstructured data are represented by clinical texts, such as discharge summaries and pathology reports. Much of the information in EHRs is unstructured, limiting its secondary use in improving medical research and developing tools to assist patient care. For instance, the authors of [2] observed during a study related to six hospitals that, on an average, 75% of all data elements were not available in a structured format or computable database fields. The preference for free text relies on the fact that it facilitates communication among the care team and enables health professionals to provide more detailed information because they are not restricted to structured fields [3].

Natural language processing (NLP) tools enable the secondary use of clinical texts by developing frameworks that automatically analyze and transform textual information into structured representations [4]. The application of NLP to texts written by healthcare professionals in a healthcare environment is called clinical NLP [5]. It extracts rich and contextual information not available elsewhere and involves rich temporal and background information about current status/conditions, even information about a patient's past (e.g., a treatment that occurred a long time ago) [6, 7]. It can also provide information about the future (e.g., foreseen interventions and treatments).

Temporal relation extraction aims to provide order among mentions over texts, representing medical events or temporal expressions. In the clinical domain, events are clinically relevant situations (e.g., treatments, problems, tests), and temporal expressions allude to temporal mentions (e.g., duration or date mentions). A temporal expression can be either a time mention in the free text or document creation time (DCT).

Research on temporal relation extraction is opportune because of the longitudinal data present in the EHRs, with several clinical texts on the same patient written at different times. Clinical texts that reflect a specific time frame, such as discharge summaries that cover the temporal window from patient admission until discharge, are also relevant.

Noncommunicable diseases (NCDs), such as cardiovascular diseases and cancers, have a longitudinal nature and provide extensive and continuous dataflows relevant to temporal relation extraction [8]. Research that can improve or supplement clinical decision-making related to NCDs is valuable, as NCDs are the leading cause of death globally, accounting for over 70% of deaths [9]. Another research topic of interest is related to adverse events, since symptoms and signs tend to appear over time after the start of a specific treatment (e.g., medication).

However, there are some challenges related to temporal relation annotation and extraction. Temporal relations can be implicit and vague, which is troublesome for both extraction and annotation [10]. In general, text annotation is a complicated process, but the annotation of temporal relations is much more complicated. For instance, temporal relation extraction in the clinical domain has a lower inter-annotator agreement (IAA) than other clinical annotation tasks, such as event and temporal expression annotation tasks [7]. Aspects such as lack of formalism and writing quality may make the extraction of temporal relations in the clinical domain more complicated than in the

general domain [11]. Furthermore, for clinical domain corpora, both annotation and extraction can require specific medical expertise, which can be expensive for the annotation process and difficult in the extraction step.

Further, clinical texts can exhibit specific characteristics that can directly impact the text preprocessing steps and extraction results. There is an extensive use of abbreviations and acronyms, particularly in individual institutions or medical specialties. Domain-specific vocabulary and assumptions are also present [4, 12]. In addition, texts may contain flexible formatting and atypical grammatical constructions [13, 14]. Moreover, the need for specific knowledge and tools may be a limiting factor, especially in the clinical domain, owing to the lack of resources and available data. Sophisticated NLP tools, which can be used for preprocessing and aggregating information, are typically provided by language-dependent frameworks, hindering the use of languages other than English. The amount of available data is also a limiting factor, and deep learning approaches rely on a large amount of data to address generality. Additionally, access to clinical domain data is difficult because of data privacy.

Many studies on temporal relation extraction are related to datasets that have become available to researchers because of shared tasks. Hopefully, several shared tasks have been organized to provide data that the research community can use to develop temporal extraction techniques and compare extraction performance. The interest in temporal relation extraction from clinical narratives began to grow with the Informatics for Integrating Biology and Bedside (i2b2) 2012 challenge [15], and then with Clinical TempEval in semEval2015 [16], semEval2016 [17], and semEval2017 [18] shared tasks. With the intent of discussing the approaches used (both shared task–related or not), highlighting the main aspects, and pointing out the best methods in studies, we performed a systematic review that followed the PRISMA statement [19].

Although there are two reviews on extracting temporal relationships in clinical texts, some topics still need to be covered. The authors of [20] highlighted some preliminary studies between 2006 and 2012, while the authors of [21] presented studies between 2006 and 2018. Owing to recent discoveries, the state-of-the-art changed over these two years, which was not covered by the authors of [21]. Currently, the state-of-the-art for several NLP tasks involves attention-based models and contextualized word representations. Hence, we aim to address this gap by considering the most recent approaches and discoveries. Further, a limitation of the review presented by the authors of [21] was that they only considered free text written in English, which limits the review power of providing insights about research in other countries. Thus, we aim at covering this gap in our review with no language restrictions. Additionally, using our publication selection criteria, we analyze a significantly more extensive set of articles than that contemplated by the authors of [21], covering the research topic evolution over the years and providing a deeper analysis of the approaches. Further, we provide visual representations to improve the understanding of the temporal relation types and their importance, highlighting their differences and their applicability in the clinical domain.

The objective of this study is to present a review of the state-of-the-art temporal relation extractions from clinical texts. It aims to answer the following question: "What is the effectiveness of machine learning and rule-based techniques in identifying temporal relations in clinical texts?" Our secondary objective is to provide insights into the domain evolution over time by leveraging temporal relation extraction objectives and developing frameworks. A reader of this review can expect an analysis of temporal relations and investigate the best performing techniques and frameworks for temporal relation extraction.

The remainder of this article is structured as follows. Section 2 provides an overview of temporal relation extraction, including explanations of temporal relation representations and an example highlighting its importance for the clinical domain. In Section 3, the methodological steps are

detailed, and global quantitative results and details of the datasets are provided. We divided the task of temporal relation extraction into two distinct types: (i) DocTimeRel, a temporal relation between an event and the DCT, a temporal expression referring to a date in the document header that indicates when the document was created/written; and (ii) TLINK, a temporal relation between mentions that occur over the text, where mentions can be events and temporal expressions (do not involve the DCT). We adopted this strategy based on the previous temporal relation shared tasks for both clinical and general domains and because each type has different extraction characteristics and task complexities. We elaborate on the DocTimeRel-related articles in Section 4 and the TLINK-related articles in Section 5. In Section 6, we present an overview of the datasets and relevant approaches for the general domain. In Section 7, we present our conclusions.

2 TEMPORAL RELATION EXTRACTION

Temporal relation extraction can be summarized in two steps: (i) identifying a relation between pairs of mentions (e.g., event and temporal expressions) and (ii) classifying this relation into a temporal relation type among a predefined set. Depending on the application, only the first step is sufficient, but a more detailed representation can be obtained only by using both steps.

In Section 2.1, we explain temporal relation representations and discuss the differences between temporal relation sets. In Section 2.2, we explain the event and temporal expression characteristics in both clinical and general domains while providing a concrete example of the importance of temporal relation extraction for the clinical domain.

2.1 Temporal Relation Representations

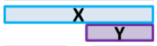
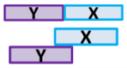
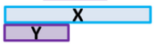
The interval-based algebra proposed by Allen in 1983 was used as a framework for temporal relation extraction. Several studies adopted Allen's representation [10], which quickly became a temporal modeling pattern [11]. Allen's representation assumes that, given two points in time or intervals of time, any relationship between them could be represented by seven relations: BEFORE, MEET, OVERLAP, DURING, START, FINISH, and EQUAL [10]. Considering the inverse relations (EQUAL does not have an inverse relation), there are 13 possible relations. Allen's relations are listed in Table 1 (Allen's Algebra column).

Several annotation standards have been developed based on Allen's representation. We highlight TimeML [22], a reference for temporal annotation for the general domain, and THYME-TimeML [6], an adaptation of TimeML for the clinical domain.

TimeML is a temporal markup language developed exclusively to annotate events, temporal expressions, and relations in the text [22]. Researchers in the NLP community have developed TimeML to move temporal information from a free-text format to a structured data format [23]. In TimeML, events are situations that occur, and temporal expressions are mentions of dates, times (specific time during a day), durations, and sets [24]. The TLINK tag represents a temporal relationship between events and temporal expressions. The main difference between Allen's representation and TimeML is that TimeML does not address OVERLAP relations. The relation EQUAL in Allen's algebra is represented over four relations in TimeML: IDENTITY, SIMULTANEOUS, HOLD, and HELD BY [25]. The IDENTITY relation is similar to the SIMULTANEOUS relation but is used only in event co-reference cases [26]. The TimeML relations are displayed in Table 1 (TimeML column).

THYME-TimeML was developed to annotate the temporal history of our medical events (THYME) corpus, which comprises clinical notes from patients with cancer and pathology reports. Thus, the event definition involves a clinical vocabulary, with mentions such as medical problems (e.g., signs and symptoms), treatments (e.g., medications), and tests (e.g., laboratory exams). Some of these events are particular to oncology. In THYME-TimeML, events are mentions relevant to

Table 1. Temporal Relation Types with Their Respective Graphical Representation and Identification in Allen's Representation, TimeML, THYME-TimeML, and i2b2 Annotation Schemas

Graphical Representation	Allen's representation	TimeML schema	I2b2 corpus schema	THYME-TimeML schema (TLINK)	THYME-TimeML schema (DocTimeRel)
	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y	X BEFORE Y
	X MEETS Y	X I_BEFORE Y	X BEF_OVERLAP Y	-	X BEF_OVERLAP Y
	X OVERLAPS Y	-	X OVERLAPS Y	X OVERLAPS Y	X OVERLAPS Y
	X DURING Y	X IS_INCLUDED Y	X DURING Y	-	-
	X STARTS Y	X BEGINS Y	-	X BEGINS_ON Y	-
	X FINISHES Y	X ENDS Y	-	-	-
	X FINISHED_BY Y	X ENDS_BY Y	X ENDS_BY Y	X ENDS_ON Y	-
	X AFTER Y	X AFTER Y	X AFTER Y	-	X AFTER Y
	X MEET_BY Y	X I_AFTER Y	-	-	-
	X OVERLAPPED_BY Y	-	-	-	-
	X CONTAINS Y	X INCLUDES Y	-	X CONTAINS Y	-
	X STARTED_BY Y	X BEGINS_BY Y	X BEGINS_BY Y	-	-
	X EQUALS Y	X IDENTIFIES Y X SIMULT. Y X HOLDS Y X HELD_BY Y	X SIMULT. Y	-	-

Note: Relation types annotated but not used for the shared tasks are marked in gray.

constructing a clinical timeline. The temporal expression definitions are similar to TimeML, with the addition of a new category for preoperative, intraoperative, and postoperative mentions [6]. The significant differences between TimeML and THYME-TimeML for temporal relation annotation are as follows: (i) THYME-TimeML created the DocTimeRel category, while the relations between events and the DCT are considered as common TLINKs in TimeML, and (ii) THYME-TimeML introduces the narrative container concept.

The DocTimeRel relations are considered an event attribute and have the following relation set: BEFORE, AFTER, OVERLAP, BEFORE/OVERLAP, and AFTER. The THYME-TimeML DocTimeRel relations are displayed in Table 1 (THYME-TimeML DocTimeRel column). BEFORE/OVERLAP indicates that the event occurred in the past and still occurs in the DCT. For instance, depending on the annotation schema, chronic diseases can be annotated as BEFORE/OVERLAP because they exist before the clinical document creation and continue to exist during its writing.

The narrative container concept is used to annotate the TLINKs. The THYME-TimeML TLINK relation set are BEFORE, OVERLAPS, BEGINS_ON, ENDS_ON, and CONTAINS. The THYME-TimeML TLINK relations are displayed in Table 1 (THYME-TimeML TLINKs column). The narrative container concept introduced in [27] involves the CONTAINS relation. The authors of [27] emphasize the importance of an annotation schema that resulted in maximally annotated temporal relation information while not relying on models that were too difficult to apply.

The choice of using narrative containers comes from the difficulty in capturing every possible relation and the rise in disagreement that occurs when annotators try to do so [6]. By using this choice of annotation, whenever possible, time expressions and events are connected to a narrative container (event or temporal expression anchor) that defines their temporal interval. Several events or temporal expressions can be connected to the same anchor, which contains them (represented in

the CONTAINS row in Table 1). Events and temporal expressions in the same narrative container can be related, as a single element, with other containers [28]. The most significant advantage is a reduction in the number of required annotations [28]. The narrative container strategy is suitable in the clinical domain because there are central mentions of the texts, such as temporal expressions of date and time types, or more comprehensive events, such as mentions of exams.

There are different definitions of temporal annotation schemes for clinical and general domains. Even in the clinical domain, depending on the clinical text type, medical specialty, and task extraction objective, the definitions are different. For instance, if the objective is to extract drug-adverse event (DAE) patterns from clinical texts, the events could be restricted to medications and experienced symptoms. Additionally, temporal expressions could be restricted to only precise dates, and a reduced temporal relation set could be used.

Different annotation schemes will have a temporal relation set based on the annotation requirements. For instance, the temporal relation OVERLAP is generic, implying that the two mentions somehow overlap. However, specific relations such as IDENTITY and SIMULTANEOUS indicate a particular OVERLAP case in which both events coincide, having the same start and endpoints. There is a trade-off between the amount of information represented by a relation set and the task complexity in both the annotation and extraction steps. A more elaborate temporal relation set may enable a more accurate representation of the temporal information. However, the annotators may need to distinguish between temporal relation types with slightly different concepts, which may cause disagreements and create a low number of annotations for certain relation types.

Additionally, the information necessary to distinguish between close temporal relation types may not be mentioned in the text or may need specific knowledge or interpretation. In the clinical domain, we often see that text writing quality and size—as certain clinical text types are short and objective—may limit an extended set of relations because of the number of disagreements and implicit information. For instance, several temporal relation types were annotated in the THYME corpus according to the THYME-TimeML scheme. However, only the CONTAINS relation type was used in Clinical TempEval shared tasks because of the low number of annotations for the other relation types. To provide a proper visualization of this aspect, we marked all relations annotated but not used during Clinical TempEval shared tasks in gray and the used relations in bold.

The same happened for the i2b2 2012 shared task annotation process, another essential corpus for temporal relation extraction in the clinical domain. The corpus was annotated with an extended set of relations (Table 1, column i2b2 2012 schema): BEFORE, BEFORE/OVERLAP, OVERLAPS, DURING, ENDS_BY, AFTER, BEGINS_BY, and SIMULTANEOUS. However, owing to a low IAA and a low number of annotations for specific types, the shared task's relation set was restricted to AFTER, BEFORE, and OVERLAP. To provide a proper visualization of this aspect, we marked all relations that were annotated but not used during the i2b2 2012 shared task in gray and the used relations in bold. Additional details regarding shared-task datasets are provided in Section 3.3.

Thus, an extended relation set is ideal, but the trade-off between temporal information and task complexity must be considered.

2.2 Temporal Relation Extraction Example

In this section, we provide an example in Figure 1 to justify the benefit of extracting temporality from clinical texts. In this example, we show the different temporal extraction levels applied to the same sentences. The sentences were created to simulate sentences written during a patient's clinical consultation with a cardiologist.

In this example, the patient had a history of hypertension and myocardial infarction. The patient underwent two procedures: myocardial revascularization and transluminal angioplasty. In the patient action plan, the medication Selozok was prescribed. These events are specific to cardiology.

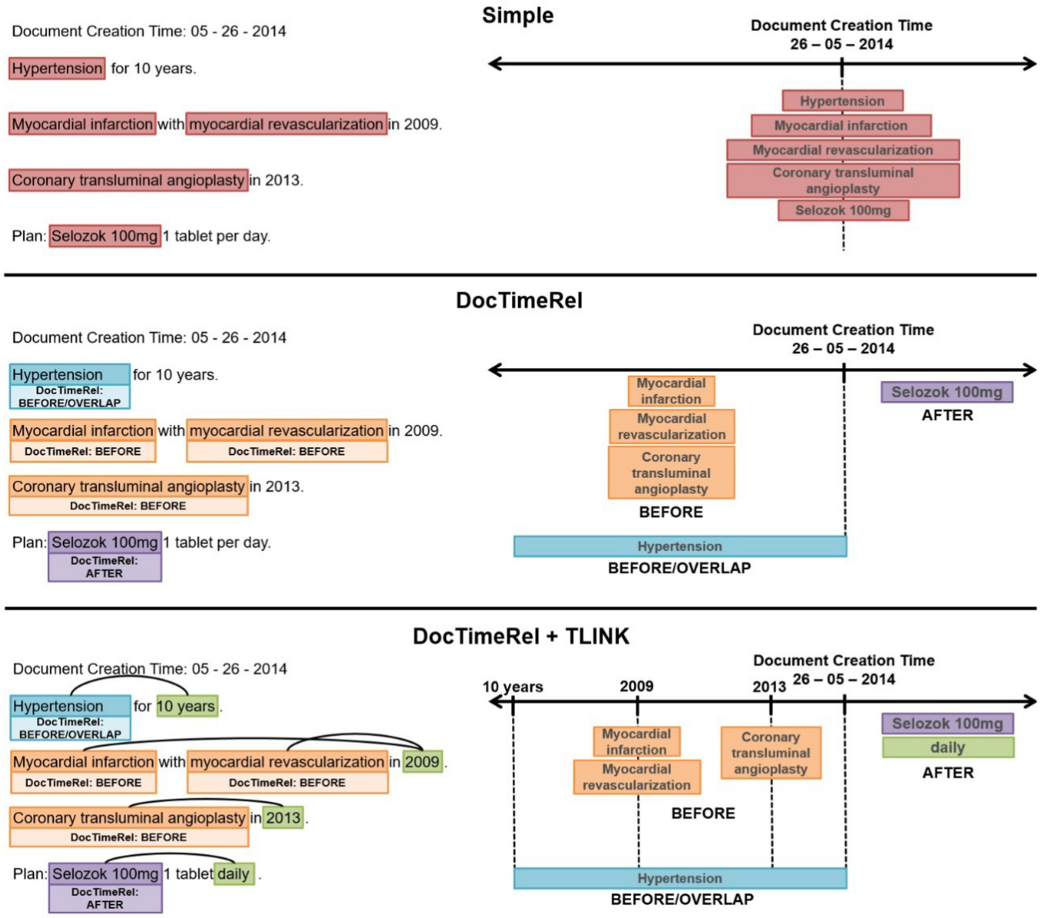


Fig. 1. Example of the benefit of temporal relation extraction levels.

Each specialty will have events—such as particular medical problems, symptoms, treatments, medicaments, and exams—that are not commonly mentioned in other medical specialties.

Using a simple approach of merely connecting every event to its DCT (Figure 1, Simple row), we cannot infer any order. Thus, in this scenario, all events coincided with time, which is not valid. Of course, this can provide sufficient information for specific temporal relation extraction tasks, especially when dealing with substantial clinical texts in a no-annotation scenario, using only automatically generated annotations by frameworks such as cTAKES [29] and Metamap [30].

By adding more information with the annotation of DocTimeRel relations (Figure 1, DocTimeRel row), we can provide a coarse ordering. In this example, we use the categories BEFORE, OVERLAP, AFTER, and BEFORE/OVERLAP in THYME-TimeML, for didactical explanations. Unlike before, it is shown that myocardial infarction, myocardial revascularization, and transluminal angioplasty are related to the patient's past medical history because they were annotated as BEFORE. Additionally, we can infer hypertension as a condition from the patient's past that still occurs during the DCT, a BEFORE/OVERLAP annotation, demonstrating the characteristics of chronic diseases. Further, it is evidenced that the medication, Sezolok, is related to the patient's future because of the AFTER annotation.

Table 2. Inclusion and Exclusion Criteria

Criteria	Inclusion criteria	Exclusion Criteria
Title and Abstract	Must mention temporality extraction in the abstract Must mention working with clinical free text	Review or update articles Articles not written in English
Full text	Must provide information about the method used to address temporality extraction Must provide at least one quantitative measure to evaluate the experiments	Do not provide information about the dataset size and data source

However, DocTimeRel is too generic for certain temporal relation extraction studies. For instance, BEFORE categories are too extensive because they do not refer to a certain point or closed period but rather to a broader period. DocTimeRel relation usage enables some event ordering, as not each event or temporal expression has associated TLINKs, but TLINKs provide a more detailed representation.

Adding TLINKs (Figure 1, DocTimeRel + TLINK row) to anchor events to specific periods of time represented by temporal expressions improves the timeline representation. For example, it is now evident that both myocardial infarction and myocardial revascularization occurred in 2009. However, as indicated, the coronary transluminal angioplasty happened only 4 years later, in 2013. Temporal expressions referring to dates in the patient's medical history can be underspecified, not containing all information required for normalization (year, month, and day information). In this example, both 2009 and 2013 are underspecified temporal expressions.

It is evident that the patient had hypertension for 10 years. This period of 10 years is somewhat imprecise because it does not reflect a specific period, being only an approximation. Furthermore, the medication is associated with its frequency, which is daily. Temporal expressions regarding medication frequency can be tricky and specific to the clinical domain. For instance, we could have the same medication with different dosages on different days of the week, or expressions such as b.i.d. (twice a day) and q.i.d. (once a day) from Latin, which indicates frequency.

3 METHODOLOGY

PubMed Central (MedLine), ScienceDirect, and ACL Anthology databases were selected for this review. The inclusion and exclusion criteria for the title and abstract analysis and the full-text analysis are provided in Table 2. The search expression was: ("temporal relation" OR "temporal relations" OR "temporal extraction" OR "temporal information" OR "temporal relationship" OR "temporal relationships" OR "timeline") AND ("clinical text" OR "clinical texts" OR "clinical narratives" OR "clinical narrative" OR "clinical reports" OR "clinical report").

We considered all published articles till October 23, 2020, with no limitations on the publication year. The PRISMA flow diagram is shown in Figure 2. With the search expression, we retrieved 2,728 articles: 1,232 from PubMed, 917 from ScienceDirect, and 579 from ACL Anthology. We identified 22 additional articles relevant to the review's scope by reading the selected articles and their references. From these 2,750 articles, 171 duplicated articles were excluded. The 2,579 remaining articles were subjected to a title and abstract analysis, followed by a full-text analysis. After analyzing the title and abstract, we selected 229 articles; after the full-text analysis, only 105 remained. All 105 articles were analyzed, and the most important ones are summarized in the tables. Important studies were defined as those that could be directly compared to infer the most effective strategies. In the tables, we have divided the approaches by dataset, sorted them by their performance, and visually emphasized those with the highest performance in bold.

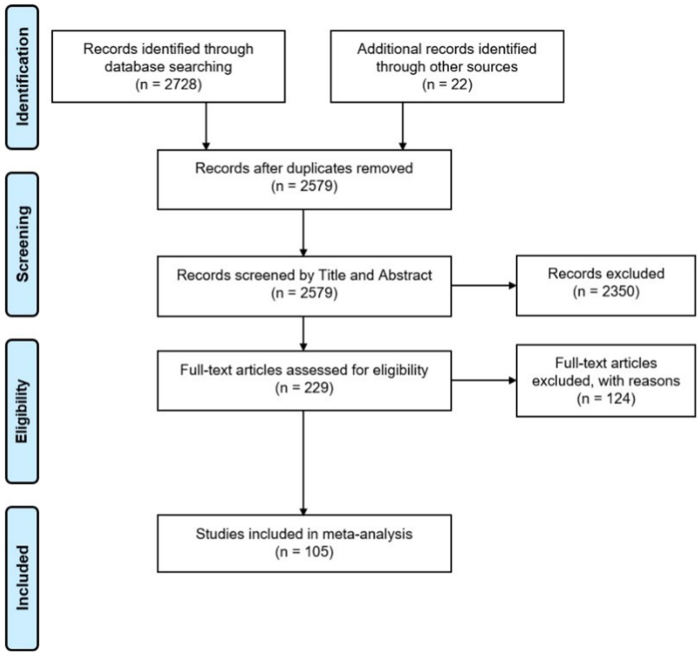


Fig. 2. Methodological steps used for this systematic review.

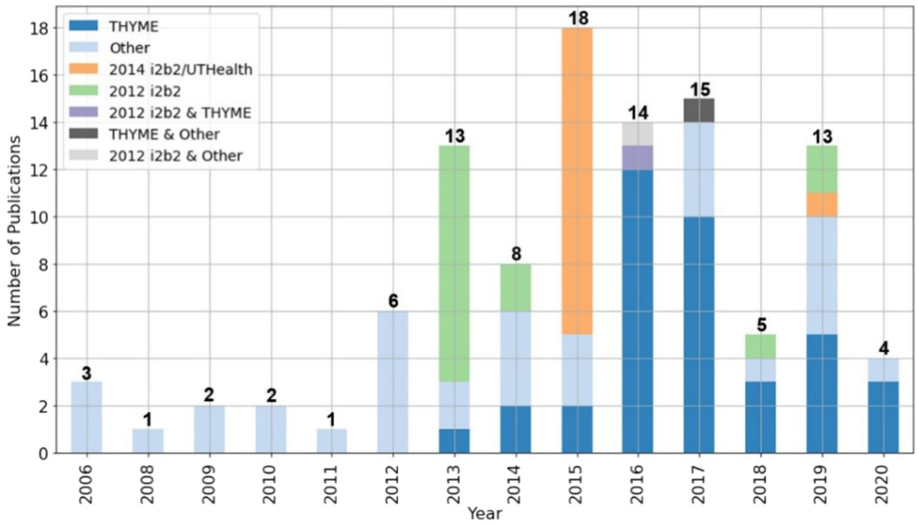


Fig. 3. Number of publications by year, separated by dataset, in chronological order.

3.1 Global Quantitative Results

Figure 3 shows the number of publications by year according to the dataset, differentiating the datasets available to the community through shared tasks from others. There were studies before 2013, but an increase in the number of publications occurred in 2013 following the i2b2 2012 challenge. Furthermore, in 2015, most of the publications were related to the i2b2/UTHealth 2014

challenge dataset, which focused on the extraction of risk factors for cardiovascular diseases with a multi-label DocTimeRel extraction task.

Of the 105 reviewed publications, 70 dealt with shared-task datasets. Except for [31] and [32], which additionally used another dataset, the other 68 publications only used shared task–related datasets. Hence, most publications were related to shared task datasets. Further, 17 were related to the i2b2 2012 dataset, 14 to the 2014 i2b2/Health dataset, and 40 to the THYME corpus.

In 2013 and 2014, there were some preliminary studies on the THYME corpus [6, 33, 34]. However, the number of publications related to the THYME corpus has started to grow with the Clinical TempEval challenges. In Clinical TempEval 2015, only two teams participated in the shared task owing to the long authorization process. Therefore, only two publications were related to the THYME corpus in 2015. Nevertheless, with the Clinical TempEval 2016 and Clinical TempEval 2017 challenges, the number of publications related to the THYME corpus increased from 2016 onward.

Publications on the i2b2 2012 dataset have been around since 2013 and those regarding the i2b2/UTHealth 2014 dataset have been around since 2014. However, since 2016, only [35] and [36] attempted to improve the reported results over these datasets, in contrast to [37] and [38], by focusing on specific relation types. The datasets used to push the state-of-the-art are the 2016 and 2017 editions of Clinical TempEval, especially the 2016 edition. The 2017 edition was aimed at cross-domain extraction with different training and testing data domains.

Most of the selected publications involved corpora written in English: 94 of the 105 reviewed articles. Publications in languages other than English include [39–42] in Chinese, [43, 44] in Korean, [45] in Dutch, [46] in Italian, [47] in Swedish, and [48] in Spanish. Additionally, [31] dealt with English and French, extracting temporal relations from both the THYME corpus and the MERLOT corpus [49], which are from medical texts in French. One can conclude that there is room for research in languages other than English.

3.2 Datasets

In this section, we provide the details of the datasets. Table 3 describes each dataset, providing information about the data origin and clinical document type, temporal annotation schema, dimension, and all related studies among the reviewed articles.

For the Clinical TempEval datasets, there is a clear difference between TLINKs and DocTimeRel, with separate annotations and evaluations in the evaluation script. DocTimeRel is considered an event attribute, with one DocTimeRel annotation for every event. The Clinical TempEval 2015 dataset contains 440 documents, averaging 136.05 events, 13.43 temporal expressions, and 37.43 TLINKs per document. The Clinical TempEval 2016 has more annotated data, with a total of 591 documents, averaging 133.42 events, 13.30 temporal expressions, and 39.33 TLINKs. The aim shifted toward a cross-domain extraction from Clinical TempEval 2017 with different training and testing domains. The Clinical TempEval 2017 dataset comprises 769 documents, averaging 120.83 events, 12.70 temporal expressions, and 33.28 TLINKs per document.

The i2b2/UTHealth 2014 challenge [126] was related to heart disease mentions and focused on discovering potential risk factors. However, there was no separate evaluation of temporality extraction.

4 EXTRACTION OF DOCTIMEREL RELATIONS

In this section, we analyze the articles that extracted DocTimeRel relations. We consider DocTimeRel relations between the event and a DCT, even if the authors identify only a relation between these two arguments, not classifying it into any category. We sort the results according to the strategy used to deal with temporality with three categories: rule-based systems (see Section 4.1), machine learning systems (see Section 4.2), and hybrid systems (see Section 4.3). As

Table 3. Datasets

Dataset description	Temporal Annotation (labels/categories)	Dimension (Train: Test)	Related studies
Reports from Stanford Translational Research Integrated Database Environment (STRIDE)	—	—	[50–52]
Reports from Palo Alto Medical Foundation (PAMF) dataset	—	—	[50]
Reports from Synthetic Derivative (SD) database	—	2,268 patients (1,512:759)	[53]
Reports from patients in the Intensive Care Unit (ICU)	—	1,040 reports (5-fold cross-validation)	[54]
Training Reports from Mayo Clinic sick-child daycare program Testing: Reports from Mayo Clinic pediatric patients	—	237 patients (125:112)	[55]
Reports from diverse types from the University of Pittsburgh Medical Center's MARS repository	DocTimeRel (HISTORICAL, RECENT, HYPOTHETICAL)	240 reports with 4,654 annotations (2,377:2,277)	[56]
General practitioner entries, specialist letters, radiology reports, and discharge letters from the Erasmus Medical Center (EMC) corpus	DocTimeRel (HISTORICAL, RECENT, HYPOTHETICAL)	7,500 reports (3,750:3,750)	[45]
Patient reports from Clinical e-Science Framework Services (CLEF-S) project	DocTimeRel (BEFORE, AFTER, IS_INCLUDED) and TLINK (BEFORE, AFTER, IS_INCLUDED)	98 reports	[57]
Reports from the Research Patient Data Repository of Partners Healthcare (i2b2/UTHealth 2014 shared task)	DocTimeRel (BEFORE, AFTER, DURING) multi-label	1,304 reports (790:514)	[36, 58–70]
Discharge summaries from Partners Healthcare and the Beth Israel Deaconess Medical Center (i2b2 2012 shared task)	DocTimeRel (BEFORE, OVERLAP, AFTER) and TLINK (BEFORE, OVERLAP and AFTER)	310 reports (190:120)	[7, 11, 12, 23, 32, 35, 37, 38, 71–76, 127, 130, 135]
Reports from Stockholm adverse drug event (ADE) corpus	DocTimeRel (PAST, FUTURE)	400 reports (320:80)	[47]
Reports from diverse types from the University of Pittsburgh Medical Center	DocTimeRel (HISTORICAL, RECENT)	42 reports	[78]
Reports from diverse types of MRSA cases	DocTimeRel (WAY BEFORE ADMISSION, BEFORE ADMISSION, ON ADMISSION, AFTER ADMISSION, AFTER DISCHARGE)	51 reports (10-fold cross-validation)	[80]
Reports	DocTimeRel (WAY BEFORE ADMISSION, BEFORE ADMISSION, ON ADMISSION, AFTER ADMISSION, AFTER DISCHARGE)	1,613 medical concepts (968:645)	[81]
Clinical notes and pathology reports from colon cancer patients from the Mayo Clinic (THYME corpus)	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER)	78 reports	[33, 34, 82]
Clinical notes and pathology reports from colon cancer patients from the Mayo Clinic (THYME corpus)	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER) and TLINK (CONTAINS, OVERLAP, BEFORE, BEGINS_ON, ENDS_ON)	107 reports	[6]
Clinical notes and pathology reports from patients with colon cancer from the Mayo Clinic (THYME corpus—**Clinical TempEval 2015 shared task**)	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER) and TLINK (CONTAINS)	440 reports (293:147)	[12, 128]
Clinical notes and pathology reports from patients with colon cancer from the Mayo Clinic (THYME corpus—**Clinical TempEval 2016 shared task**)	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER) and TLINK (CONTAINS)	590 reports (440:150)	[28, 31, 77, 83–107]

(Continued)

Table 3. Continued

Dataset description	Temporal Annotation (labels/categories)	Dimension (Train: Test)	Related studies
Clinical notes and pathology reports from patients with colon and brain cancer from Mayo Clinic (THYME corpus—**Clinical TempEval 2017 shared task**)	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER) and TLINK (CONTAINS)	759 reports (621:148)	[77, 99, 102, 103, 108–113]
Reports from Gastroenterology, Hepatology, and Nutrition departments (MERLOT corpus)	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER) and TLINK (CONTAINS)	500 reports	[31]
Cardiology texts from Molecular Cardiology Laboratories of the Istituti Clinici Scientifici Maugeri (ICSM) hospital	DocTimeRel (BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER)	75 reports (60:15)	[46]
Reports	DocTimeRel (CURRENT, HISTORY, FUTURE, UNKNOWN)	1,089 reports	[114]
Discharge summaries, and clinical progress notes from the cardiovascular diseases risk factor corpus (CVDsRFC)	DocTimeRel (CONTINUING, DURING, BEFORE, AFTER)	600 reports (420:180)	[40]
Reports in the Spanish language	—	200 reports	[48]
Reports from patients in the ICU from the Royal Prince Alfred Hospital	—	200 reports (10-fold cross-validation)	[115]
Reports	—	200 patients	[116]
Reports from Mayo Clinic	—	20 patients	[117]
Reports from Mayo Clinic	—	1507 patients	[118]
Reports from Record Interactive Search (CRIS) database	—	70 reports	[119]
Vaccine Adverse Event Reporting System (VAERS) reports and US Food and Drug Administration (FDA) Adverse Event Reporting System (FAERS) reports	—	140 reports	[32]
Discharge summaries from the Seoul National University Hospital HER	—	200 reports (170:30)	[44]
Reports from Guangdong Provincial Hospital of Traditional Chinese Medicine (GPHTCM)	—	24,817 reports (24,417:400)	[39]
Discharge summaries from the University Hospital in Korea	—	150 reports	[43]
Discharge summaries	—	354 reports	[120]
Discharge summaries from Columbia University Medical Center	—	20 reports	[121]
Reports from the MMIC-II dataset	—	100 reports	[122]
Discharge summaries from the *New England Journal of Medicine* (*NEJM*)	TLINK (AFTER, BEFORE, INCOMPARABLE)	60 reports	[123, 124]
Reports from diverse types	TLINK (AFTER, BEFORE, OVERLAP)	80 reports (cross-validation)	[125]
Reports	TLINK (BEGINS, END, SIMULTANEOUS, INCLUDES, BEFORE)	47 reports (10-fold cross-validation)	[79]
Reports from a hospital in China	TLINK (SIMULTANEOUS, BEFORE, AFTER)	563 reports (413:150)	[41]
Reports of diverse types from a hospital in China	—	100 patients	[42]

We highlight the datasets used in shared tasks in bold. The i2b2 2012 dataset contains 310 discharge summaries, averaging 86.6 events, 12.4 temporal expressions, and 176 TLINKs per note. There is no distinction between TLINKs and DocTimeRel in the annotations; the shared-task evaluation script evaluated them jointly. Thus, it is difficult to evaluate each temporal relation category's contribution in the final result.

Table 4. Articles Related to DocTimeRel that Used Full Rule-Based Systems

Authors	Best Strategy	SE	Results
	I2b2 2012 dataset		
[71]	Rules		Fm 0.5628
[32]	Rules	√	Match ratio 0.69 (NTC)
	I2b2/UTHealth 2014 dataset		
[70]	df + specific rules		Fm 0.915
[67]	Rules		Fm 0.907
[58]	df + context-aware refinement approach		Fm 0.897
[69]	Df		Fm 0.890
[64]	df + specific rules		Fm 0.8776
[65]	Df		Fm 0.875

Legend: SE, separated evaluation; AUROC, area under the receiver operator curve; DAE, drug-adverse events; DCT, document creation time; DDI, drug-drug interaction; Fm, F-measure; ICU, intensive care unit; Regex, regular expression; TRE, temporal relation extraction; RFE, risk factor extraction; df, "default value" strategy; NTC, not comparable.

we provide an in-depth overview and evaluation of all selected articles, we compiled a summary (see Section 4.4) with highlights and conclusions.

4.1 Rule-Based Systems

Systems that exclusively extracted DocTimeRel relations with a rule-based approach are listed in Table 4. The table contains information about the article's primary objective, the strategy used to extract temporality, the obtained results, and an indicator of a separate evaluation for the temporality extraction with a separate evaluation (SE) column. If the article had a separate evaluation for temporality, the obtained results were related to the extraction. Otherwise, the obtained results were related to the system's primary objective.

Rule-based systems can be divided into two types: (i) those that only identify a relationship between the event and the DCT by connecting both and (ii) those that also classify the relation into a category.

The first type is usually associated with systems in which the temporal extraction is just a step in the information extraction methodology, and no complex temporal information is required. The authors of [48, 50–55] used this approach.

One of the research topics in which this strategy has been widely used is the identification of adverse events with the extraction of drug-drug interactions (DDIs) and drug-adverse events. The authors of [50] and [51] focused on creating a timeline for each patient, using statistics to extract DDIs, and compared the results with structured data, the standard information source. In contrast, the authors of [52] created a framework to differentiate DAE mentions from indications by creating pairs of drug diseases. The authors of [53] also extracted DAE but restricted it to interactions between clopidogrel and mentions of bleeding, using a temporal feature based on the difference between the mentions' DCTs. Another research topic was the identification of occurrence dates for a specific condition in patient longitudinal data, for example, pneumonia in [54], and asthma in [55].

The second type provides more detailed temporal information, where the task difficulty depends on the number of categories; this is because each category needs its own specific rules. All remaining articles mentioned next are from the second type.

We highlight ConText [56], a regular-expression-based tool to extract event attributes. It extracts the experiencer, negation, and temporality (DocTimeRel). Experiencers and negation are relevant for extraction because they completely change the event's context. There was an

adaptation of ConText to Dutch with additional rules, named ConTextD [45]. The authors of [57] present a preliminary CLEF study on temporal extraction.

For the i2b2/UTHealth challenge dataset, rules were popular for extracting DocTimeRel as the scope of events was limited to specific predefined risk factors, facilitating the creation of rules. For corpora such as the THYME corpus and i2b2 2012 corpus, where there is a wide range of events with different characteristics, the creation of rules is more challenging and prone to overfitting.

A widely used strategy for the i2b2/UTHealth 2014 challenge was to use the most frequent label in the training set for each risk favor, the default value strategy. This strategy was used alone or in combination with additional rules to deal with specific cases. This strategy was used alone in [65], with superior results in the training set over ConText. Additional rules were used in [58, 64, 69, 70]. However, the authors of [69] verified that the system's most significant error source was attribute extraction, mainly, the DocTimeRel component. The authors of [67] used rules but did not rely on the default value strategy, creating rules based on observations on the training set and the ConText output.

For the i2b2 2012 dataset, the authors of [32] and [71] used rule-based systems. However, their performance was inferior to machine learning–based or hybrid systems.

There was no separate evaluation of temporal relations in the i2b2/UTHealth 2014 challenge script. However, we believe that a well-constructed machine learning–based system or hybrid system outperforms rule-based systems. Additionally, rule-based systems are not robust enough to deal with datasets in which event annotations involve several different aspects (treatments, symptoms, medical problems, and exams) and ensure generalization.

4.2 Machine Learning

This section analyzes the articles that used machine learning–based systems for DocTimeRel (summarized in Table 5).

Unlike the previous section, all machine learning–based approaches identified the relation and classified it into a specific category. In addition to DocTimeRel, the authors of [47] extracted attributes such as negation and speculation. In contrast, the authors of [82] used only DocTimeRel as a feature for DAE identification, with the temporal feature improving the classification results. The DocTimeRel relation was also used as a feature in [81] for co-reference resolution. The DocTimeRel system was developed in [80], based on a CRF classifier.

There was a preference for support vector machines (SVMs) and conditional random fields (CRFs) among the traditional machine learning classifiers. However, other machine learning classifiers were also used in the reviewed publications: random forest (RF) classifiers by the authors of [47] and [106], RIPPER classifiers by the authors of [68] and [78], OneRule classifiers by the authors of [62], and logistic regression (LR) classifiers by the authors of [86].

The use of CRF classifiers is widespread in shared task–related and regular datasets. One of the advantages of CRF is the possibility of extracting the entities and classifying the relation simultaneously in a sequence-labeling task with a single classifier. A single CRF classifier for DocTimeRel extraction was used in regular datasets by the authors of [80] and [81]. For the i2b2 2012 dataset, the authors of [73] used a single CRF classifier, even in a scenario with two DCTs (admission and discharge dates). For the Clinical TempEval shared tasks, a single CRF classifier was used by the authors of [28, 84, 85, 91, 111, 112, 128].

When considering the number of publications here and the hybrid systems sections, SVM was the most used machine learning algorithm. For shared task–related datasets, the SVMs held or maintained the best performance. In regular datasets, a single SVM classifier was used by the authors of [46] and [82]. For the i2b2/UTHealth 2014 dataset, several publications used SVM classifiers, but most combined it with rules or the default value strategy. The exception was [59], which

Table 5. Articles Related to DocTimeRel that Used Machine Learning Systems

Authors	Best strategy	SE	Results
I2b2 2012 dataset			
[72]	SVM clf		**Fm 0.6954**
[12]	2 SVM clfs		**Fm 0.695**
[11]	2 SVM clfs		Fm 0.6932
[73]	CRF clf		Fm 0.693
[127]	2 SVM clfs		Fm 0.6849
I2b2/UTHealth 2014 dataset			
[59]	label-powerset strategy + SVM clfs		**Fm 0.9268**
[68]	21 RIPPER clfs + voting		**Fm 0.9185**
[36]	BI-LSTM		Fm 0.9081
[62]	OneRule clfs		Fm 0.857
Clinical TempEval 2015 dataset			
[12]	SVM clf	√	**Fm 0.807**
[128]	CRF clf	√	Fm 0.791
Clinical TempEval 2016 dataset			
[103]	BERT + MTL		Fm 0.86 (NTC)
[31]	SVM clf	√	**Fm 0.87**
[97]	Structured perceptron + ILP	√	**Fm 0.846**
[28]	CRF clf	√	Fm 0.844
[94]	SVM clf	√	Fm 0.835
[86]	LR clfs	√	Fm 0.815
[106]	RF clf	√	Fm 0.807
[98]	CNN + MLP	√	Fm 0.788
[85]	CRF clf	√	Fm 0.714
[84]	CRF clf	√	Fm 0.712
[91]	CRF clf	√	Fm 0.687
Clinical TempEval 2017 dataset			
[113]	SVM clf	√	**Fm 0.519 UDA, 0.591 SDA**
[109]	Structured perceptron + ILP	√	**Fm 0.49 UDA, 0.56 SDA**
[112]	CRF clf	√	Fm 0.45 UDA, 0.52 SDA
[111]	CRF clf	√	Fm 0.40 UDA, 0.50 SDA
[108]	SVM clf	√	Fm 0.49 SDA
[110]	RNNs	√	Fm 0.32 SDA

Legend: SE, separate evaluation; DAE, drug adverse events; IE, information extraction; RF, random forest; clf, classifier; Fm, f-measure; ML, machine learning; CRF, conditional random fields; AD, after discharge; BA, before admission; OA, on admission; WBA, way before admission; AA, after admission; TRE, temporal relation extraction; SVM, support vector machine; GRU, gated recurrent unit; ATT, attention; CNN, convolutional neural network; RFE, risk factor extraction; LSTM, long short-term memory; MTL, multitask learning; ILP, integer linear programming; LR, logistic regression; MLP, multilayer perceptron; UDA, unsupervised domain adaptation; SDA, supervised domain adaptation; RNN, recurrent neural network; NTC, not comparable.

used a label-powerset strategy to transform the multi-label classification into several binary classification tasks addressed with SVM classifiers.

For the i2b2 2012 dataset, the authors of [72] used a single SVM classifier, while the authors of [11], [12], and [127] involved two SVM classifiers, one for each DCT. Additionally, the authors

Table 6. Features Used by the Machine Learning Systems

Feature	Explanation
Nearby tokens	Nearby tokens around the mention over a token window (e.g., 2 or 3 tokens)
Tense	Tense of the verbs in the same sentence of the mention
Nearby events	Surrounding events with their respective attributes
Nearby temporal expressions	Surrounding temporal expressions and their respective attributes
Nearby part-of-speech (POS) tags	Nearby POS around the mention over a token window (e.g., 2 or 3 tokens)
Event information	Event tokens, POS tags, and attributes (e.g., category and polarity)
Event position	Event position in the document, generally associated with the section header (e.g., medical history)
Lexicon searching	Semantic features based on search for event terms in crafted lexicons or the Unified Medical Language System (UMLS)

of [127] also used two SVM classifiers for relations between temporal expressions and the DCT, which most authors ignored.

For preliminary THYME corpus studies and Clinical TempEval–related articles, a single SVM was used by the authors of [6, 12, 31, 94, 108, 113]. Among these, we highlight [12] and [31]. The first developed a system with features that were fully extracted by cTAKES and experimented with both i2b2 2012 and Clinical TempEval 2015 datasets. The second experiment was conducted in a multilingual setting, extracting DocTimeRel from the MERLOT (French) corpus and THYME corpus.

The classifiers that achieved the best performance used features to better understand the context and the event. The features generally associated with the best performing systems are listed in Table 6.

Owing to the clinical text characteristics, specialized tools to preprocess the text and generate features are widely used. For instance, cTAKES provides several components, such as a sentence boundary detector, tokenizer, and part-of-speech tagger. Further, semantic features can be obtained by cTAKES, named entity recognition (NER) components, or by mapping tools such as Metamap. A more detailed analysis of the specialized tools commonly used in the clinical domain can be found in [21].

Over the years, approaches based on deep learning have emerged. For the Clinical TempEval 2016 dataset, the authors of [98] used a convolutional neural network (CNN) with a multilayer perceptron (MLP). For the 2017 edition dataset, the authors of [110] used a recurrent neural network (RNN) classifier for each relation type. For the i2b2/UTHealth 2014 dataset, the authors of [36] jointly extracted the entities and DocTimeRel with a bidirectional long short-term memory (Bi-LSTM)–based architecture. In addition to Bi-LSTM, they tested standard RNNs, CNNs, and LSTMs, achieving superior results with Bi-LSTM. However, the results were still not comparable with those of traditional machine learning algorithms. For regular datasets, the authors of [40] extracted risk factors for cardiovascular diseases, similar to the i2b2/UTHealth 2014 shared-task objective, using a CNN-based model.

All of these approaches dealt only with the DocTimeRel task. However, some authors developed frameworks that jointly predicted DocTimeRel with another NLP task. For instance, the authors of [114] proposed a framework based on GRU, deep residual networks, and attention to jointly predict DocTimeRel and presence attributes. The authors of [97] and [109] focused on structured machine learning, jointly predicting DocTimeRel and TLINKs using a structured perceptron model and integer linear programming (ILP) and achieving consistent results over Clinical TempEval

Table 7. Articles Related to DocTimeRel that Used Hybrid Systems

Authors	Best strategy	SE	Results
	I2b2 2012 dataset		
[7]	SVM clf + rules		Fm 0.63
[130]	SVM clf + rules		Fm 0.5594
	I2b2/UTHealth 2014 dataset		
[66]	3 SVM clfs + df + rules + ann refinement		**Fm 0.9277**
[60]	CART DT + df		Fm 0.917
[61]	Markov networks + rules		Fm 0.9098
[63]	NB clf + rules		Fm 0.8302
	Clinical TempEval 2016 dataset		
[88]	LR clf + rules	√	Fm 0.743

Legend: SE, separate evaluation; TRE, temporal relation extraction; SVM, support vector machine; clf, classifier; RFE, risk factor extraction; df, default value strategy; ann, annotation; DT, decision tree; NB, naïve Bayes; LR, logistic regression.

2016 and 2017 datasets. Recently, the authors of [103] have developed a one-pass model based on bidirectional encoder representations from transformers (BERT) [129] that leverages global embeddings to jointly predict TLINKs and DocTimeRel. As the system was developed at the entity level, considering both events and temporal expressions as inputs, the model had to classify the entity into the BEFORE, OVERLAP, BEFORE/OVERLAP, AFTER, and TIMEX3 categories, where the first four events are DocTimeRel categories and the last is a label only for time-related entities.

4.3 Hybrid Systems

This section analyzes the articles that used hybrid systems for DocTimeRel (summarized in Table 7).

There were fewer hybrid systems than rule-based and machine learning–based systems. Considering the i2b2 2012 dataset, the authors of [7] and [130] used an SVM classifier with crafted rules. The authors of [88] used an LR classifier with craft rules for the Clinical TempEval 2016 dataset. However, systems relying only on machine learning achieved superior results for these datasets.

The i2b2/UTHealth 2014 dataset had predefined risk factor categories with specific patterns over the training set. Hence, the authors widely used both the default value strategy and manually crafted rules. Manually crafted rules were also used by the authors of [61] and [63] to complement machine learning classifiers. The default value strategy was used by the authors of [60] to complement machine-learning classifiers. Additionally, the authors of [66] used manually crafted rules, default value strategy, and SVM classifiers but improved the performance by annotating the training set, providing a grainier set of annotations.

Thus, there is no evidence that hybrid systems are superior to systems that rely only on machine learning, especially in a scenario with no predefined categories and predominant labels over the training set.

4.4 DocTimeRel Conclusions

Connecting an event to its corresponding DCT can provide sufficient temporal information for specific extraction tasks. However, a more detailed representation that also classifies the relation into specific categories provides additional temporal information at the cost of increasing task difficulty. For instance, the DocTimeRel categories in [47] were PAST and FUTURE, while

the THYME corpus categories were BEFORE, AFTER, OVERLAP, and BEFORE/OVERLAP. In addition to having more categories to classify events, differentiating between them can become complicated because they depend on textual tips and clinical knowledge. For instance, to label BEFORE/OVERLAP, the event's continuity until the DCT must be ensured.

Rule-based systems or hybrid systems were adequate for the i2b2/UTHealth 2014 dataset. However, for datasets such as the i2b2 2012 and THYME corpus, rule coverage would be low because of different patterns and event diversity over the text. SVM and CRF classifiers are widely used for traditional machine learning, with SVM providing slightly superior results for the analyzed datasets. DocTimeRel is a classification problem with well-defined categories, and a feature set that leverages information about the entity and context and does not require an extensive set of features can achieve excellent performance. Among the best performing systems, we noted a preference for specialized tools such as cTAKES, which provides several components, such as a sentence boundary detector, tokenizer, and part-of-speech tagger. We highlight the SVM-based approaches of [12] and [31]. The authors of [12] developed a system with features that were fully extracted by cTAKES while conducting experiments on two datasets, and the authors of [31] conducted experiments in a multilingual setting, developing systems for French and English languages.

Recent publications have experimented with CNN-, LSTM-, Bi-LSTM-, and attention-based models. However, the volume of research over the last few years is far from TLINK extraction, and most of the approaches are not directly comparable to previous studies because of different evaluation settings. In addition to SVM, we highlight the multi-task learning (MTL) approaches of [97], [109], and [103], which jointly predict DocTimeRel and TLINK relations. The first two developed a system based on the structured perceptron model and ILP, while the third developed a one-pass model based on BERT. Because TLINK and DocTimeRel extraction are dependent tasks, joint learning can improve results. For instance, if one event has a DocTimeRel annotation of BEFORE and another one has AFTER, no TLINK should be marked between them when we consider the THYME annotation guidelines.

5 EXTRACTION OF TLINK RELATIONS

In this section, we analyze the articles that extracted TLINK relations. We sort the results according to the strategy used to deal with temporality with three categories: rule-based systems (Section 5.1), machine learning systems (Section 5.2), and hybrid systems (Section 5.3). As we provide an in-depth overview and evaluation of all selected articles, we compiled a summary (Section 5.4) with highlights and conclusions.

5.1 Rule-Based Systems

In rule-based systems, the creation of candidate pairs to feed the classifiers is not needed. Therefore, the aspects of creating rules for entities that are on a single sentence or abroad sentences are underspecified.

A strategy used by the authors of [48, 57, 116–119] was to create rules to connect events to dates, a specific type of temporal expression that reflects calendar times [131]. The authors of [57] classified the relation into a specific type, which is an additional step in connecting the event to its respective date. In [119], rules jointly connected symptoms to dates and normalized the date mentions.

Developing strategies to deal with low-quality and noisy texts, which are common characteristics of clinical texts, has been addressed by the authors of [32] and [115]. The authors of [32] aimed at extracting temporal information from low-quality texts, such as medical product safety surveillance reports, connecting dates, and time intervals to events. Further, the authors of [115] focused on developing a question answering-system based on noisy texts.

The studies from [43] and [44] focused on creating temporal snippets of texts. The authors of [44] aimed at extracting temporal segments, where temporal segments were text segments containing topics with the same temporal or topical content. The authors of [43] extracted clinical semantic units (CSU), which are segments of text based on temporal expression position with rules. These segments contain events based on their position in the text. The CSUs were then classified into problem-action relations.

There are other relevant studies, such as [39, 120–122]. The authors of [39] created triples of events, temporal expressions, and descriptions, where descriptions were elaborations or outcomes. The authors of [120] adapted the TARSQI Toolkit, built for newswire texts, to clinical texts, predicting whether the patients were in statins when they were admitted. The authors of [121] evaluated the performance of a system developed to enable question answering from discharge summaries. The authors of [122] tested a temporal query system to identify acute kidney injuries of patients in intensive care units.

5.2 Machine Learning Systems

This section analyzes the articles that used machine learning–based systems for TLINKs (summarized in Table 8). Most articles in the TLINK extraction section were machine learning or hybrid systems.

Most articles here and in the next section (hybrid systems) are related to shared task datasets. The datasets were not shared-task related in [41, 42, 79, 123–125]. In [123] and [124], the focus was on ordering with temporal segments. The authors of [79] and [125] focused on ordering events by considering the relations between event mentions only. The authors of [79] tested both pairwise classification and event ranking, and achieved better ranking results. The authors of [41] focused on temporal indexing, predicting TLINKs between events and temporal expressions while keeping the most relevant pair for each event. The authors of [42] focused on extracting several entities and relations from clinical texts using a BERT model.

The remaining articles are detailed according to the candidate pair selection strategy and the approaches used to extract the relations. The task of training classifiers to extract relations consists of generating training samples. Positive samples are provided through annotations, but negative samples need to be generated by developers. For instance, a strategy for generating instances can be creating all possible pairs among the entities within a document. However, this approach would generate a much higher ratio of negative samples than positive samples, especially for datasets such as the THYME corpus and i2b2 2012 with diverse types of events and temporal expression annotations. There were close to 133 events and 13 temporal expression annotations per document in the Clinical TempEval 2016 dataset. Creating all possible pairs would be unrealistic, especially when considering the relations between events. Thus, the premise of temporal relation extraction is that it is not possible to cover all positive samples without creating too many negative samples. Hence, there is a trade-off between the number of positive samples covered and the number of negative samples generated.

A widely used strategy was to restrict within-sentence relations by considering all possible pairs within the same sentence. This strategy was used in [6, 12, 31, 33, 34, 77, 84, 86, 87, 89, 90, 99–101, 104, 108, 111]. Most of these studies are related to the THYME corpus, either initial publications about the THYME corpus or publications dealing with Clinical TempEval datasets. For the Clinical TempEval 2016 dataset, approximately 74% of the TLINKs in the training set were related to within-sentence relations. Hence, if the testing set follows the same pattern as the training set, approximately 26% of the positive instances would be false negatives because the frameworks would not predict these relations.

Table 8. Articles Related to TLINK that Used Machine Learning Systems

Authors	Best strategy	Candidate pair selection	WS	CS	NS	SE	Results
I2b2 2012 dataset							
[127]	WS: 3 SVM clfs; CS: 3 SVM clfs	WS: APP, CS: rules	√	√			**Fm 0.6849**
[35]	WS: NB clf; CS: NB clf	WS: rules, CS: rules	√	√			Fm 0.671
[130]	2 SVM clfs	Rules	√	√			Fm 0.5594
[37]	BI-LSTM; TS expansion	-			√		Fm 0.6217 (NTC)
Clinical TempEval 2015 dataset							
[12]	2 SVM clfs. CSL; TS expansion	WS: APP	√			√	Fm 0.321
Clinical TempEval 2016 dataset							
[103]	BERT; 3 class; MTL	APP over TK	√	√		√	**Fm 0.686**
[102]	BERT; TS expansion; 3-class	APP over TK	√	√		√	**Fm 0.684**
[107]	Context segmentation; Associated ATT; Position weights	APP over TK	√	√		√	Fm 0.643
[89]	WS: tree-based Bi-LSTM-RNN	WS: APP	√			√	Fm 0.633
[99]	Bi-LSTM; TS expansion; 3-class; XML markup	WS: APP	√			√	Fm 0.630
[90]	Tree-based Bi-LSTM-RNN	WS: APP	√			√	Fm 0.629
[95]	LSTM; MTL	APP over TK	√	√		√	Fm 0.628
[101]	SVM clf + CNN; XML markup	WS: APP	√			√	Fm 0.621
[105]	WS: Bi-LSTM; CS: Bi-LSTM; 3-class	WS: APP, CS: rules	√	√		√	Fm 0.613
[87]	2 CNNs; 3-class; XML markup	WS: APP	√			√	Fm 0.515 event-event, 0.700 event-time (NTC)
[77]	RNN; ATT; Piece representation	WS: APP	√			√	Fm 0.729 (NTC)
[104]	GRU + ATT; 3-class; XML markup	WS: APP	√			√	Fm 0.690 (NTC)
[97]	Structured perceptron; ILP; MTL	APP over TK + rules	√	√		√	Fm 0.608
[93]	Classifier ensemble; ILP	-			√	√	Fm 0.595
[100]	2 SVM clfs; TS expansion	WS: APP	√			√	Fm 0.594
[94]	WS: 2 SVM clfs; CS: 4 SVM clfs; CSL	WS: APP + pair filtering, CS: rules	√	√		√	Fm 0.573
[96]	2 SVM clfs	WS: APP + restrictions, CS: rules	√	√		√	Fm 0.551
[31]	SVM clf. 3-class	WS: APP	√			√	Fm 0.53
[28]	WS: 2 SVM clfs; CS: 2 SVM clfs	WS: APP, CS: rules	√	√		√	Fm 0.511
[86]	LR clf	WS: APP	√			√	Fm 0.506
[84]	2 CRF clfs	WS: APP	√			√	Fm 0.453
[85]	CRF clf	-			√	√	Fm 0.313
[83]	4 CRF clfs	WS: APP, CS: rules	√	√		√	Fm 0.264
[92]	List-Net [132]	-			√	√	MSE 0.072 (NTC)

(Continued)

Table 8. Continued

Authors	Best strategy	Candidate pair selection	WS	CS	NS	SE	Results
Clinical TempEval 2017 dataset							
[103]	BERT; 3 class; MTL	APP over TK	√	√		√	**Fm 0.582 UDA**
[102]	BERT; TS expansion; 3-class	APP over TK	√	√		√	**Fm 0.565 UDA**
[99]	Bi-LSTM; TS expansion; 3-class; XML markup	WS: APP	√			√	Fm 0.547 UDA
[77]	RNN; Attention; Piece representation	WS: APP	√			√	Fm 0.63 (NTC)
[111]	XGBoost clf	WS: APP	√			√	Fm 0.34 UDA, 0.25 SDA
[113]	WS: Bi-LSTM; CS: Bi-LTM; 3-class	WS: APP, CS: rules	√	√		√	Fm 0.328 UDA, 0.316 SDA
[109]	Structured perceptron; ILP; MTL	APP over TK	√	√		√	Fm 0.32 UDA, 0.28 SDA
[108]	2 SVM clfs	WS: APP	√			√	Fm 0.26 (SDA)
[112]	WS: 2 clf ensembles; CS: 2 clf ensembles; CSL	WS: APP, CS: rules. Pair filtering; rules	√	√		√	Fm 0.23 UDA, 0.15 SDA

Legend: WS, within-sentence; CS, cross-sentence; NS, not specified; SE, separate evaluation; TO, temporal ordering; ILP, integer linear programming; ME, maximum entropy; clf, classifier; Fm, f-measure; SVM, support vector machine; TRE, temporal relation extraction; APP, all possible pairs; LSTM, long short-term memory; CNN, convolutional neural network; NTC, not comparable; NB, naïve Bayes; TS, training set; CTE, Clinical TempEval; CSL, cost-sensitive learning; 3-class, transforming into a 3-class classification task; MSE, mean squared error; MTL, multi-task learning; TK, token window; ATT, attention; RNN, recurrent neural network; GRU, gated recurrent unit; LR, logistic regression; CRF, conditional random fields; UDA, unsupervised domain adaptation; SDA, supervised domain adaptation.

Some publications considered all possible pairs within a sentence and added specific heuristics to cover cross-sentence relations. This strategy was used in [28, 83, 94, 96, 105, 112, 113, 127]. The authors of [83] and [127] considered pairs between entities in neighbor sentences, restricting them to a one-sentence window. The authors of [105] and [113] also considered pairs between entities in neighbor sentences, but increased the range to a three-sentence window. The authors of [28] also defined heuristics based on a sentence window and considered entity position, such as considering only the first and last event mentions in the sentence. The strategy used in [112] was based on the previous studies in [28] and [94]. The authors of [96] added a restriction that considers all possible pairs in a sentence or line.

The approach of [94] created all possible pairs within a sentence but added restrictions to exclude those unlikely to have a relation. The filtering rules involved section information and event attributes. The same strategy was used to create pairs within a two-sentence window, but for pairs with greater sentence window values, additional rules were created to create fewer pairs.

A widespread strategy in the latest publications, being recurrent over the current state-of-the-art approaches, restricts candidate pairs based on token windows. Thus, there are no criteria based on the same or different sentences; they are based only on the token distance. This strategy was used in [95, 97, 102, 103, 107, 109]. A token window of 30 was used in [95], [97], and [109], with the authors of [97] and [109] restricting it to entities in the same paragraph. A token window of 60 was used in [102] and 100 in [107]. Further, the authors of [103] tested token windows of 60, 70, and 100.

Overall, traditional machine learning approaches restricted to within-sentence pairs achieved the best results. However, the authors of [96] and [97] achieved competitive results with the token window strategy. Previous deep learning–based state-of-the-art approaches restricted candidate

pairs to within-sentence pairs, but the latest considered all possible pairs over a certain token window.

In addition to the candidate pair selection strategy, another important topic is the strategy used to extract TLINKs. We summarize the approaches in traditional machine learning and deep learning, aiming to provide an overview of the evolution of approaches over time.

For traditional machine learning, most approaches use SVM classifiers. SVM classifiers were used in [6, 12, 28, 31, 33, 34, 94, 96, 100, 108, 127, 130]. CRF classifiers were used in [83–85]. Additionally, the authors of [111] used an XGboost classifier, and the authors of [35] used naïve Bayes classifiers.

SVM classifiers outperformed the previously mentioned traditional machine learning classifiers, with separate classifiers for TLINKs between events (event-event) and between events and temporal expressions (event-time). Event-event relations are more complicated because of lower annotation quality and because they suffer more from imbalance with a higher number of negative samples when training the classifiers [100]. Additionally, event-event and event-time TLINKs have different characteristics because they occur in different contexts. Thus, creating separate classifiers with different sets of features is effective.

Separate SVM classifiers for within-sentence event-event and event-time TLINKs were used in [6, 12, 28, 94, 96, 100, 108, 127]. This approach was also valid for articles dealing with cross-sentence relations. The authors of [28], [94], and [127] used an SVM classifier for event-event and another SVM classifier for event-time relations. The authors of [94] used two more SVM classifiers to deal with event-event and event-time pairs across more than two sentences.

Cost-sensitive learning was used in [12], [94], and [112] to mitigate the class imbalance by adding different costs for each class misclassification.

The feature set for traditional machine learning classifiers is similar to the features detailed in Section 4.2, with additional features representing the relation between entities. These features generally rely on extracting information about the dependency path, conjunctions, number of words, and words between entities. For event-event TLINKs, the presence of overlapped heads is generally used to detect co-references.

A combination of different classifiers was used in [93] and [112]. This strategy was beneficial in [93], with comparable results to the state-of-the-art SVM-based models for the Clinical TempEval 2016 dataset. The authors of [93] combined classifiers from different publications with ILP. The classifiers were obtained from [28, 86, 91, 94, 96].

Approaches based on MTL have also been effective in [97] and [109]. These approaches have already been detailed in Section 4.2, as they jointly predict DocTimeRel and TLINKs. Furthermore, Leeuwenberg and Moens [96, 97] had the best performance for both Clinical TempEval 2016 and 2017 datasets when deep learning–based systems were not considered.

However, machine learning systems do not perform as well as deep learning–based systems. In addition to the algorithms, some strategies have improved the results for deep learning–based systems.

One widely used strategy for the THYME corpus was developed in [106] and consisted of transforming the two-class classification task (CONTAINS and NO RELATION) into a three-class classification task (CONTAINS, NO RELATION, and IS CONTAINED). All pairs from left to right were considered, and the label was changed to IS CONTAINED when necessary. Further, not considering all possible permutations by only considering pairs that occur from left to right reduces the number of candidate pairs to half, mitigating the class imbalance problem. This strategy was used in [20, 31, 34, 87, 99, 102, 103, 113].

Another popular strategy is to expand the training set with additional examples. This strategy is helpful for both machine learning and deep learning systems. The authors of [12] and [100]

developed a training set expansion technique based on the UMLS, looking at the UMLS entities that overlapped with the annotated event spans. The authors of [37] proposed creating additional artificial training data using a transformer model with language generation. In addition, The authors of [99] and [102] used unlabeled THYME corpus additional data to generate more training instances with self-training, using cTAKES to generate events and temporal expressions over the unlabeled data. However, self-training was based on a Bi-LSTM model in [99], while a strategy based on fine-tuning the BERT was used in [102].

The encoding of relation arguments by XML tags was first introduced in [87]. It was modified in [101] to represent the temporal expressions with a single pseudo-token. This modified version was used in [99, 101, 104].

Among the architectures used, we differentiate between publications that addressed only within-sentence relations and those that addressed cross-sentence relations. The conclusions are based on the complete TLINKs set, but the comparison is fairer this way.

Among publications that addressed only within-sentence relations, the authors of [89] and [90] used tree-based Bi-LSTM-RNNs, the authors of [99] used BI-LSTM with self-training, the authors of [87] used CNNs, the authors of [101] used a hybrid approach based on a CNN and an SVM model, the authors of [104] used GRUs and attention, and the authors of [77] used RNNs, attention, and piecewise representation. Based on these results, we highlight [89] and [99]. The authors of [89] adapted the tree-based Bi-LSTM-RNN model in [133], making new sentence-level annotations to adapt the input, relying on the dependency structure between the pair and the output. The authors of [99] combined several factors that were successful in the previous approaches. The approach used a Bi-LSTM model, additionally encoding relations with XML tags, transforming into a 3-class classification task, and adding training samples with self-training.

Among publications that addressed within-sentence and cross-sentence relations, [95] combined LSTM and MTL, [105] and [113] used BI-LSTM models, [107] used context segmentation and associate attention, [102] fine-tuned BERT and used self-training, and [103] fine-tuned BERT with MTL. Based on the results, we highlight [102] and [103]. The authors of [102] combined the fine-tuning of BioBERT [134], a pre-trained model on biomedical texts, self-training and transforming into a 3-class classification problem. The authors of [103] used a one-pass BERT model that leverages global embeddings and MLT to jointly predict TLINKs and DocTimeRel.

5.3 Hybrid Systems

This section analyzes the articles that used hybrid systems for TLINKs (summarized in Table 9).

Most articles in this section are related to the i2b2 2012 dataset, where specific TLINKs, especially cross-sentence TLINKs, were extracted with rules. In this section, we analyze the candidate pair selection strategy and the approach used to extract the TLINKs.

For candidate pair selection, the most successful approaches have developed different strategies to generate pairs for within-sentence and cross-sentence TLINKs.

To create pairs for within-sentence relations, a common strategy is to create all possible pairs within a sentence. This strategy was used in [7, 12, 38, 72, 106, 135]. The authors of [11] considered all consecutive pairs in a sentence or pairs with a dependency relation. The authors of [73] and [74] used the strategy proposed in [11]. Both strategies were successful, with [11] being more restrictive in terms of the number of created pairs.

To create pairs for cross-sentence relations, the typical strategies were to restrict the pairs to all possible pairs in a sentence range or develop strategies focused on creating pairs for specific cases, such as co-references. The first strategy was used in [12] and [106], with a restriction for consecutive sentences in [12] and a restriction of a three-sentence window in [106]. For the second strategy, the authors of both [11] and [72] focused on co-referencing event-event pairs, creating

Table 9. Articles Related to TLINK that Used Hybrid Systems

Authors	Best strategy	Candidate pair selection	WS	CS	NS	SE	Results
I2b2 2012 dataset							
[74]	[11] + rules + additional features	WS: rules, CS: rules	√	√			**Fm 0.702**
[72]	WS: 2 ME clfs; CS: 1 ME clf + rules	WS: APP, CS: rules	√	√			**Fm 0.6954**
[12]	WS: 2 SVM clfs; CS: 2 SVM clfs + rules; CSL; TS expansion	WS: APP, CS: rules	√	√			**Fm 0.695**
[11]	WS: 2 SVM clfs; CS: 2 SVM clfs; Rules	WS: rules, CS: rules	√	√			Fm 0.6932
[73]	[11] + rules + additional features	WS: rules, CS: rules	√	√			Fm 0.693
[7]	WS: 2 SVM clfs + temporal graph; CS: rules	WS: APP	√	√			Fm 0.63
[76]	9 clfs + rules	Cross-product	√	√			Fm 0.6231
[71]	2 ME clfs + rules	Rules	√	√			Fm 0.5628
[135]	SVM clf + rules	WS: APP, CS: rules	√	√			Fm 0.537
[23]	WS: ME clf + conflict resolution, CS: rules	WS: rules	√	√			Fm 0.43
[75]	[73, 74]	[73, 74]		√			Fm 0.341 (NTC)
[38]	SVM clfs + rules + CSL	WS: APP	√				Fm 0.6377 (NTC)
Clinical TempEval 2015 dataset							
[128]	CRF clf + rules	Rules	√	√		√	Fm 0.181
Clinical TempEval 2016 dataset							
[106]	WS: SVM clf; CS: SVM clf; Rules; 3-class	WS: APP, CS: rules	√	√		√	Fm 0.538

Legend: WS, within-sentence; CS, cross-sentence; NS, not specified; SE, separate evaluation; Fm, f-measure; ME, maximum entropy; clf, classifier; APP, all possible pairs; CSL, cost-sensitive learning; TS, training set; CTE, Clinical TempEval; CRF, conditional random fields; 3-class, transforming to a 3-class classification.

pairs of events with matching attributes. Additionally, the authors of [11] added a criterion to consider only events with the same head noun. The authors of [11] also focused on the main events, considering pairs involving all first and last events in two consecutive sentences. The authors of [73] and [74] also used the strategy found in [11].

The approaches used to extract the TLINKs were SVM classifiers in [7, 11, 12, 38, 73–75, 106, 135], CRF classifiers in [128], and ME classifiers in [72]. All of these approaches also used rules to infer TLINKs or solve conflicts between classifier predictions. Strategies that were effective in the previous section, such as cost-sensitive learning and training set expansion, were used in [12] (further details in Section 5.2).

For within-sentence relations, as in the previous section, the most successful approach was to create separate classifiers for event-event and event-time TLINKs. This strategy was used in [7, 11, 12, 72–74].

We highlight some approaches for cross-sentence relations. The authors of [77] used a classifier for event-event and another for event-time. The authors of [11] and [72] used a classifier to detect co-references, but the authors of [11] used an additional classifier to detect the main events. The authors of [73] and [74] also used the strategy proposed in [11].

Regarding the rules created by the publications, we highlight [11, 12, 72–74]. The authors of [11] developed a rule-based result-merging module. The authors of [12] focused on creating rules

for detecting co-references in different sentences. The authors of [72] also developed rules for inferring cross-sentence relations. The framework proposed by the authors of [73] and [74] initially attempted prediction with rules and then relied on the machine learning–based models if no rule was applicable.

We highlight [11, 12, 72–74] according to their performance over the i2b2 2012 dataset. Further, the state-of-the-art for i2b2 2012 belongs to [74]. The authors of [74] developed a hybrid system based on [11], with a preference for inferring TLINKs with rules and a more elaborate feature set. They employed discourse-based features along with domain-independent and domain-dependent semantics.

5.4 TLINK Conclusions

Most publications on TLINK extraction were based on datasets made available by shared tasks. The datasets were the THYME corpus, which is related to the Clinical TempEval shared tasks and the i2b2 2012 corpus. However, the most recent searches were related to the THYME corpus. For instance, the state-of-the-art for the i2b2 2012 corpus belongs to an approach developed in 2014, evidencing that this dataset is not as widely used for TLINK extraction as the THYME corpus.

However, a downside is that few recent studies have addressed different datasets from the clinical domain in their evaluations; primarily, they have extracted TLINKs from different THYME corpus portions (Clinical TempEval 2016, 2017 edition corpora). This is unlike the NER-related publications, which generally provide evaluations for multiple datasets, such as the evaluations for BioBERT or ClinicalBERT [136] pre-trained models. Thus, there is a need for more annotated datasets for TLINK extraction, especially for different medical specialties and clinical text types. This way, the approaches can be evaluated over different scenarios, and a more solid evaluation can be obtained.

Two factors were relevant when defining approaches to extract TLINK: a strategy to generate candidate pairs and a strategy to extract TLINKs. Recently, the most common approach has been to delimit within-sentence pairs or operate over a token window. Current state-of-the-art approaches restrict pairs based on a token window ranging from 60 to 100 tokens. However, there was no further analysis of these strategies' effects over the patient timeline, such as when evaluating any essential positive pair about the patient condition that was entirely ignored by the candidate pair generation technique.

There has been an evident evolution of TLINK extraction techniques over the years, from completely rule-based systems to traditional machine learning-based systems with different heuristics and several specialized classifiers, and then to deep learning–based systems. First, models based on CNN, LSTM, and Bi-LSTM were developed, but attention-based models started achieving superior results.

Regarding the embeddings, besides word embeddings, we noticed character embeddings in [105] and contextualized embeddings in [102] and [103]. Some authors pre-trained word embeddings based on medical and biomedical corpora. A comparison between the concatenation of word embeddings is provided in [99], with the best results involving combining word embeddings from concatenated general and clinical domains, with the clinical word embeddings being pre-trained on the MIMIC-III corpus [137] and unbalanced THYME corpus notes. For contextualized embeddings, the authors of [103] used the $\text{BERT}_{\text{base}}$ model, while the authors of [102] conducted experiments with BERTbase, BioBERT, and a pre-trained BERT model in MIMIC-III clinical notes. BioBERT achieved slightly superior results for the Clinical TempEval 2016 datasets when compared with the $\text{BERT}_{\text{base}}$.

The state-of-the-art now resides on BERT pre-trained models by creating candidate pairs over token windows and transforming the classification task into a three-class classification task

(detailed in Section 5.2). This approach was used in [102] and [103]. Each publication has its own strategies. The authors of [102] focused on generating additional self-training instances, while [103] jointly predicted TLINKs and DocTimeRel using an MTL-based approach.

These results show the power of language models pre-trained using transformers, such as BERT, which can replace word embeddings; this is because the embeddings are contextualized. They can be fine-tuned to several NLP tasks using an additional output layer [138].

6 TEMPORAL RELATION EXTRACTION IN THE GENERAL DOMAIN

This section provides an overview of the datasets and publications relevant to the general domain. The TempEval-3 (TE-3) corpus was related to the TempEval 2013 shared task and was based on the AQUAINT and TimeBank [139] corpora. According to the authors of [140], the annotators only labeled relations key to understanding the document during the TimeBank annotation process, which resulted in sparse annotations. Therefore, they annotated the TimeBank-Dense, which increased the number of annotations, considered additional relation categories, simplified the relation types, and added a VAGUE relation type.

Regarding the TE-3 corpus, we highlight the ClearTK-TimeML system, which is generally used as a baseline for the TE-3 corpus and has been applied to the clinical domain in [6] (see Section 5.2). For TimeBank-Dense, we highlight the CAEVO system [141], which is generally used as a baseline for TimeBank-Dense, a sieve-based approach that uses smaller specialized classifiers while leveraging rules.

Among recent publications, we have highlighted [142–144]. All of them used contextualized word embedding as an input in their systems. The authors of [142] used ELMo [145] contextual embeddings and attention mechanisms to jointly predict event duration and temporal relations with MLP. The authors of [143] used BERT and POS embeddings as inputs for a model based on Bi-LSTM and structured SVM, and verified the performance improvement with contextualized embeddings. The authors of [144] combined contextualized word embeddings, Siamese networks, and ILP, and verified that both BERT and ELMo improved the results. Hence, the contextual representations were only used as embeddings but improved the results.

7 CONCLUSIONS

This article reviews existing temporal relation extraction approaches in clinical texts, dividing temporal relations into DocTimeRel and TLINKs. The DocTimeRel relation extraction is less complicated than TLINK extraction, as evidenced by the performance over the datasets. The DCT is a temporal expression of the date type, which is completed by having explicit information about the year, month, and day. Additionally, depending on the annotation scheme, an event always has a temporal relation with the DCT, with no need to create candidate pairs connecting the events to diverse temporal expressions over the document. Hence, DocTimeRel relations do not suffer from the same imbalance that TLINKs suffer. This is because, for TLINKs, most of the created pairs have no relation, being negative examples to the classifiers. In contrast to TLINK extraction, which is actively researched to push the state-of-the-art, DocTimeRel is a secondary research topic. For DocTimeRel, most of the articles relied on traditional machine learning approaches, especially SVMs. However, recent architectures based on MTL have started achieving positive results. TLINKs have been an active field of research over the past years, with the current state-of-the-art based on contextual embeddings and approaches based on BERT. In recent publications, training set expansion and MTL have positively impacted the results.

Most publications on TLINKs are based on a single dataset, limiting the evaluation of the approaches in different medical specialties, clinical text types, and languages. Research on this topic

would improve if additional datasets with different medical specialties, clinical text types, and languages were made available to the research community. For instance, in a survey for primary care consultation, it was discovered that in 18 countries, the average consultation time was 5 minutes or less [146]. It would be interesting to analyze how well a system would perform for this short and highly structured clinical text type.

Additionally, the TLINK extraction performance for clinical texts is relatively low compared with other NLP tasks, such as event and temporal expression extraction. For the dataset available in the Clinical TempEval 2016, the primary research target, the state-of-the-art approaches, achieved an f-measure close to 0.7, with only one relation type being considered (CONTAINS). This result would not be suitable for an actual application in the clinical domain, where every misclassification can negatively impact clinical decision-making. Hence, studies on performance improvement are necessary. For instance, simplifying the event annotation guideline to be less extensive could lower task complexity by reducing the number of events and candidate pairs. Furthermore, providing additional annotated data could also improve the results.

Furthermore, in TLINK extraction evaluation, event and temporal expression inputs are generally gold-standard annotations. This would directly impact the performance in an end-to-end system scenario by adding noise to the TLINK extraction system. Thus, considering the current state of the research field, creating a real-time use in a clinical configuration is still a long way in the future.

Based on our analysis, we identified directions for future research in temporal relation extraction from clinical texts. One research topic involves fine-tuning pre-trained models with clinical texts, such as ClinicalBERT, which could improve the ability of the model to understand the clinical context. Furthermore, contextual representations such as BERT and its variants (e.g., distilBERT [147] and RoBERTa [148]) or representations such as XLNet [149] could be used for relation extraction. Additionally, several studies have demonstrated the positive effect of extending the training set. Therefore, studying and evaluating data augmentation techniques could benefit future research. One research direction that could be beneficial not only to relation extraction but also to several other NLP tasks would be to develop models pre-trained on biomedical and medical texts, especially for languages other than English.

Although we achieved our review goals, we did not discuss which areas within the clinical domain are directly affected by temporal relation extraction and how improving the temporal relation extraction framework results could benefit them in the future.

REFERENCES

[1] Peter B. Jensen, Lars J. Jensen, and Søren Brunak. 2012. Mining electronic health records: Towards better research applications and clinical care. *Nature Reviews Genetics* 13, 6 (2012), 395–405. DOI: https://doi.org/10.1038/nrg3208

[2] Daniel Capurro, Meliha Yetisgen, Erik van Eaton, Robert Black, and Peter Tarczy-Hornoch. 2014. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: A multisite assessment. *EGEMS* 2, 1 (2014), 1079. DOI: https://doi.org/10.13063/2327-9214.1079

[3] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics* 77 (2018), 34–49. DOI: https://doi.org/10.1016/j.jbi.2017.11.011

[4] Kory Kreimeyer, Matthew Foster, Abhishek Pandey, Nina Arya, Gwendolyn Halford, Sandra F Jones, Richard Forshee, Mark Walderhaug, and Taxiarchis Botsis. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics* 73 (2017), 14–29. DOI: https://doi.org/10.1016/j.jbi.2017.07.012

[5] Sumithra Velupillai, Danielle Mowery, Brett South, Maria Kvist, and Hercules Dalianis. 2015. Recent advances in clinical natural language processing in support of semantic analysis. *Yearbook of medical informatics* 10 (2015), 183–93. DOI: https://doi.org/10.15265/IY-2015-009

[6] William F. Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, and James Pustejovsky. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2 (2014), 143–154. DOI: https://doi.org/10.1162/tacl_a_00172

[7] Azadeh Nikfarjam, Ehsan Emadzadeh, and Graciela Gonzalez. 2013. Towards generating a patient's timeline: Extracting temporal relationships from clinical notes. *Journal of Biomedical Informatics* 46 (2013), S40–S47. DOI: https://doi.org/10.1016/j.jbi.2013.11.001

[8] Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T. Dudley, Alberto Lavelli, Fabio Rinaldi, and Venet Osmani. 2019. Natural language processing of clinical notes on chronic diseases: Systematic review. *JMIR Med Inform* 7, 2 (2019), e12239. DOI: https://doi.org/10.2196/12239

[9] World Health Organization. 2020. *Noncommunicable Diseases: Progress Monitor 2020.* World Health Organization. vi, 224 pages.

[10] James F. Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26, 11 (1983), 832–843. DOI: https://doi.org/10.1145/182.358434

[11] Buzhou Tang, Yonghui Wu, Min Jiang, Yukun Chen, Joshua C. Denny, and Hua Xu. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association* 20, 5 (2013), 828–835. DOI: https://doi.org/10.1136/amiajnl-2013-001635

[12] Chen Lin, Dmitriy Dligach, Timothy A. Miller, Steven Bethard, and Guergana K. Savova. 2016. Multilayered temporal modeling for the clinical domain. *Journal of the American Medical Informatics Association* 23, 2 (2016), 387–395. DOI: https://doi.org/10.1093/jamia/ocv113

[13] Robert Leaman, Ritu Khare, and Zhiyong Lu. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics* 57, C (2015), 28–37. DOI: https://doi.org/10.1016/j.jbi.2015.07.010

[14] Stephane M. Meystre, Guergana K. Savova, Karin C. Kipper-Schuler, and John F. Hurdle. 2008. Extracting information from textual documents in the electronic health record: A review of recent research. *Yearbook of Medical Informatics* (2008), 128–144.

[15] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association* 20, 5 (2013), 806–813. DOI: https://doi.org/10.1136/amiajnl-2013-001628

[16] Steven Bethard, Leon Derczynski, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. SemEval-2015 Task 6: Clinical TempEval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15).* Association for Computational Linguistics, Denver, Colorado, 806–814. DOI: https://doi.org/10.18653/v1/S15-2136

[17] Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. SemEval-2016 Task 12: Clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16).* Association for Computational Linguistics, San Diego, California, 1052–1062. DOI: https://doi.org/10.18653/v1/S16-1165

[18] Steven Bethard, Guergana Savova, Martha Palmer, and James Pustejovsky. 2017. SemEval-2017 Task 12: Clinical TempEval. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17).* Association for Computational Linguistics, Vancouver, Canada, 565–572. DOI: https://doi.org/10.18653/v1/S17-2093

[19] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, and The PRISMA Group. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *PLOS Medicine* 6, 7 (2009), 1–6. DOI: https://doi.org/10.1371/journal.pmed.1000097

[20] Weiyi Sun, Anna Rumshisky, and Ozlem Uzuner. 2013. Temporal reasoning over clinical text: The state of the art. *Journal of the American Medical Informatics Association* 20, 5 (2013), 814–819. DOI: https://doi.org/10.1136/amiajnl-2013-001760

[21] Ghada Alfattni, Niels Peek, and Goran Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics* 108 (2020), 103488. DOI: https://doi.org/10.1016/j.jbi.2020.103488

[22] James Pustejovsky, Kiyong Lee, Harry Bunt, and Laurent Romary. 2010. ISO-TimeML: An international standard for semantic annotation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC'10).* European Language Resources Association (ELRA), Valletta, Malta, 394–397.

[23] Yao Cheng, Peter Anick, Pengyu Hong, and Nianwen Xue. 2013. Temporal relation discovery between events and temporal expressions identified in clinical narrative. *Journal of Biomedical Informatics* 46 (2013), S48–S53. DOI: https://doi.org/10.1016/j.jbi.2013.09.010

[24] James Pustejovsky, Robert Knippen, Jessica Littman, and Roser Saurí. 2005. Temporal and event information in natural language text. *Language Resources and Evaluation* 39, 2 (2005), 123–164. DOI: https://doi.org/10.1007/s10579-005-7882-7

[25] Marc Verhagen. 2005. Temporal closure in an annotation environment. *Language Resources and Evaluation* 39, 2 (2005), 211–241. DOI: https://doi.org/10.1007/s10579-005-7884-5

[26] Leon Derczynski. 2016. Representation and learning of temporal relations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. The COLING 2016 Organizing Committee*, Osaka, Japan, 1937–1948.

[27] James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*. Association for Computational Linguistics, Portland, Oregon, 152–160.

[28] Abdulrahman Khalifa, Sumithra Velupillai, and Stephane Meystre. 2016. UtahBMI at SemEval-2016 Task 12: Extracting temporal information from clinical text. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1256–1262. DOI: https://doi.org/10.18653/v1/S16-1195

[29] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (2010), 507–513. DOI: https://doi.org/10.1136/jamia.2009.001560

[30] Alan R. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proceedings AMIA Symposium* (2001), 17–21.

[31] Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. Temporal information extraction from clinical text. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 739–745.

[32] Wei Wang, Kory Kreimeyer, Emily Jane Woo, Robert Ball, Matthew Foster, Abhishek Pandey, John Scott, and Taxiarchis Botsis. 2016. A new algorithmic approach for the extraction of temporal associations from clinical narratives with an application to medical product safety surveillance reports. *Journal of Biomedical Informatics* 62 (2016), 78–89. DOI: https://doi.org/10.1016/j.jbi.2016.06.006

[33] Timothy Miller, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, Chen Lin, and Guergana Savova. 2013. Discovering temporal narrative containers in clinical text. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Sofia, Bulgaria, 18–26.

[34] Chen Lin, Timothy Miller, Alvin Kho, Steven Bethard, Dmitriy Dligach, Sameer Pradhan, and Guergana Savova. 2014. Descending-path convolution kernel for syntactic structures. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 81–86. DOI: https://doi.org/10.3115/v1/P14-2014

[35] Gandhimathi Moharasan and Tu-Bao Ho. 2019. Extraction of temporal information from clinical narratives. *Journal of Healthcare Informatics Research* 3, 2 (2019), 220–244. DOI: https://doi.org/10.1007/s41666-019-00049-0

[36] Thanat Chokwijitkul, Anthony Nguyen, Hamed Hassanzadeh, and Siegfried Perez. 2018. Identifying risk factors for heart disease in electronic medical records: A deep learning approach. In *Proceedings of the BioNLP 2018 Workshop*. Association for Computational Linguistics, Melbourne, Australia, 18–27. DOI: https://doi.org/10.18653/v1/W18-2303

[37] Zixu Wang, Julia Ive, Sumithra Velupillai, and Lucia Specia. 2019. Is artificial data useful for biomedical Natural Language Processing algorithms?. In *Proceedings of the 18th BioNLP Workshop and Shared Task*. Association for Computational Linguistics, Florence, Italy, 240–249. DOI: https://doi.org/10.18653/v1/W19-5026

[38] Hee Jin Lee, Yaoyun Zhang, Min Jiang, Jun Xu, Cui Tao, and Hua Xu. 2018. Identifying direct temporal relations between time and events from clinical notes. *BMC Medical Informatics and Decision Making* 18, Suppl 2 (2018), 49. DOI: https://doi.org/10.1186/s12911-018-0627-5

[39] Dong Xu, Meizhuo Zhang, Tianwan Zhao, Chen Ge, Weiguo Gao, Jia Wei, and Kenny Q. Zhu. 2015. Data-driven information extraction from Chinese electronic medical records. *PLOS ONE* 10, 8 (2015), 1–18. DOI: https://doi.org/10.1371/journal.pone.0136270

[40] Jia Su, Jinpeng Hu, Jingchi Jiang, Jing Xie, Yang Yang, Bin He, Jinfeng Yang, and Yi Guan. 2019. Extraction of risk factors for cardiovascular diseases from Chinese electronic medical records. *Computer Methods and Programs in Biomedicine* 172 (2019), 1–10.

[41] Zengjian Liu, Xiaolong Wang, Qingcai Chen, Buzhou Tang, and Hua Xu. 2019. Temporal indexing of medical entity in Chinese clinical notes. *BMC Medical Informatics and Decision Making* 19, 1 (2019), 17. DOI: https://doi.org/10.1186/s12911-019-0735-x

[42] Xiaohui Zhang, Yaoyun Zhang, Qin Zhang, Yuankai Ren, Tinglin Qiu, Jianhui Ma, and Qiang Sun. 2019. Extracting comprehensive clinical information for breast cancer using deep learning methods. *International Journal of Medical Informatics* 132 (2019), 103985. DOI: https://doi.org/10.1016/j.ijmedinf.2019.103985

[43] Jae-Wook Seol, Wangjin Yi, Jinwook Choi, and Kyung Soon Lee. 2017. Causality patterns and machine learning for the extraction of problem-action relations in discharge summaries. *International Journal of Medical Informatics* 98 (2017), 1–12. DOI: https://doi.org/10.1016/j.ijmedinf.2016.10.021

[44] Wangjin Lee and Jinwook Choi. 2018. Temporal segmentation for capturing snapshots of patient histories in Korean clinical narrative. *Healthcare Informatics Research* 24, 3 (2018), 179–186. DOI: https://doi.org/10.4258/hir.2018.24.3.179

[45] Zubair Afzal, Ewoud Pons, Ning Kang, Miriam C. J. M. Sturkenboom, Martijn J. Schuemie, and Jan A. Kors. 2014. ContextD: an algorithm to identify contextual properties of medical terms in a Dutch clinical corpus. *BMC Bioinformatics* 15, 1 (2014), 373. DOI: https://doi.org/10.1186/s12859-014- 0373-3

[46] Natalia Viani, Timothy A. Miller, Carlo Napolitano, Silvia G. Priori, Guergana K. Savova, Riccardo Bellazzi, and Lucia Sacchi. 2019. Supervised methods to extract clinical events from cardiology reports in Italian. *Journal of Biomedical Informatics* 95 (2019), 103219. DOI: https://doi.org/10.1016/j. jbi.2019.103219

[47] Aron Henriksson, Maria Kvist, Hercules Dalianis, and Martin Duneld. 2015. Identifying adverse drug event information in clinical notes with distributional semantic representations of context. *Journal of Biomedical Informatics* 57 (2015), 333–349. DOI: https://doi.org/10.1016/j.jbi.2015.08.013

[48] Marjan Najafabadipour, Massimiliano Zanin, Alejandro Rodríguez-González, Maria Torrente, Beatriz Nuñez García, Juan Luis Cruz Bermudez, Mariano Provencio, and Ernestina Menasalvas. 2020. Reconstructing the patient's natural history from electronic health records. *Artificial Intelligence in Medicine* 105 (2020), 101860. DOI: https://doi.org/10.1016/j.artmed.2020.101860

[49] Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A French clinical corpus with comprehensive semantic annotations: Development of the Medical Entity and Relation LIMSI annOtated Text corpus (MERLOT). *Language Resources and Evaluation* 52, 2 (2018), 571–601. DOI: https://doi.org/10.1007/s10579-017-9382-y

[50] Srinivasan V. Iyer, Rave Harpaz, Paea LePendu, Anna Bauer-Mehren, and Nigam H. Shah. 2014. Mining clinical text for signals of adverse drug-drug interactions. *Journal of the American Medical Informatics Association* 21, 2 (2014), 353–362. DOI: http://dx.doi.org/10.1136/amiajnl-2013-001612

[51] Srinivasan V. Iyer, Paea Lependu, Rave Harpaz, Anna Bauer-Mehren, and Nigam H. Shah. 2013. Learning signals of adverse drug-drug interactions from the unstructured text of electronic health records. *AMIA Joint Summits on Translational Science Proceedings* 2013 (2013), 83–87.

[52] Yi Liu, Paea Lependu, Srinivasan Iyer, and Nigam H. Shah. 2012. Using temporal patterns in medical records to discern adverse drug events from indications. *AMIA Joint Summits on Translational Science Proceedings* 2012 (2012), 47–56.

[53] Hee-Jin Lee, Min Jiang, Yonghui Wu, Christian M. Shaffer, John H. Cleator, Eitan A. Friedman, Joshua P. Lewis, Dan M. Roden, Josh Denny, and Hua Xu. 2017. A comparative study of different methods for automatic identification of clopidogrel-induced bleedings in electronic health records. *AMIA Joint Summits on Translational Science Proceedings* 2017 (2017), 185–192.

[54] Cosmin A. Bejan, Lucy Vanderwende, Heather L. Evans, Mark M. Wurfel, and Meliha Yetisgen-Yildiz. 2013. On-time clinical phenotype prediction based on narrative reports. *AMIA Annual Symposium Proceedings* 2013 (2013), 103–110.

[55] Stephen T. Wu, Young J. Juhn, Sunghwan Sohn, and Hongfang Liu. 2014. Patient-level temporal aggregation for text-based asthma status ascertainment. *Journal of the American Medical Informatics Association* 21, 5 (2014), 876–884. DOI: https://doi.org/10.1136/amiajnl-2013-002463

[56] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* 42, 5 (2009), 839–851. DOI: https://doi.org/10.1016/j.jbi.2009.05.002

[57] Rob Gaizauskas, Henk Harkema, Mark Hepple, and Andrea Setzer. 2006. Task-oriented extraction of temporal information: The case of clinical narratives. In *13th International Symposium on Temporal Representation and Reasoning (TIME'06)*. 188–195. DOI: https://doi.org/10.1109/TIME. 2006.27

[58] Nai-Wen Chang, Hong-Jie Dai, Jitendra Jonnagaddala, Chih-Wei Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2015. A context-aware approach for progression tracking of medical concepts in electronic medical records. *Journal of Biomedical Informatics* 58, S (2015), S150–S157. DOI: https://doi.org/10.1016/j.jbi.2015.09.013

[59] Qingcai Chen, Haodi Li, Buzhou Tang, Xiaolong Wang, Xin Liu, Zengjian Liu, Shu Liu, Weida Wang, Qiwen Deng, Suisong Zhu, Yangxin Chen, and Jingfeng Wang. 2015. An automatic system to identify heart disease risk factors in clinical texts over time. *Journal of Biomedical Informatics* 58 (2015), S158–S163. DOI: https://doi.org/10.1016/j.jbi.2015.09.002

[60] James Cormack, Chinmoy Nath, David Milward, Kalpana Raja, and Siddhartha R. Jonnalagadda. 2015. Agile text mining for the 2014 i2b2/UTHealth Cardiac Risk Factors Challenge. *Journal of Biomedical Informatics* 58 (2015), S120–S127. DOI: https://doi.org/10.1016/j.jbi.2015.06.030

[61] Travis Goodwin and Sanda M. Harabagiu. 2015. A probabilistic reasoning method for predicting the progression of clinical findings from electronic medical records. *AMIA Joint Summits on Translational Science Proceedings* 2015 (2015), 61–65.

[62] Cyril Grouin, Véronique Moriceau, and Pierre Zweigenbaum. 2015. Combining glass box and black box evaluations in the identification of heart disease risk factors and their temporal relations from clinical records. *Journal of Biomedical Informatics* 58 (2015), S133–S142. DOI: https://doi.org/10.1016/j.jbi.2015.06.014

[63] Jitendra Jonnagaddala, Siaw-Teng Liaw, Pradeep Ray, Manish Kumar, Hong-Jie Dai, and Chien-Yeh Hsu. 2015. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. *BioMed Research International* 2015 (2015), 636371. DOI: https://doi.org/10.1155/2015/636371

[64] George Karystianis, Azad Dehghan, Aleksandar Kovacevic, John A. Keane, and Goran Nenadic. 2015. Using local lexicalized rules to identify heart disease risk factors in clinical notes. *Journal of Biomedical Informatics* 58 Suppl, Suppl (2015), S183–S188. DOI: https://doi.org/10.1016/j.jbi.2015.06.013

[65] Abdulrahman Khalifa and Stéphane Meystre. 2015. Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes. *Journal of Biomedical Informatics* 58, S (2015), S128–S132. DOI: https://doi.org/10.1016/j.jbi.2015.08.002

[66] Kirk Roberts, Sonya E. Shooshan, Laritza Rodriguez, Swapna Abhyankar, Halil Kilicoglu, and Dina Demner-Fushman. 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *Journal of Biomedical Informatics* 58, S (2015), S111–S119. DOI: https://doi.org/10.1016/j.jbi.2015.06.010

[67] Chaitanya P. Shivade, Pranav Malewadkar, Eric Fosler-Lussier, and Albert M. Lai. 2015. Comparison of UMLS terminologies to identify risk of heart disease using clinical notes. *Journal of Biomedical Informatics* 58 (2015), S103–S110. DOI: https://doi.org/10.1016/j.jbi.2015.08.025

[68] Manabu Torii, Jung wei Fan, Wei li Yang, Theodore Lee, Matthew T. Wiley, Daniel S. Zisook, and Yang Huang. 2015. Risk factor detection for heart disease by applying text analytics in electronic medical records. *Journal of Biomedical Informatics* 58 (2015), S164–S170. DOI: https://doi.org/10.1016/j.jbi.2015.08.011

[69] Jay Urbain. 2015. Mining heart disease risk factors in clinical text with named entity recognition and distributional semantic models. *Journal of Biomedical Informatics* 58 (2015), S143–S149. DOI: https://doi.org/10.1016/j.jbi.2015.08.009

[70] Hui Yang and Jonathan M. Garibaldi. 2015. A hybrid model for automatic identification of risk factors for heart disease. *Journal of Biomedical Informatics* 58 Suppl, Suppl (2015), S171–S182. DOI: https://doi.org/10.1016/j.jbi.2015.09.006

[71] Yung-Chun Chang, Hong-Jie Dai, Johnny Chi-Yang Wu, Jian-Ming Chen, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2013. TEMPTING System: A hybrid method of rule and machine learning for temporal relation extraction in patient discharge summaries. *Journal of Biomedical Informatics* 46 (2013), S54–S62. DOI: https://doi.org/10.1016/j.jbi.2013.09.007

[72] Colin Cherry, Xiaodan Zhu, Joel Martin, and Berry de Bruijn. 2013. À la recherche du temps perdu: Extracting temporal relations from medical text in the 2012 i2b2 NLP challenge. *Journal of the American Medical Informatics Association* 20, 5 (2013), 843–848. DOI: https://doi.org/10.1136/amiajnl-2013-001624

[73] Jennifer D'Souza and Vincent Ng. 2013. Classifying temporal relations in clinical data: A hybrid, knowledge-rich approach. *Journal of Biomedical Informatics* 46 (2013), S29–S39. DOI: https://doi.org/10.1016/j.jbi.2013.08.003

[74] Jennifer D'Souza and Vincent Ng. 2014. Knowledge-rich temporal relation identification and classification in clinical notes. *Database* 2014 (11 2014). DOI: https://doi.org/10.1093/database/bau109

[75] Jennifer D'Souza and Vincent Ng. 2014. Annotating inter-sentence temporal relations in clinical notes. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, 2758–2765.

[76] Cyril Grouin, Natalia Grabar, Thierry Hamon, Sophie Rosset, Xavier Tannier, and Pierre Zweigenbaum. 2013. Eventual situations for timeline extraction from clinical reports. *Journal of the American Medical Informatics Association* 20, 5 (2013), 820–827. DOI: https://doi.org/10.1136/amiajnl-2013-001627

[77] Zhijing Li, Chen Li, Yu Long, and Xuan Wang. 2020. A system for automatically extracting clinical events with temporal information. *BMC Medical Informatics and Decision Making* 20, 1 (2020), 198. DOI: https://doi.org/10.1186/s12911-020-01208-9

[78] Danielle Mowery, Henk Harkema, John Dowling, Jonathan Lustgarten, and Wendy Chapman. 2009. Distinguishing historical from current problems in clinical reports—which textual features help?. In *Proceedings of the BioNLP 2009 Workshop*. Association for Computational Linguistics, Boulder, Colorado, 10–18.

[79] Preethi Raghavan, Albert Lai, and Eric Fosler-Lussier. 2012. Learning to temporally order medical events in clinical text. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 70–74.

[80] Preethi Raghavan, Eric Fosler-Lussier, and Albert Lai. 2012. Temporal classification of medical events. In *BioNLP: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Montréal, Canada, 29–37.

[81] Preethi Raghavan, Eric Fosler-Lussier, and Albert Lai. 2012. Exploring semi-supervised coreference resolution of medical concepts using semantic and temporal features. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 731–741.

[82] Chen Lin, Elizabeth W. Karlson, Dmitriy Dligach, Monica P. Ramirez, Timothy A. Miller, Huan Mo, Natalie S. Braggs, Andrew Cagan, Vivian Gainer, Joshua C. Denny, and Guergana K. Savova. 2015. Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record. *Journal of the American Medical Informatics Association* 22, e1 (2015), e151–161. DOI: https://doi.org/10.1136/amiajnl-2014-002642

[83] Marcia Barros, Andre Lamurias, Gonçalo Figueiro, Marta Antunes, Joana Teixeira, Alexandre Pinheiro, and Francisco M. Couto. 2016. ULISBOA at SemEval-2016 Task 12: Extraction of temporal expressions, clinical events and relations using IBEnt. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1263–1267. DOI: https://doi.org/10.18653/v1/S16-1196

[84] Tommaso Caselli and Roser Morante. 2016. VUACLTL at SemEval 2016 Task 12: A CRF pipeline to clinical TempEval. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1241–1247. DOI: https://doi.org/10.18653/v1/S16-1193

[85] Veera Raghavendra Chikka. 2016. CDE-IIITH at SemEval-2016 Task 12: Extraction of temporal information from clinical documents using machine learning techniques. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1237–1240. DOI: https://doi.org/10.18653/v1/S16-1192

[86] Arman Cohan, Kevin Meurer, and Nazli Goharian. 2016. GUIR at SemEval-2016 Task 12: Temporal information processing for clinical narratives. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1248–1255. DOI: https://doi.org/10.18653/v1/S16-1194

[87] Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Association for Computational Linguistics, Valencia, Spain, 746–751.

[88] Jason Fries. 2016. Brundlefly at SemEval-2016 Task 12: Recurrent neural networks vs. Joint Inference for Clinical Temporal Information Extraction. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1274–1279. DOI: https://doi.org/10.18653/v1/S16-1198

[89] Diana Galvan, Koji Matsuda, Naoaki Okazaki, and Kentaro Inui. 2020. Empirical exploration of the challenges in temporal relation extraction from clinical text. *Journal of Natural Language Processing* 27, 2 (2020), 383–409. DOI: https://doi.org/10.5715/jnlp.27.383

[90] Diana Galvan, Naoaki Okazaki, Koji Matsuda, and Kentaro Inui. 2018. Investigating the challenges of temporal relation extraction from clinical text. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Brussels, Belgium, 55–64. DOI: https://doi.org/10.18653/v1/W18-5607

[91] Cyril Grouin and Véronique Moriceau. 2016. LIMSI at SemEval-2016 Task 12: Machine-learning and temporal information to identify clinical events and time expressions. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1225–1230. DOI: https://doi.org/10.18653/v1/S16-1190

[92] Serena Jeblee and Graeme Hirst. 2018. Listwise temporal ordering of events in clinical notes. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Brussels, Belgium, 177–182. DOI: https://doi.org/10.18653/v1/W18-5620

[93] Catherine Kerr, Terri Hoare, Paula Carroll, and Jakub Mareček. 2020. Integer programming ensemble of temporal relations classifiers. *Data Mining and Knowledge Discovery* 34, 2 (2020), 533–562. DOI: https://doi.org/10.1007/s10618-019-00671-x

[94] Hee-Jin Lee, Hua Xu, Jingqi Wang, Yaoyun Zhang, Sungrim Moon, Jun Xu, and Yonghui Wu. 2016. UTHealth at SemEval'16 Task 12: An end-to-end system for temporal information extraction from clinical notes. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1292–1297. DOI: https://doi.org/10.18653/v1/S16-1201

[95] Artuur Leeuwenberg and Marie-Francine Moens. 2018. Word-level loss extensions for neural temporal relation classification. In *Proceedings of the 27th International Conference on Computational Linguistics*. Association for Computational Linguistics, Santa Fe, New Mexico, 3436–3447.

[96] Artuur Leeuwenberg and Marie-Francine Moens. 2016. KULeuven-LIIR at SemEval 2016 Task 12: Detecting narrative containment in clinical records. In *Proceedings of the 10th International Workshop on Semantic Evaluation*

(SemEval'16). Association for Computational Linguistics, San Diego, California, 1280–1285. DOI: https://doi.org/10.18653/v1/S16-1199

[97] Artuur Leeuwenberg and Marie-Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain, 1150–1158.

[98] Peng Li and Heng Huang. 2016. UTA DLNLP at SemEval-2016 Task 12: Deep learning based natural language processing system for clinical information identification from clinical notes and pathology reports. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1268–1273. DOI: https://doi.org/10.18653/v1/S16-1197

[99] Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proceedings of the 9th International Workshop on Health Text Mining and Information Analysis*. Association for Computational Linguistics, Brussels, Belgium, 165–176. DOI: https://doi.org/10.18653/v1/W18-5619

[100] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2016. Improving temporal relation extraction with training instance augmentation. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Berlin, Germany, 108–113. DOI: https://doi.org/10.18653/v1/W16-2914

[101] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP 2017*. Association for Computational Linguistics, Vancouver, Canada, 322–327. DOI: https://doi.org/10.18653/v1/W17-2341

[102] Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A BERT-based universal model for both within- and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 65–71. DOI: https://doi.org/10.18653/v1/W19-1908

[103] Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020. A BERT-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*. Association for Computational Linguistics, Online, 70–75. DOI: https://doi.org/10.18653/v1/2020.bionlp-1.7

[104] Sijia Liu, Liwei Wang, Vipin Chaudhary, and Hongfang Liu. 2019. Attention neural model for temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 134–139. DOI: https://doi.org/10.18653/v1/W19-1917

[105] Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2017. Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 224–230. DOI: https://doi.org/10.18653/v1/P17-2035

[106] Julien Tourille, Olivier Ferret, Aurélie Névéol, and Xavier Tannier. 2016. LIMSI-COT at SemEval-2016 Task 12: Temporal relation identification using a pipeline of classifiers. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval'16)*. Association for Computational Linguistics, San Diego, California, 1136–1142. DOI: https://doi.org/10.18653/v1/S16-1175

[107] Shiyi Zhao, Lishuang Li, Hongbin Lu, Anqiao Zhou, and Shuang Qian. 2019. Associative attention networks for temporal relation extraction from electronic health records. *Journal of Biomedical Informatics* 99 (2019), 103309. DOI: https://doi.org/10.1016/j.jbi.2019.103309

[108] Po-Yu Huang, Hen-Hsen Huang, Yu-Wun Wang, Ching Huang, and Hsin-Hsi Chen. 2017. NTU-1 at SemEval-2017 Task 12: Detection and classification of temporal events in clinical data with domain adaptation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada, 1010–1013. DOI: https://doi.org/10.18653/v1/S17-2177

[109] Artuur Leeuwenberg and Marie-Francine Moens. 2017. KULeuven-LIIR at SemEval'17 Task 12: Cross-domain temporal information extraction from clinical records. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada, 1030–1034. DOI: https://doi.org/10.18653/v1/S17-2181

[110] Yu Long, Zhijing Li, Xuan Wang, and Chen Li. 2017. XJNLP at SemEval-2017 Task 12: Clinical temporal information extraction with a hybrid model. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada, 1014–1018. DOI: https://doi.org/10.18653/v1/S17-2178

[111] Sean MacAvaney, Arman Cohan, and Nazli Goharian. 2017. GUIR at SemEval-2017 Task 12: A framework for cross-domain clinical temporal information extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada, 1024–1029. DOI: https://doi.org/10.18653/v1/S17-2180

[112] Sarath P. R. Manikandan Ravikiran, and Yoshiki Niwa. 2017. Hitachi at SemEval-2017 Task 12: System for temporal information extraction from clinical notes. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada, 1005–1009. DOI: https://doi.org/10.18653/v1/S17-2176

[113] Julien Tourille, Olivier Ferret, Xavier Tannier, and Aurélie Névéol. 2017. LIMSI-COT at SemEval-17 Task 12: Neural architecture for temporal information extraction from clinical narratives. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval'17)*. Association for Computational Linguistics, Vancouver, Canada, 597–602. DOI: https://doi.org/10.18653/v1/S17-2098

[114] Li Rumeng, Jagannatha Abhyuday N, and Yu Hong. 2017. A hybrid neural network model for joint prediction of presence and period assertions of medical events in clinical notes. *AMIA Annual Symposium Proceedings* 2017 (2017), 1149–1158.

[115] Min Li and Jon Patrick. 2012. Extracting temporal information from electronic patient records. *AMIA Annual Symposium Proceedings* 2012 (2012), 542–551.

[116] Joshua C. Denny, Josh F. Peterson, Neesha N. Choma, Hua Xu, Randolph A. Miller, Lisa Bastarache, and Neeraja B. Peterson. 2010. Extracting timing and status descriptors for colonoscopy testing from electronic medical records. *Journal of the American Medical Informatics Association* 17, 4 (2010), 383–388. DOI: https://doi.org/10.1136/jamia.2010.004804

[117] Sijia Liu, Liwei Wang, Donna Ihrke, Vipin Chaudhary, Cui Tao, Chunhua Weng, and Hongfang Liu. 2017. Correlating lab test results in clinical notes with structured lab data: A case study in HbA1c and glucose. *AMIA Joint Summits on Translational Science Proceedings* 2017 (2017), 221–228.

[118] Guergana K. Savova, Janet E. Olson, Sean P. Murphy, Victoria L. Cafourek, Fergus J. Couch, Matthew P. Goetz, James N. Ingle, Vera J. Suman, Christopher G. Chute, and Richard M. Weinshilboum. 2012. Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record. *Journal of the American Medical Informatics Association* 19, e1 (2012). DOI: https://doi.org/10.1136/amiajnl-2011-000295

[119] Natalia Viani, Joyce Kam, Lucia Yin, Somain Verma, Robert Stewart, Rashmi Patel, and Sumithra Velupillai. 2019. Annotating temporal relations to determine the onset of psychosis symptoms. *Studies in Health Technology and Informatics* 264 (2019), 418–422. DOI: https://doi.org/10.3233/SHTI190255

[120] Amber Stubbs and Benjamin Harshfield. 2010. Applying the TARSQI toolkit to augment text mining of EHRs. In *Proceedings of the 2010 Workshop on Biomedical Natural Language Processing*. Association for Computational Linguistics, Uppsala, Sweden, 141–143.

[121] Li Zhou, Simon Parsons, and George Hripcsak. 2008. The evaluation of a temporal reasoning system in processing clinical discharge summaries. *Journal of the American Medical Informatics Association* 15, 1 (2008), 99–106. DOI: https://doi.org/10.1197/jamia.M2467

[122] Daniel Capurro, Mario Barbe, Claudio Daza, Josefa Santa María, Javier Trincado, and Ignacio Gomez. 2015. ClinicalTime: Identification of patients with acute kidney injury using temporal abstractions and temporal pattern matching. *AMIA Joint Summits on Translational Science Proceedings* 2015 (2015), 46–50.

[123] Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Finding temporal order in discharge summaries. *AMIA Annual Symposium Proceedings* 2006 (2006), 81–85.

[124] Philip Bramsen, Pawan Deshpande, Yoong Keok Lee, and Regina Barzilay. 2006. Inducing temporal graphs. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Sydney, Australia, 189–198.

[125] Preethi Raghavan, Eric Fosler-Lussier, Noémie Elhadad, and Albert M. Lai. 2014. Cross-narrative temporal ordering of medical events. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 998–1008. DOI: https://doi.org/10.3115/v1/P14-1094

[126] Amber Stubbs and Özlem Uzuner. 2015. Annotating risk factors for heart disease in clinical narratives for diabetic patients. *Journal of Biomedical Informatics* 58 (2015), S78–S91. DOI: https://doi.org/10.1016/j.jbi.2015.05.009

[127] Yan Xu, Yining Wang, Tianren Liu, Junichi Tsujii, and Eric I-Chao Chang. 2013. An end-to-end system to identify temporal relation in discharge summaries: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 20, 5 (2013), 849–858. DOI: https://doi.org/10.1136/amiajnl-2012-001607

[128] Sumithra Velupillai, Danielle L Mowery, Samir Abdelrahman, Lee Christensen, and Wendy Chapman. 2015. BluLab: Temporal information extraction for the 2015 clinical TempEval challenge. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval'15)*. Association for Computational Linguistics, Denver, Colorado, 815–819. DOI: https://doi.org/10.18653/v1/S15-2137

[129] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. DOI: https://doi.org/10.18653/v1/N19-1423

[130] Kirk Roberts, Bryan Rink, and Sanda M. Harabagiu. 2013. A flexible framework for recognizing events, temporal expressions, and temporal relations in clinical text. *Journal of the American Medical Informatics Association* 20, 5 (2013), 867–875. DOI: https://doi.org/10.1136/amiajnl-2013-001619

[131] Roser Saurí, Jessica Moszkowicz, Bob Knippen, Rob Gaizauskas, Andrea Setzer, and James Pustejovsky. 2006. TimeML Annotation Guidelines Version 1.2.1. (2006).

[132] Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: From pairwise approach to listwise approach. In *Proceedings of the 24th International Conference on Machine Learning*. ACM, New York, NY, 129–136. DOI: https://doi.org/10.1145/1273496.1273513

[133] Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using LSTMs on sequences and tree structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 1105–1116. DOI: https://doi.org/10.18653/v1/P16-1105

[134] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2019), 1234–1240. DOI: https://doi.org/10.1093/bioinformatics/btz682

[135] Sunghwan Sohn Dr., Kavishwar B. Wagholikar, Dingcheng Li, Siddhartha R. Jonnalagadda, Cui Tao, Ravikumar Komandur Elayavilli, and Hongfang Liu. 2013. Comprehensive temporal information detection from clinical text: Medical events, time, and TLINK identification. *Journal of the American Medical Informatics Association* 20, 5 (2013), 836–842. DOI: https://doi.org/10.1136/amiajnl-2013-001622

[136] Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 72–78. DOI: https://doi.org/10.18653/v1/W19-1909

[137] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (2016), 160035. DOI: https://doi.org/10.1038/sdata.2016.35

[138] Jing Li, Aixin Sun, Jianglei Han, and Chenliang Li. 2020. A survey on deep learning for named entity recognition. *IEEE Transactions on Knowledge Data Engineering* 01 (2020), 1–1. DOI: https://doi.org/10.1109/TKDE.2020.2981314

[139] James Pustejovsky, Patrick Hanks, Roser Saurí, Andrew See, Rob Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, and Marcia Lazo. 2003. The TimeBank corpus. *Proceedings of Corpus Linguistics* (Jan 2003).

[140] Taylor Cassidy, Bill McDowell, Nathanael Chambers, and Steven Bethard. 2014. An annotation framework for dense event ordering. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 501–506. DOI: https://doi.org/10.3115/v1/P14-2082

[141] Nathanael Chambers, Taylor Cassidy, Bill McDowell, and Steven Bethard. 2014. Dense event ordering with a multi-pass architecture. *Transactions of the Association for Computational Linguistics* 2 (2014), 273–284. DOI: https://doi.org/10.1162/tacl_a_00182

[142] Siddharth Vashishtha, Benjamin Van Durme, and Aaron Steven White. 2019. Fine-grained temporal relation extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 2906–2919. DOI: https://doi.org/10.18653/v1/P19-1280

[143] Rujun Han, I-Hung Hsu, Mu Yang, Aram Galstyan, Ralph Weischedel, and Nanyun Peng. 2019. Deep structured neural network for event temporal relation extraction. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, Hong Kong, China, 666–676. DOI: https://doi.org/10.18653/v1/K19-1062

[144] Qiang Ning, Sanjay Subramanian, and Dan Roth. 2019. An improved neural baseline for temporal relation extraction. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP'19)*. Association for Computational Linguistics, Hong Kong, China, 6203–6209. DOI: https://doi.org/10.18653/v1/D19-1642

[145] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 2227–2237. DOI: https://doi.org/10.18653/v1/N18-1202

[146] Greg Irving, Ana Luisa Neves, Hajira Dambha-Miller, Ai Oishi, Hiroko Tagashira, Anistasiya Verho, and John Holden. 2017. International variations in primary care physician consultation time: A systematic review of 67 countries. *BMJ Open* 7, 10 (2017). DOI: https://doi.org/10.1136/bmjopen-2017-017902

[147] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. (2019). arXiv:1910.01108. Retrieved from https://arxiv.org/abs/1910.01108.

[148] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. (2019). arXiv:1907.11692. Retrieved from https://arxiv.org/abs/1907.11692.

[149] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems*, 32. Curran Associates, Inc.

Received February 2019; revised March 2021; accepted April 2021